RADIO-FREQUENCY INTEGRATED-CIRCUIT ENGINEERING

WILEY SERIES IN MICROWAVE AND OPTICAL ENGINEERING

KAI CHANG, Editor
Texas A&M University

A complete list of the titles in this series appears at the end of this volume.

RADIO-FREQUENCY INTEGRATED-CIRCUIT ENGINEERING

CAM NGUYEN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Nguyen, Cam.
Radio-frequency integrated-circuit engineering / Cam Nguyen.
1 online resource. – (Wiley series in microwave and optical engineering ; 128)
Includes bibliographical references and index.
Description based on print version record and CIP data provided by publisher; resource not viewed.
ISBN 978-1-118-93648-1 (ePub) – ISBN 978-1-118-90047-5 (Adobe PDF) – ISBN 978-0-471-39820-2 (hardback) 1. Radio frequency integrated circuits. I. Title.
TK7874.78
621.382–dc23

2014024757

Cover Images: Courtesy of the Editor

Typeset in 11/13pt TimesTenLTStd by Laserwords Private Limited, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

1 2015

This book is dedicated to my parents (Mr. and Mrs. Nguyễn Xuân Sưởng), my wife (Trần Ngọc-Diệp), and my children (Christine Nhã-Uyên, Devon, and Andrew Đình-An).

CONTENTS

PREFACE

Radio Frequency Integrated Circuits (RFICs) implemented using silicon-based technologies such as complementary metal–oxide–semiconductor (CMOS) and bipolar and complementary metal–oxide–semiconductor (BiCMOS) offer competitive performance with much lower cost and better integration capability than their non-silicon based counterparts. RFIC has become one of the most exciting areas in the radio frequency (RF) domain with contributions and impacts far reaching into the millimeter-wave range and advancing into the sub-millimeter-wave regime. Studies and research for RFIC, particularly those extending into the millimeter-wave region and beyond, across the World have exploded in the past decade and are indeed increasing rapidly.

Several years ago, when my research interests shifted from the then more well-known microwave-integrated circuits and systems to RFICs, I was looking for possible books that address RFIC design, especially from the microwave design point of view, which I consider as absolutely essential for RF operation. As a result, the long journal for this book began and its birth, long overdue, is just now matured.

As RF is moving into very high frequencies now, reaching THz, RF (as it is practiced now) is not different from microwave. RF at present implies frequencies from a few KHz up to hundreds of GHz (not a few GHz as considered before). Therefore, knowledge in electromagnetics (EM) and microwave engineering, together with passive and active RFICs, RFIC analysis and design techniques, and RF systems, is vital for RFIC engineers. Without EM and microwave engineering foundation, RFIC engineers would lack the essential background needed for designing RFICs at high frequencies. The primary objective of the book is to present the theory, analysis, and design of passive and active RFICs, including those at high frequencies beyond those in the traditional RF spectrum, aiming toward providing essential knowledge in RFIC design to graduate students and engineers. The materials in this book are self-contained and presented in such details that allow readers with only undergraduate electrical engineering knowledge in EM, RF and circuits to understand and design RFICs. The book includes problems at the end of each chapter, allowing readers to reinforce their knowledge and practice their understanding. Some of these problems are relatively long and difficult, and may thus be more suitable for class projects. The book can serve not only as a textbook for graduate students and senior undergraduate students (to some extent), but also as a reference book for practicing RFIC and microwave engineers. It is written based partly on the materials of some graduate courses on active RFICs and microwave circuits offered at the Texas A&M University and partly on the RFIC research conducted at the University. The majority of the book can be covered in two graduate semester courses (or two undergraduate courses with reduced load): one for passive RFICs and another for active RFICs.

I sincerely appreciate some of my former students (Drs. M. Miao, Y. Jin, X. Guan, M. Chirala, R. Xu, and S. Lee) for their enthusiasm in venturing into the RFIC area with me and for their contributions, and my

current Ph.D. students (C. Huynh, J. Lee, D. Lee, K. Kim, S. Jang, Y. Luo, Y. Um, J. Bae, and C. Geha) for continuing carrying out our passion in RFICs and for their help in preparing the book. Without them, our venture into RFICs would not have succeeded and this book would hence never been completed. Finally, I wish to express my deepest appreciation to the person I forever owe my indebtedness to: my wife, Ngoc-Diep Tran, for her support during the writing of this book.

CAM NGUYEN
College Station, TX, USA
Newport Beach, CA, USA

CHAPTER 1

INTRODUCTION

Wireless systems, including communication, networking, and sensing systems, play a critical role in our information-age society in many areas, from public service and safety, consumer, industry, sports, gaming, and entertainment, asset and inventory management, medicine, banking to government and military operations. The key to enabling effective wireless communications, sensing and networking is radio-frequency (RF) integrated circuits (ICs).

Radio-frequency integrated circuits (RFICs) typically refer to RF monolithic ICs fabricated on silicon (Si) substrates using complementary metal oxide semiconductor (CMOS) or BiCMOS technology. From a general perspective, however, RFICs are not and should not be limited to only Si-based CMOS and BiCMOS circuits; others like microwave monolithic integrated circuits (MMICs) using III–V semiconductors such as GaAs MMICs can also be classified as RFICs. Nevertheless, in this book, to emphasize the main objective of the book and to distinguish Si-based RFICs from other non-Si based RFICs, we will use the term RFIC to indicate Si-based CMOS/BiCMOS RFIC. The readers should, however, keep in mind that the presented materials are not limited to Si-based RFICs; they are also applicable to non-Si based RFICs such as GaAs MMICs.

The frequencies used to indicate the RF range, in general, and for RFICs, in particular, are not strictly defined in practice. To some extent, particularly in the past, the frequencies in the RF range are known as a few kilohertz to a few gigahertz and, hence, RF is clearly distinct from microwave. Since the frequencies for radio waves are normally known as between 3 KHz and 300 GHz, to a broader extent, the frequencies in the RF range can be considered from 3 KHz to 300 GHz. As the name RF implies, however, these frequencies should not be limited to below 300 GHz. In this book, we will consider all the RFs in the electromagnetic (EM) spectrum up to terahertz (THz) as RF — in other words, we view the RF range as including all frequencies from 3 KHz to microwave, millimeter-wave and sub-millimeter-wave frequencies. Therefore, RF, as it is practiced or should be practiced now, is not different from microwave, millimeter-wave and sub-millimeter-wave frequencies. The boundary between RF and microwave, millimeter-wave and sub-millimeter-wave indeed no longer exists or should not exist. As the technologies for RFICs advance toward the terahertz region of the RF spectrum, it is expected that RFICs will find many useful applications in both the commercial and the defense sectors at terahertz — for instance, medical imaging or personal-health

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

monitoring in the medical field and extremely wide bandwidth and ultrahigh data-rate for wireless communications.

Over the past several decades, RF components and systems in the microwave, millimeter-wave and sub-millimeter-wave ranges have been dominated with circuits employing III–V compound semiconductor devices, such as GaAs metal semiconductor field effect transistor (MESFET), high electron mobility transistor (HEMT), InP HEMT, GaAs heterojunction bipolar transistor (HBT), and InP HBT, etc. due to their superior performance as compared to Si-based technologies. RFICs based on III–V semiconductors, however, are expensive and have limited integration capability for single-ship systems. Exploding demand for low cost, low power, compactness, and high integration ability, far exceeding those provided by the III–V semiconductor technologies, has led the wireless industry to focus on better Si-based technologies capable of operating in the RF range. Si-based RF technologies have advanced significantly during the past few decades and are increasingly important for wireless communications, sensing, and networking due to their low cost, low power, and excellent integration ability, notably CMOS, that facilitates various applications requiring miniature, low cost, low power systems with high volume throughput. Presently, Si-based technologies can offer good performance up to the millimeter-wave regime with much lower cost and better integration capability than their non-silicon based counterparts, hence opening up many opportunities in wireless communications, sensing and networking.

Various RFICs, including single components and single-chip subsystems and systems, have been successfully developed with good performance up to millimeter-wave frequencies, demonstrating the potential capability of RFICs and their possible applications in the higher end of the RF spectrum. As an example, Figure 1.1 shows the schematic and microphotograph of a single-chip millimeter-wave RFIC transmitter operating concurrently in two frequency bands at 24.5 ± 0.5 GHz and 35 ± 0.5 GHz using a 0.18-μm SiGe BiCMOS process, showing the integration of several RF components. RFICs have played a significant role in advancing the state of the art of RF circuits and systems for various applications from sensing and imaging to communications across a few hundred MHz to millimeter-wave frequencies, and potentially beyond. It is foreseen that RFICs and systems with 3D (vertical and horizontal) integrations can perform at very high frequencies in the RF range in the future. RFIC is now inevitable in RF systems, and it is expected that it will dominate the RF territory, particularly for commercial applications, just like III–V semiconductors based MMICs have done, but with much lower costs and better abilities for direct integration with digital ICs. Although the performance of many Si-based RFICs still presently does not match that of RFICs implemented using III–V compound semiconductor devices, particularly in the higher frequency end of the RF spectrum such as millimeter-wave frequencies, due to lower f_T and $f_{\max}$, higher substrate loss, more noise, and so on of current CMOS/BiCMOS devices, they have lower cost and better abilities for direct integration with digital ICs (and hence better potential for complete system-on-chip). RFICs are also small and low power, making them suitable for battery-operated wireless communication, sensing and networking devices and systems. RFICs are thus attractive for systems and, in fact, the principal choice for commercial wireless markets.

Typical RFIC design based on traditional analog design approach is not very suitable at high frequencies of the RF range that is currently practiced. As the RF spectrum moves toward the multi-GHz realm, the need of incorporating microwave design techniques into analog circuits and systems becomes increasingly important and is, in fact, inevitable. Consequently, the knowledge of EM and microwave engineering becomes vital for RFIC engineers in order to understand and design RFICs properly. This is a fact that is recognized by RF researchers and engineers in both academia and industry. High frequencies in the RF range, especially those approaching the frequency limits of CMOS/BiCMOS technology, make RFIC design challenging. The design of RFICs at high frequencies poses further challenging as circuits and devices become extremely small and the interactions between elements within a circuit or between circuits in an integrated system become so immense. Typical RFICs, especially those at low frequencies, use exclusively lumped elements. While lumped elements are useful for RF circuitry and, in some cases, mandatory (e.g., resistive terminations, bias bypass capacitors), it is difficult to realize a truly lumped element in lossy silicon substrates at high frequencies because of significant parasitics to ground associated with Si substrates and high frequency EM effects. At these frequencies, the need of incorporating transmission lines, distributed elements (e.g., transmission-line components) and microwave design techniques, besides lumped elements and (low frequency) analog design

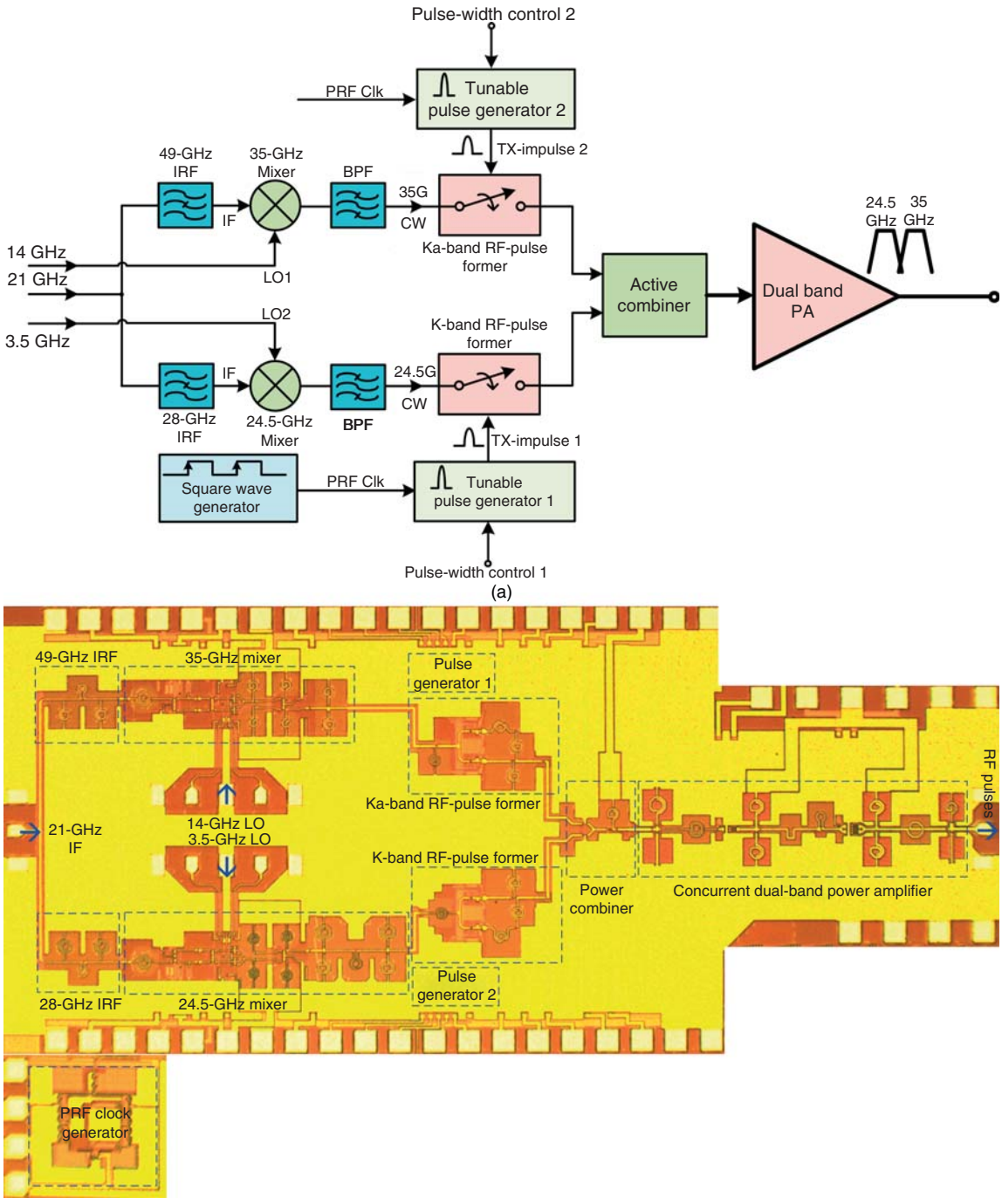


Figure 1.1. Schematic (a) and microphotograph (b) of a single-chip 0.18-μm SiGe BiCMOS millimeter-wave transmitter operating concurrently at two bands around 24.5 and 35 GHz. IRF: image-reject filter; BPF: band-pass filter; PA: power amplifier; TX: transmitter; RX: receiver; PRF: pulse repetition frequency; Clk: clock.

techniques, into RFICs becomes essential. Furthermore, besides circuit simulation, EM simulation needs to be effectively utilized to accurately model all effects occurring at these high frequencies. In view of these, it is crucial that the design of RFIC needs to be approached from the microwave design point of view. The design of Si-based RFICs is in general similar to the design of GaAs MMICs; the main difference is the use

of Si instead of GaAs as the processing means. In other words, the design of RFICs is essentially executed using the microwave design principles in Si-based "analog" environment.

This book revolves around the philosophy that RFIC engineers need knowledge in EM and microwave engineering, passive RFICs, active RFICs, RFIC analysis and design techniques, and RF systems. To that end, the book is aimed to address the theory, analysis, and design of passive and active RFICs using Si-based CMOS and BiCMOS technologies, in particular, and other non-silicon based technologies, in general, at high frequencies beyond those in the traditionally considered RF range. It intends to provide a comprehensive coverage for RFICs from passive to active circuits with particular emphasis on using microwave analysis and design techniques, which distinguishes itself from other RFIC books. It attempts to present the materials in details with a self-contained concept to allow graduate students and engineers with basic knowledge in RF and circuits to understand RFICs and their design. The book also includes problems for each chapter so readers can reinforce and practice their knowledge. Some of the problems are rather difficult and time-consuming and they can also be used as class projects for students. An important remark, yet may be redundant to RF engineers, is that many RF applications and systems are generally based on the same fundamentals and similar RF components. Knowledge of an RF system (e.g., pulsed system) and its RF components (e.g., mixer) for one application (e.g., sensing) can be used for the design of other systems (e.g., frequency-modulated continuous wave (FMCW) system) and for other applications (e.g., wireless communications.)

The book is organized into 16 chapters blending analog and microwave engineering with particular emphasis on the microwave engineering approach for RFICs, which is essential but not implemented in typical RFIC books. Chapter 2 provides the fundamentals of EM theory needed for RF engineers to understand basic yet relevant EM principles and effects on RFICs. Chapter 3 covers the design and analysis of on-chip lumped elements typically used in RFICs, including inductors, capacitors and resistors. Chapter 4 discusses the fundamentals of transmission lines for both single and multiconductor transmission lines including transmission-line equations and important transmission-line parameters, as well as synthetic transmission lines and commonly used printed-circuit transmission lines. Chapter 5 covers the analysis and design of both lumped-element and distributed resonators. Chapter 6 presents some fundamental design techniques for impedance-matching networks. Chapter 7 presents the formulation and characteristics of the scattering parameters as well as important parameters related to them. Chapter 8 presents the analysis and design of various basic RF passive components including directional couplers, hybrids, power dividers, and filters. Chapter 9 provides the fundamentals of CMOS transistors that are useful for the design of CMOS RFICs. Chapter 10 presents an analysis of stability for RFICs employing transistors. Chapter 11 covers the fundamentals and design of RF amplifiers, low noise amplifiers, power amplifiers (PAs), balanced amplifiers, and broad-band amplifiers. Chapter 12 discusses the fundamentals of oscillators, the theory of phase noise, and the design of both single-ended and balanced oscillators for RFICs. Chapter 13 presents the fundamentals of mixers, their topologies, analysis, and design for RFICs. Chapter 14 discusses the fundamentals and analyses of switches, and the design of SPST (single pole single throw) and T/R switches for RFICs. It also addresses ultra-wideband distributed switches, ultrahigh isolation switches, and switches implementing filtering functions. Chapter 15 presents the simulation, layout, and measurement for RFICs, as well as the calibration and de-embedding for on-wafer measurement. Chapter 16 addresses the commonly used pulsed and FMCW systems along with the widely used receiver architectures of homodyne and super-heterodyne as a way to introduce RF systems. Finally, the Appendix presents the design of an RFIC double-balanced mixer based on the Gilbert cell as an example to illustrate the design process of RFICs.

It is particularly noted that, in this book, for the sake of simplicity, we will use the term CMOS RFIC often but this, by no means, implies that the book only addresses CMOS RFICs. The design of other Si-based RFICs such as BiCMOS RFICs (and in fact the design of other non-silicon RFICs like GaAs MMICs as stated earlier) is equally applicable.

PROBLEMS

The objective of these problems is to familiarize readers with some of the current and potential applications/systems of RFICs and systems.

1.1 Search the 802.11b wireless Local Area Network (WLAN) applications. Describe the IEEE 802.11b standards, some systems currently employed and components used in these systems, their applications, performance, operating frequencies, and CMOS/BiCMOS technologies used, etc.

1.2 Repeat Problem 1.1 for un-licensed ultra-wideband (UWB) applications from 3.1 to 10.6 GHz.

1.3 Repeat Problem 1.1 for Bluetooth.

1.4 Repeat Problem 1.1 for millimeter-wave (MMW) radio applications including 60 GHz and E-band (71–76 GHz and 81–86 GHz bands).

1.5 Search for current Si-based CMOS and BiCMOS processes and compile on a table the following: (i) device technology (i.e., 30, 45, 90, 130, 180, and 250 nm), (ii) f_T (the cut-off frequency or the frequency of unity gain), (iii) $f_{\max}$ (the maximum frequency of oscillation or the frequency at which the maximum available gain is 0 dB), (iv) foundry, and (v) other pertinent information.

1.6 Describe current and potential applications of Si-based RFICs and systems (from microwave to millimeter wave frequencies). What do you think are the future trends and applications and at what frequencies?

CHAPTER 2

FUNDAMENTALS OF ELECTROMAGNETICS

Radio-frequency integrated circuits (RFICs) involve high frequencies that cause electromagnetic (EM) or high frequency effects to circuit performance which, if not properly accounted for, can disrupt or even ruin the performance, particularly at frequencies in the high end of the radio-frequency (RF) spectrum. EM therefore plays a crucial role in the RFIC design. It influences not only the circuit analysis and simulation, but also the selection or derivation of circuit topologies and schematics as well as the circuit layout. Knowledge of EM and what EM can do to improve or inadvertently degrade the performance of RFIC is thus absolutely essential for RF engineers. This implies that well-rounded RF engineers should acquire sufficient education in basic and advanced EM. In this chapter, we will present the fundamentals of EM which, although are relatively basic, would help RF engineers to understand some of the EM effects on RFIC, if properly interpreted, and/or to acquire further EM information that is relevant for RFIC.

2.1 EM FIELD PARAMETERS

The EM fields are separated into three cases: time-varying case for fields changing in any time fashion or, loosely speaking, fields in the time domain; sinusoidal time-varying case for fields varying sinusoidally, loosely defined as fields in the frequency domain; and static or DC case for fields independent of time or frequency. The following field parameters and notations[1] will be used in this chapter:

	Time Varying	Sinusoidal Time Varying (Phasor)	Static
Electric field intensity, V/m	$\boldsymbol{E}(x,y,z,t)$	$\hat{\boldsymbol{E}}(x,y,z)$	$\overline{\boldsymbol{E}}(x,y,z)$
Magnetic field intensity, A/m	$\boldsymbol{H}(x,y,z,t)$	$\hat{\boldsymbol{H}}(x,y,z)$	$\overline{\boldsymbol{H}}(x,y,z)$
Electric flux density, C/m^2	$\boldsymbol{D}(x,y,z,t)$	$\hat{\boldsymbol{D}}(x,y,z)$	$\overline{\boldsymbol{D}}(x,y,z)$
Magnetic flux density, Tesla (T)	$\boldsymbol{B}(x,y,z,t)$	$\hat{\boldsymbol{B}}(x,y,z)$	$\overline{\boldsymbol{B}}(x,y,z)$
Current density, A/m^2	$\boldsymbol{J}(x,y,z,t)$	$\hat{\boldsymbol{J}}(x,y,z)$	$\overline{\boldsymbol{J}}(x,y,z)$
Volume charge density, C/m^3	$\rho(x,y,z,t)$	$\rho(x,y,z)$	$\rho(x,y,z)$

[1]Throughout this chapter, vectors are written in boldface.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

The charge (ρ) and current ($\boldsymbol{J}, \hat{\boldsymbol{J}}, \overline{\boldsymbol{J}}$) density are for free charges and free currents, respectively.

In the above notations, (x, y, z) in the rectangular coordinate system is assumed for all the parameters. It is noted that any electrical parameters including the above field parameters are also a function of frequency; however, since time and frequency are related, frequency is not included in the independent parameters for simplification. The electric flux density is also known as the displacement field vector, and the current and charge density are the electric density. We also recognize that signals or waves and parameters associated with them such as power, electric field, and magnetic field all have magnitude and direction, and hence they are described as vectors, enabling us to know their strength and direction. For instance, power, as we know, is given by $\boldsymbol{P} = \boldsymbol{E} \times \boldsymbol{H}$ which shows precisely the magnitude and direction of a signal traveling in a medium such as air or propagating in an electrical circuit such as RFIC. This equation describes exactly the nature of the propagation of signals and is much more powerful than the conventional power equation $P = VI$ from the circuit theory, where V and I are voltage and current, respectively, which does not show the propagation of signals.

2.2 MAXWELL'S EQUATIONS

Maxwell's equations form the foundation for EM in particular and electrical engineering in general and are given in the following differential and integral forms:

	Differential (or Point) Form		Integral Form	
Gauss's law	$\nabla.\boldsymbol{D} = \rho$	(2.1a)	$\iint_S \boldsymbol{D}.\boldsymbol{ds} = \iiint_V \rho dv$	(2.2a)
	$\nabla.\boldsymbol{B} = 0$	(2.1b)	$\iint_S \boldsymbol{B}.\boldsymbol{ds} = 0$	(2.2b)
Faraday's law	$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}$	(2.1c)	$\int_C \boldsymbol{E}.\boldsymbol{dl} = -\frac{\partial}{\partial t}\iint_S \boldsymbol{B}.\boldsymbol{ds}$	(2.2c)
Ampere's law	$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$	(2.1d)	$\int_C \boldsymbol{H}.\boldsymbol{dl} = \iint_S \boldsymbol{J}.\boldsymbol{ds} + \frac{\partial}{\partial t}\iint_S \boldsymbol{D}.\boldsymbol{ds}$	(2.2d)

where V is volume enclosed by surface S, $\boldsymbol{ds} = ds\,\boldsymbol{a}_n$ is differential surface vector with $\boldsymbol{a}_n$ being a unit vector perpendicular to surface ds and pointing away from the surface, $\boldsymbol{dl} = dl\,\boldsymbol{a}_l$ is differential length vector with $\boldsymbol{a}_l$ being a unit vector along $\boldsymbol{dl}$, and $\frac{\partial \boldsymbol{D}}{\partial t}$ represents the displacement current density ($\mathrm{A/m^2}$).

The differential and integral Maxwell's equations are related by the divergence theorem described by

$$\iint_S \boldsymbol{A}.\boldsymbol{ds} = \iiint_V \nabla.\boldsymbol{A}dv \tag{2.3}$$

where A is an arbitrary field vector, and Stokes's theorem given by

$$\int_C \boldsymbol{A}.\boldsymbol{dl} = \iint_S \nabla \times \boldsymbol{A}.\boldsymbol{ds} \tag{2.4}$$

Maxwell's equations can be simplified under special conditions as follows. Static fields ($f = 0$, $\partial/\partial t = 0$):

$$\nabla.\overline{\boldsymbol{D}} = \rho \quad \nabla \times \overline{\boldsymbol{E}} = 0 \quad \nabla.\overline{\boldsymbol{B}} = 0 \quad \nabla \times \overline{\boldsymbol{H}} = \overline{\boldsymbol{J}} \tag{2.5}$$

Steady-state sinusoidal time-varying fields:

$$\nabla.\hat{\boldsymbol{D}} = \rho \quad \nabla \times \hat{\boldsymbol{E}} = -j\omega\hat{\boldsymbol{B}} \quad \nabla.\hat{\boldsymbol{B}} = 0 \quad \nabla \times \hat{\boldsymbol{H}} = \hat{\boldsymbol{J}} + j\omega\hat{\boldsymbol{D}} \tag{2.6}$$

Maxwell's equations along with the following auxiliary relations enable many equations to be derived, from which not only many electrical phenomena and problems from DC to high frequencies can be explained and solved, but also many applications can be evolved.

2.3 AUXILIARY RELATIONS

2.3.1 Constitutive Relations

The constitutive relations describe the properties of materials. The electric flux density and electric field intensity in a material are related by

$$\boldsymbol{D} = \varepsilon_o\boldsymbol{E} + \boldsymbol{P} = \varepsilon_o\boldsymbol{E} + \varepsilon_o\chi_e\boldsymbol{E} = \varepsilon_o(1 + \chi_e)\boldsymbol{E} \tag{2.7}$$

where $\varepsilon_o \simeq 8.854 \times 10^{-12}$ F/m is the dielectric constant or permittivity of free space, $\boldsymbol{P} = \varepsilon_o\chi_e\boldsymbol{E}$ is the electric polarization vector of the material assuming the material's electrical properties are linear and isotropic, and χ_e is the (dimensionless) electric susceptibility of the material. Equation (2.7) can rewritten as

$$\boldsymbol{D} = \varepsilon_o\varepsilon_r\boldsymbol{E} = \varepsilon\boldsymbol{E} \tag{2.8}$$

where ε_r and ε are the relative dielectric constant (or relative permittivity) and dielectric constant (or permittivity) of material, respectively. ε_r is different for different materials and is normally considered the most important parameter characterizing (nonmagnetic) materials such as dielectric layers or substrates in RFIC. Note that air has $\varepsilon_r = 1.00059$ and is therefore typically used in place of free space.

Similarly, the magnetic flux density and magnetic field intensity in material are related by

$$\boldsymbol{B} = \mu_o\boldsymbol{H} + \boldsymbol{M} = \mu_o\boldsymbol{H} + \mu_o\chi_m\boldsymbol{H} = \mu_o(1 + \chi_m)\boldsymbol{H} \tag{2.9}$$

where $\mu_o = 4\pi \times 10^{-7}$ H/m is the permeability of free space, $\boldsymbol{M} = \mu_o\chi_m\boldsymbol{H}$ is the magnetic polarization vector of the material assuming the magnetic properties of the material are linear and isotropic, and χ_m is the (dimensionless) magnetic susceptibility of the material. Equation (2.9) can rewritten as

$$\boldsymbol{B} = \mu_o\mu_r\boldsymbol{H} = \mu\boldsymbol{H} \tag{2.10}$$

where μ_r and μ are the relative permittivity and permittivity of material, respectively. Most materials used for RFIC such as SiO_2 or Si (and in fact materials used in most of electrical circuits) are nonmagnetic and so have μ_r close to 1.

In general ε_r and μ_r, and hence ε and μ, are function of location (in the material), frequency, and signal, and hence the electric and magnetic fields, applied to the material. If these parameters are not a function of location in a material, the material is called homogeneous. A linear material is characterized by ε_r and μ_r not dependent on the strength of the applied signals or electric and magnetic fields. A material is called isotropic if ε_r and μ_r are independent of the direction of the applied signals. A material that is linear, homogeneous and isotropic is called a simple material. A simple material has constant ε_r and μ_r. Most materials used for circuits such as dielectrics and substrates used in RFIC are simple materials.

2.3.2 Current Relations

The most widely known current is the conduction current given as

$$\boldsymbol{J} = \sigma \boldsymbol{E} \tag{2.11}$$

where σ is the conductivity of the material. The current obtained from (2.11) is basically the current described in Ohm's law. The other less well-known current is the convection current whose density is described as

$$\boldsymbol{J} = \rho \boldsymbol{v} \tag{2.12}$$

where ρ is the (volume) charge density (C/m^3) in the material and $\boldsymbol{v}$ is the velocity of the charge carrier. As expected, the current density $\boldsymbol{J}$ in Maxwell's equations may consist of both conduction and convection currents. The convection current is typically neglected in most RF circuits. Another current is the displacement current density $\boldsymbol{J} = \partial \boldsymbol{D}/\partial t$ or $\boldsymbol{J} = j\omega\varepsilon \boldsymbol{E}$ (for sinusoidal time-varying steady state). This current is much smaller than the conduction current in good conductors even at RF and is usually neglected.

The conduction current and charge are related by the continuity or conservation-of-charge equation:

$$\nabla \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t} \tag{2.13}$$

for general time-varying case, and

$$\nabla \cdot \boldsymbol{J} = -j\omega\rho \tag{2.14}$$

under the sinusoidal time-varying steady state.

2.4 SINUSOIDAL TIME-VARYING STEADY STATE

Sinusoidal waveform is the most widely known and used signal type in electrical engineering. Various signal waveforms can be developed from sinusoidal waveforms. One of the most attractive features of sinusoidal waveforms with respect to analysis is their mathematical simplification resulting from the separation of the amplitude and phase of signals.

The electric field, or any field components, of a sinusoidal signal can be expressed as a sinusoidal expression (with reference to cosine):

$$\begin{aligned} \boldsymbol{E} &= E_x(x,y,z,t)\boldsymbol{a}_x + E_y(x,y,z,t)\boldsymbol{a}_y + E_z(x,y,z,t)\boldsymbol{a}_z = \widehat{\boldsymbol{E}}_o(x,y,z)\cos(\omega t + \theta) \\ &= \mathrm{Re}[\widehat{\boldsymbol{E}}(x,y,z)e^{j\omega t}] \end{aligned} \tag{2.15}$$

where $\widehat{\boldsymbol{E}}_o(x,y,z)$ and $\theta(x,y,z)$ represent the maximum amplitude and the phase of the electric field, respectively, and

$$\widehat{\boldsymbol{E}} \equiv \widehat{\boldsymbol{E}}(x,y,z) = \widehat{E}_x(x,y,z)\boldsymbol{a}_x + \widehat{E}_y(x,y,z)\boldsymbol{a}_y + \widehat{E}_z(x,y,z)\boldsymbol{a}_z = \widehat{\boldsymbol{E}}_o e^{j\theta} \tag{2.16}$$

is the phasor representation of the electric field intensity $\boldsymbol{E}$ in the time domain. $\widehat{\boldsymbol{E}}$ is called the electric field phasor, which is a vector independent of time, and represents the electric field in the frequency domain. Expanding (2.16) leads to the relationship between the components of the electric field in the time and frequency domains as

$$E_{x,y,z}(x,y,z,t) = \widehat{E}_{x,y,z}(x,y,z)\cos(\omega t + \theta_{x,y,z}) = \mathrm{Re}[\widehat{E}_{x,y,z}(x,y,z)e^{j\omega t}] \tag{2.17}$$

where $\theta_{x,y,z}(x,y,z)$ is the phase of $\widehat{E}_{x,y,z}(x,y,z)$, respectively.

Taking the derivative of (2.15) or (2.17) with respect to time (t) leads to the same electric field intensity with an additional term of $j\omega$, implying that $\partial/\partial t$ (in the time domain) for sinusoidal signals is equivalent to $j\omega$ (in the frequency domain). Maxwell's equations under general time variation (or the time domain) in (2.1c, d) can hence be transferred directly to the steady-state sinusoidal time variation (or the frequency domain) as described in (2.6) by replacing $\partial/\partial t$ with $j\omega$. Similarly, the time-domain continuity equation (2.13) becomes (2.14) in the frequency domain. It is noted that the constitutive relations, described in (2.8) and (2.10), for general time variation also hold for sinusoidal time variation.

As can be seen, Maxwell's equations in the frequency domain are simpler than those in the time domain and hence are preferred when the signal is a sinusoidal signal. It is reminded that the fields obtained from the frequency-domain Maxwell's equations and auxiliary relations are phasors. As these fields are time-varying sinusoidal fields, the phasors need to be multiplied by $e^{j\omega t}$ and the real parts of the results are taken to give the final time-dependent fields.

2.5 BOUNDARY CONDITIONS

We recall the Kirchhoff's current and voltage laws that every sophomore electrical engineering students know: the total current entering and leaving a particular node or the total voltages around a particular loop must be equal to zero. This theory actually set two conditions that the currents across components (such as resistors, capacitors, inductors) in branches and the voltages across components in a loop must follow in order for them to be valid. These rules set the fundamental for circuit solutions. Similar principle holds for EM and high frequency problems under the so-called boundary conditions described as follows.

To illustrate the significance of the boundary conditions, we consider a Lange coupler fabricated on a 0.25-μm complementary metal oxide silicon (CMOS) process, discussed in Section 8.2.2.3, as shown in Figure 2.1. This RFIC consists of conductors and dielectrics in different conductor and dielectric layers above a Si substrate, and can be decomposed into separate regions. In general, the electric and magnetic fields existing in the different regions of the RFIC, which affect the performance of the RFIC, can be determined utilizing Maxwell's equations and some auxiliary relations. Such fields obtained in each region are only valid in that region and can also be used in the same region in other RFIC that is different from the considered RFIC. In other words, these regional fields are common for all RFICs provided that the regions are identical. That means that these fields do not allow a unique solution to be obtained for the considered RFIC. To obtain a unique solution, the relations of these regional electric and magnetic fields at all the interfaces existing in the RFIC must be imposed. These relations are known as the boundary conditions.

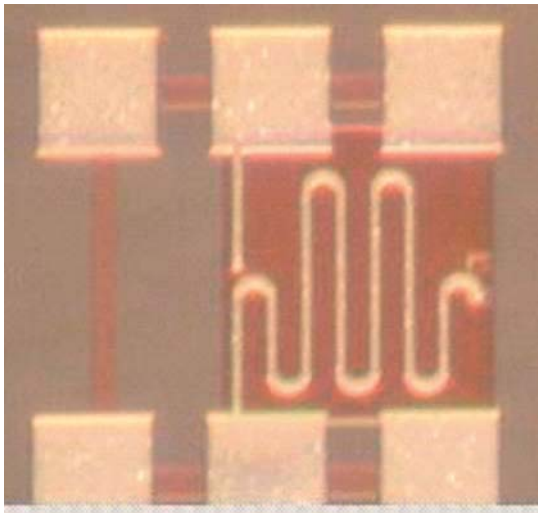


Figure 2.1. Photograph of a 0.25-μm CMOS Lange coupler.

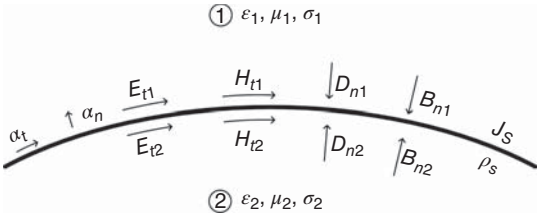


Figure 2.2. Boundary between two media. $\boldsymbol{a}_n$ is the unit vector perpendicular to the interface and pointing into medium 1; $\boldsymbol{J}_s$ is the (linear) surface current density at the interface (A/m); and ρ_s is the surface charge density at the interface (C/m^2).

Only with an enforcement of the boundary conditions, the calculated fields are valid, a unique solution can be obtained, and the calculated performance is correct.

2.5.1 General Boundary Conditions

We consider a boundary between two regions 1 and 2 consisting of any material characterized by the dielectric constant ε_1, permeability μ_1, conductor σ_1 and the dielectric constant ε_2, permeability μ_2, conductor σ_2, respectively, as shown in Figure 2.2. The electric and magnetic field, and the electric and magnetic flux density in each region are decomposed into components tangential (denoted by subscript t) and normal (denoted by subscript n) to the interface.[2]

The boundary conditions can be derived directly from Maxwell's equations and are given in the following equations:

$$\begin{aligned}
\boldsymbol{a}_n \times (\boldsymbol{E}_1 - \boldsymbol{E}_2) &= 0 \quad \text{or} \quad E_{t1} = E_{t2} \\
\boldsymbol{a}_n \cdot (\boldsymbol{D}_1 - \boldsymbol{D}_2) &= \rho_s \quad \text{or} \quad D_{n1} - D_{n2} = \rho_s \\
\boldsymbol{a}_n \times (\boldsymbol{H}_1 - \boldsymbol{H}_2) &= \boldsymbol{J}_s \quad \text{or} \quad H_{t1} - H_{t2} = \boldsymbol{J}_s \\
\boldsymbol{a}_n \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) &= 0 \quad \text{or} \quad B_{n1} = B_{n2}
\end{aligned} \tag{2.18}$$

These boundary conditions are the same for both time-varying and time-invariant fields, and are of course valid only at the interface. It is noted that, in each medium for each field, there are only one component normal but two components tangent to the boundary surface. For instance, considering Cartesian coordinates and assuming (x, y) is the boundary surface, then there are two tangential components along x and y, and one normal component along z. As a result, care needs to be exercised when using the scalar forms of the boundary conditions involving tangential components. Using the vector form of the boundary conditions avoids potential of missing one of the two tangential components.

2.5.2 Specific Boundary Conditions

The boundary conditions (2.18) can be simplified for special cases as follows. Boundary conditions between two perfect (or lossless) dielectrics ($\sigma_1 = \sigma_2 = 0$):

$$\begin{aligned}
\boldsymbol{a}_n \times (\boldsymbol{E}_1 - \boldsymbol{E}_2) &= 0 \quad \text{or} \quad E_{t1} = E_{t2} \\
\boldsymbol{a}_n \cdot (\boldsymbol{D}_1 - \boldsymbol{D}_2) &= 0 \quad \text{or} \quad D_{n1} = D_{n2} \\
\boldsymbol{a}_n \times (\boldsymbol{H}_1 - \boldsymbol{H}_2) &= 0 \quad \text{or} \quad H_{t1} = H_{t2} \\
\boldsymbol{a}_n \cdot (\boldsymbol{B}_1 - \boldsymbol{B}_2) &= 0 \quad \text{or} \quad B_{n1} = B_{n2}
\end{aligned} \tag{2.19}$$

[2]Only one tangential component is shown here. For instance, for the electric field $\boldsymbol{E}_1$, the tangential components designated by E_{t1} and E_{t2} (along direction $\boldsymbol{a}_t$) is shown and other tangential components along $\boldsymbol{a}'_t$, that is normal to $\boldsymbol{a}_t$, are not shown.

Boundary conditions between a perfect dielectric ($\sigma_1 = 0$) in medium 1 and a perfect conductor ($\sigma_2 = \infty$) in medium 2:

$$\begin{aligned} \boldsymbol{a}_n \times \boldsymbol{E} &= 0 \quad \text{or} \quad E_t = 0 \\ \boldsymbol{a}_n \cdot \boldsymbol{D} &= \rho_s \quad \text{or} \quad D_n = \rho_s \\ \boldsymbol{a}_n \times \boldsymbol{H} &= \boldsymbol{J}_s \quad \text{or} \quad H_t = J_s \\ \boldsymbol{a}_n \cdot \boldsymbol{B} &= 0 \quad \text{or} \quad B_n = 0 \end{aligned} \tag{2.20}$$

where $\boldsymbol{E} = \boldsymbol{E}_1$, $\boldsymbol{D} = \boldsymbol{D}_1$, $\boldsymbol{H} = \boldsymbol{H}_1$, and $\boldsymbol{B} = \boldsymbol{B}_1$. The first equation in (2.20) indicates that the tangential component of the electric field at the surface of a perfect conductor is always zero, which is a very important and well-known fact for RFIC design. This implies that the electric field at the surface a perfect conductor is always normal to that surface. The second and third equations show that the charge and current induced on the surface of a perfect conductor when fields are present can be obtained by calculating the normal electric flux density and tangential magnetic field at the surface, respectively.

2.6 WAVE EQUATIONS

Although the electric and magnetic fields in any medium can be determined from Maxwell's equations subject to boundary conditions, it is more convenient to determine them from single equations such as wave equations. To illustrate the formulation for wave equations, we consider a simple medium (homogeneous, isotropic, and linear medium) characterized by permittivity ε, permeability μ, and conductivity σ.

Taking the curl of (2.1c) and making use of (2.1d) with $\boldsymbol{B} = \mu\boldsymbol{H}$ gives

$$\nabla \times \nabla \times \boldsymbol{E} = -\frac{\partial}{\partial t}(\nabla \times \boldsymbol{B}) = -\mu \frac{\partial}{\partial t}\left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}\right) \tag{2.21}$$

Applying the vector identity $\nabla \times \nabla \times \boldsymbol{E} = \nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}$ to (2.21) and utilizing (2.1a) with $\boldsymbol{D} = \varepsilon \boldsymbol{E}$, we get

$$\nabla^2 \boldsymbol{E} - \mu\varepsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \mu \frac{\partial \boldsymbol{J}}{\partial t} - \frac{1}{\varepsilon}\nabla\rho = 0 \tag{2.22}$$

where ∇^2 is the Laplacian operator. In Cartesian coordinates, it is given as

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.23}$$

Similarly, we can also derive from (2.1d), (2.1c), and (2.1b):

$$\nabla^2 \boldsymbol{H} - \mu\varepsilon \frac{\partial^2 \boldsymbol{H}}{\partial t^2} + \nabla \times \boldsymbol{J} = 0 \tag{2.24}$$

Equations (2.22) and (2.24) are called the (inhomogeneous) wave equations for any time-varying fashion, which, along with the boundary conditions for a particular structure, can be solved to obtain unique electric and magnetic fields for that structure.

In a perfect dielectric medium ($\sigma = 0$), there is no free charge ($\rho = 0$), and accordingly the wave equations (2.22) and (2.24) are reduced to the (homogeneous) wave equations:

$$\begin{aligned} \nabla^2 \boldsymbol{E} - \mu\varepsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2} &= 0 \\ \nabla^2 \boldsymbol{H} - \mu\varepsilon \frac{\partial^2 \boldsymbol{H}}{\partial t^2} &= 0 \end{aligned} \tag{2.25}$$

If the fields are sinusoidal time varying, we can derive from Maxwell's equations (2.6), or directly from (2.22) and (2.24) recognizing the equivalence between $j\omega$ and $\partial/\partial t$, the following inhomogeneous wave equations:

$$\nabla^2\hat{\boldsymbol{E}} + k^2\hat{\boldsymbol{E}} - j\omega\mu\hat{\boldsymbol{J}} - \frac{1}{\varepsilon}\nabla\rho = 0$$
$$\nabla^2\hat{\boldsymbol{H}} + k^2\hat{\boldsymbol{H}} - \nabla\times\hat{\boldsymbol{J}} = 0 \tag{2.26}$$

where k is the wave number given as

$$k = \omega\sqrt{\varepsilon\mu} = \frac{\omega}{u} = \frac{2\pi}{\lambda} \tag{2.27}$$

with u being the (phase) velocity of the signal and λ being the wavelength.

For a perfect dielectric medium, the wave equations (2.26) are reduced to the (homogeneous) wave equations:

$$\nabla^2\hat{\boldsymbol{E}} + k^2\hat{\boldsymbol{E}} = 0$$
$$\nabla^2\hat{\boldsymbol{H}} + k^2\hat{\boldsymbol{H}} = 0 \tag{2.28}$$

which are also known as Helmholtz's equations. Expanding (2.28) in Cartesian coordinates then give six individual wave equations:

$$\frac{\partial^2\hat{E}_{x,y,z}}{\partial^2 x} + \frac{\partial^2\hat{E}_{x,y,z}}{\partial^2 y} + \frac{\partial^2 E_{x,y,z}}{\partial^2 z} + k^2\hat{E}_{x,y,z} = 0$$
$$\frac{\partial^2\hat{H}_{x,y,z}}{\partial^2 x} + \frac{\partial^2\hat{H}_{x,y,z}}{\partial^2 y} + \frac{\partial^2\hat{H}_{x,y,z}}{\partial^2 z} + k^2\hat{H}_{x,y,z} = 0 \tag{2.29}$$

which can be solved for individual field components. For instance, the electric field component along the x direction is

$$\hat{E}_x = \hat{E}_o^+ e^{-jkz} + \hat{E}_o^- e^{-jkz} \tag{2.30}$$

where $\hat{E}_o^+$ and $\hat{E}_o^-$ represent the magnitude of the x-directed electric field of the forward- and backward-traveling signals, respectively, and $\gamma = jk = j\beta$ is the propagation constant with β being the phase constant (rad/m). Note that Eqs. (2.28) and (2.29), as well as their solutions such as (2.30), although derived for a perfect dielectric medium, are also applicable for an imperfect or lossy dielectric medium with the propagation constant given as $\gamma = jk = \alpha + j\beta$, where α is the attenuation constant (neper/m or Np/m).

2.7 POWER

The power density (W/m^2) of a signal is given by the Poynting vector

$$\boldsymbol{P} = \boldsymbol{E}\times\boldsymbol{H} \tag{2.31}$$

This is the power equation for general time-varying fields and is the instantaneous power density that indicates the power flow per unit area (W/m^2). The total power flow across a surface S can then be obtained as

$$\boldsymbol{P}_T = \oint_S \boldsymbol{P}\cdot\boldsymbol{ds} = \oint_S (\boldsymbol{E}\times\boldsymbol{H})\cdot\boldsymbol{ds} \tag{2.32}$$

Equations (2.32) and (2.31) are equivalent to the conventional power equation $p(t) = v(t)i(t)$, where $v(t)$ and $i(t)$ are voltage and current, respectively, that we are familiar with in circuit theory, yet it is much more powerful in which it gives precisely the magnitude as well the direction of power or signal flow.

For sinusoidal signals, the time-average power density (W/m^2) is given as

$$\boldsymbol{P}_{\text{avg}} = \frac{1}{2}\text{Re}(\hat{\boldsymbol{E}} \times \hat{\boldsymbol{H}}^*) \tag{2.33}$$

The total time-average power flow across a surface S can be obtained as

$$\boldsymbol{P}_{T,\text{avg}} = \frac{1}{2}\text{Re}\int_S (\hat{\boldsymbol{E}} \times \hat{\boldsymbol{H}}^*) \cdot \boldsymbol{ds} \tag{2.34}$$

Equations (2.33) and (2.34) are equivalent to the conventional power equation $P = \frac{1}{2}\text{Re}(VI^*)$, where V and I are the voltage and current phasors, respectively, that we are familiar with in circuit theory. These equations allow the magnitude and direction of the real power flow in any medium to be determined.

Determination of the direction of signal propagation in a structure such as that in RFIC using the average power equation in (2.33) or (2.34) is very important to RF engineers. For instance, we consider an RFIC employing microstrip lines for signal transmission. We know from the analysis of microstrip lines that most of the electric and magnetic fields of a signal traveling along the lines are in the transverse plane to the conducting strip, with the electric field perpendicular and the magnetic field parallel to the conducting strip, while only a small portion of these fields are directed along the conducting strip. Applying (2.33) shows that most of the power would then flow along the conducting strip and only a small amount of the power flows normally to the strip representing the unwanted radiated power. Using the circuit equation $P = \frac{1}{2}\text{Re}(VI^*)$ does not provide such crucial information.

2.8 LOSS AND PROPAGATION CONSTANT IN MEDIUM

Dielectrics and substrates used for electrical circuits including RFIC are imperfect. Consequently, there is always loss present in any practical dielectrics and substrates, known as dielectric loss, due to a nonzero conductivity of the medium. We begin the formulation by stating that, as the properties of a dielectric medium is independent of the waveform type and sources (for simple medium), we can assume that a signal is sinusoidal and there is no charge ($\rho = 0$) in the medium in deriving the dielectric properties.

For a perfect or lossless simple dielectric medium ($\sigma = 0$), no free charge exists ($\rho = 0$), and the Ampere's equation in Maxwell's equations (2.6) can be written as

$$\nabla \times \hat{\boldsymbol{H}} = j\omega\varepsilon\hat{\boldsymbol{E}} \tag{2.35}$$

For a lossy dielectric medium ($\sigma \neq 0$), we can write from (2.6):

$$\nabla \times \hat{\boldsymbol{H}} = \sigma\hat{\boldsymbol{E}} + j\omega\varepsilon\hat{\boldsymbol{E}} = j\omega\left(\varepsilon - j\frac{\sigma}{\omega}\right)\hat{\boldsymbol{E}} \tag{2.36}$$

which is identical in form to (2.35), except that ε in (2.35) is replaced by

$$\varepsilon_c = \varepsilon - j\frac{\sigma}{\omega} \equiv \varepsilon' - j\varepsilon'' \tag{2.37}$$

where

$$\begin{aligned} \varepsilon' &= \varepsilon = \varepsilon_o\varepsilon_r \\ \varepsilon'' &= \frac{\sigma}{\omega} \end{aligned} \tag{2.38}$$

ε_c is called the complex dielectric constant or complex permittivity of the medium. Both ε' and ε'' are function of frequency, and ε'' accounts for the loss in the medium. We can also characterize a lossy dielectric medium by its complex relative dielectric constant

$$\varepsilon_{cr} \equiv \frac{\varepsilon_c}{\varepsilon_o} = \varepsilon_r - j\frac{\sigma}{\omega} = \varepsilon_r' - j\varepsilon_r'' \tag{2.39}$$

Note that $\varepsilon_r' = \varepsilon_r$.

The loss in a dielectric medium is typically characterized in terms of the loss tangent defined as the ratio between the imaginary and real parts of the complex dielectric constant

$$\tan\delta \equiv \frac{\varepsilon''}{\varepsilon'} = \frac{\varepsilon_r''}{\varepsilon_r'} = \frac{\sigma}{\omega\varepsilon} = \frac{\sigma}{\omega\varepsilon_o\varepsilon_r} \tag{2.40}$$

which is of course dependent upon frequency. The loss tangent of a dielectric, just like its relative dielectric constant ε_r and complex relative dielectric constant ε_{cr}, can be measured.

We know that signals propagating in a lossy dielectric medium according to the propagation constant $\gamma = jk = \alpha + j\beta$, where k is the complex wave number given by $k = \omega\sqrt{\mu\varepsilon_c}$ by making use of (2.27). The propagation constant is thus obtained, utilizing (2.37), as

$$\gamma = \alpha + j\beta = j\omega\sqrt{\mu\varepsilon_c} = j\omega\sqrt{\mu\varepsilon}\sqrt{1 - j\frac{\sigma}{\omega\varepsilon}} = j\omega\sqrt{\mu\varepsilon'}\sqrt{1 - j\frac{\varepsilon''}{\varepsilon'}} \tag{2.41}$$

or, applying (2.39),

$$\gamma = jk_o\sqrt{\varepsilon_r' - j\varepsilon_r''} = jk_o\sqrt{\varepsilon_r'}\sqrt{1 - j\frac{\varepsilon''}{\varepsilon_r'}} \tag{2.42}$$

where $k_o = \omega\sqrt{\mu_o\varepsilon_o}$ is the wave number for free space.

For good dielectrics having small loss, $\sigma/\omega\varepsilon \ll 1$ or $\varepsilon_r'' \ll \varepsilon_r'$ and hence $\tan\delta \ll 1$, and the propagation constant can be approximated using the binomial series as

$$\gamma = \alpha + j\beta \simeq \frac{k_o\varepsilon_r''}{2\sqrt{\varepsilon_r'}} + jk_o\sqrt{\varepsilon_r'} \tag{2.43}$$

Equation (2.41) can also be used for good conductors. Imposing the condition $\sigma \gg \omega\varepsilon'$ for good conductors upon (2.41) results in

$$\gamma = \alpha + j\beta \simeq j\omega\sqrt{\mu\varepsilon}\sqrt{j\frac{\sigma}{\omega\varepsilon}} = \sqrt{j}\sqrt{\omega\mu\sigma} \tag{2.44}$$

Substituting $\sqrt{j} = \cos(\pi/4) + j\sin(\pi/4) = (1 + j)/\sqrt{2}$ into (2.44) leads to

$$\gamma = \alpha + j\beta \simeq (1 + j)\sqrt{\pi f\mu\sigma} \tag{2.45}$$

from which the attenuation and phase constant for good conductors are obtained as

$$\alpha = \beta \simeq \sqrt{\pi f\mu\sigma} \tag{2.46}$$

2.9 SKIN DEPTH

We consider a sinusoidal uniform plane wave[3] incident normally onto a lossy medium, representing an imperfect dielectric or conductor, as shown in Figure 2.3. To simplify the formulation without loss of generality, we represent the signal using only the electric field component along the x direction.

As the signal propagates into the medium along the direction z, the amplitude of the electric field (along x) according to $\hat{\boldsymbol{E}} = a_x \hat{E}_o e^{-\alpha z} e^{-j\beta z}$ is reduced according to the attenuation factor $e^{-\alpha z}$. The magnitude of the electric field at $z = 1/\alpha$ is obtained, making use of the Euler's constant $e = 2.718$, as

$$\hat{E}\left(z = \frac{1}{\alpha}\right) = \hat{E}_o e^{-1} = 0.368\hat{E}_o \tag{2.47}$$

which indicates that the field magnitude is reduced by about 63% over a distance of

$$\delta_s = \frac{1}{\alpha} \tag{2.48}$$

This distance is referred to as the "skin depth" or "depth of penetration" of materials including dielectrics and conductors. As the properties of a material remain the same for any time-varying waveform, the skin depth derived in (2.48) is valid for any time-varying signals. The magnitude of the electric field along the x direction is then dependent on the skin depth as

$$\hat{E}_x = \hat{E}_o e^{-z/\delta_s} \tag{2.49}$$

Skin depth is one of the most important electrical parameters of materials. It is crucial not only for conductors but also for dielectrics since it indicates that a propagating signal is concentrated within a few skin depths of a medium, dictates how a signal is attenuated while it traverses a medium, and causes some effects to circuit performance such as coupling in lossy substrates such as Si or internal inductance for transmission lines that we discuss in Section 4.4. Particularly, for highly lossy substrates such as Si, the skin depth is small, indicating signals cannot propagate far into the substrates.

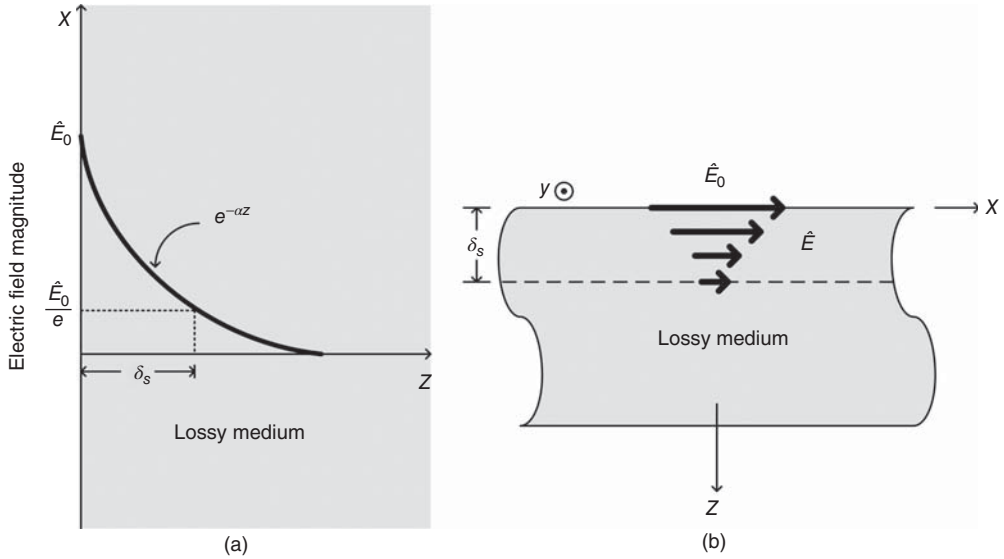


Figure 2.3. (a, b) Behavior of the electric field of a signal incident normally onto a medium. $\hat{E}_o$ is the (initial) amplitude at the medium's surface. The length of the arrows in (b) shows the relative magnitude.

[3] A uniform plane wave propagating in the z direction has its electric and magnetic fields independent of x and y.

As an example for skin-depth derivation, we consider a good conductor having permeability μ (normally equal to μ_o) and conductivity σ. The attenuation constant is given in (2.46) which, upon substituting in to (2.48), gives the skin depth of (good) conductors as

$$\delta_s \simeq \sqrt{\frac{2}{\omega\mu\sigma}} \tag{2.50}$$

which becomes zero for perfect conductors. As mentioned in Section 4.4, for a given conductor, as the frequency is increased, the skin depth reduces leading to increased resistance and attenuation, and hence loss as expected.

The density of the current in the conductor corresponding to the x-directed electric field is obtained as

$$\hat{\boldsymbol{J}}_x = \sigma\hat{\boldsymbol{E}}_x = \boldsymbol{a}_x\hat{J}_o e^{-z/\delta_s}e^{-jz/\delta_s} \tag{2.51}$$

where $\hat{J}_o = \sigma\hat{E}_o$ is the amplitude of the current density at the surface ($z = 0$) of the conductor. We can see that the electric and magnetic fields (or in turn the signal) and current are concentrated within a few skin depths of conductors. In the limit when the conductivity approaches infinity (perfect conductor), the current becomes a "true" surface current.

2.10 SURFACE IMPEDANCE

We consider an imperfect conductor. As a signal traverses the conductor, its electric field and current penetrate into the conductor as shown in Figure 2.4, effectively giving rise to internal impedance called "surface impedance." Due to the skin depth of the conductor, the electric field and current diminish quickly inside the conductor, as illustrated in Figure 2.4.

Assume the conductor is good, the density of the current along the x direction is obtained from (2.51) as

$$\hat{J}_x = \hat{J}_o e^{-(1+j)z/\delta_s} \tag{2.52}$$

Assume the conductor dimension is infinite in the z direction (infinite depth), the total x-directed current flow in the conductor per unit width (in the y direction) is obtained as

$$\begin{aligned}\hat{I}_x &= \int_0^\infty \hat{J}_x dz = \int_0^\infty \hat{J}_o e^{-(1+j)z/\delta_s} dz \\ &= \frac{\hat{J}_o\delta_s}{1+j}\end{aligned} \tag{2.53}$$

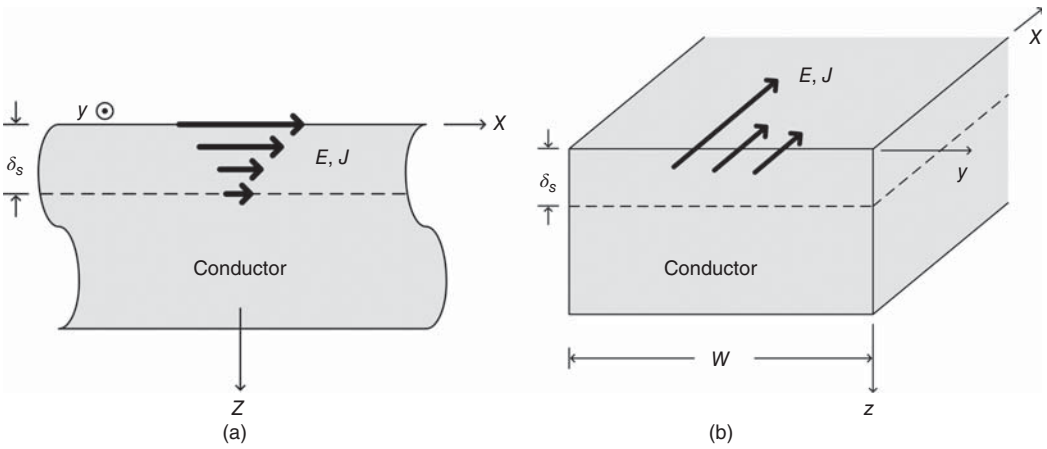


Figure 2.4. (a, b) Behavior of the electric field and current in an imperfect conductor.

The surface impedance per unit length (along z) and per unit width (along y) is defined as

$$Z_S = R_s + j\omega L_s \equiv \frac{\widehat{E}_x(z=0)}{\widehat{I}_x} \tag{2.54}$$

which becomes, after using (2.53) and substituting $\widehat{J}_o = \sigma \widehat{E}_o$:

$$Z_S = \frac{1+j}{\sigma \delta_s} \tag{2.55}$$

from which we obtain, upon using (2.50), the surface resistance per unit length

$$R_s = \frac{1}{\sigma \delta_s} = \sqrt{\frac{\omega \mu}{2\sigma}} \tag{2.56}$$

and the surface inductance per unit length

$$L_s = \frac{R_s}{\omega} = \sqrt{\frac{\mu}{2\sigma\omega}} \tag{2.57}$$

of the conductor. The surface inductance gives rise to the total inductance per unit width (along y) of the conductor through the internal inductance per unit width (along y)[4] L_i discussed in Section 4.4. Under the assumption that the current is distributed uniformly within a skin depth, we can derive

$$L_s = WL_i \tag{2.58}$$

since the conductor width W can be divided into W unit-width elements that are essentially in parallel.

The surface impedance of perfect conductors, as obtained from (2.55), is zero as expected. We also note based on (2.50) and (2.55) that, as the frequency is reduced, the skin depth increases and the surface impedance reduces, and in the limit of DC, it becomes zero as expected. On the other hand, as the frequency is increased, the skin depth reduces, and the surface impedance increases. The existence of the surface impedance, and in turn the surface resistance and surface inductance, and their escalation at high frequencies cause potential problems in RFICs which might disrupt the circuit performance if not properly taken care for. One example is a change in the characteristic impedance of transmission lines at high frequencies due to the surface impedance of the transmission line's conductors. Another example is the metal pad typically used for RFIC grounding exhibits surface impedance, which prevents a perfect (or, to a lesser extent, a good) ground connection, even a circuit point (to be grounded) is connected directly to the metal pad without interconnect. It is also further noted that the electric ($\boldsymbol{E}_t$) tangent to the surface of a conductor and the surface current ($\boldsymbol{J}_s$) are related by the surface impedance as

$$\boldsymbol{E}_t = Z_S \boldsymbol{J}_s \tag{2.59}$$

where

$$\boldsymbol{J}_s = \boldsymbol{a}_n \times \boldsymbol{H} \tag{2.60}$$

which, upon substituting into (2.59), gives the relation between the tangential electric and magnetic ($\boldsymbol{H}_t$) fields as

$$\boldsymbol{E}_t = Z_S \boldsymbol{H}_t \tag{2.61}$$

We can then see that finite surface impedance of an imperfect conductor gives rise to the tangential electric field, which is otherwise zero on a perfect conductor.

[4]The width in here is actually the length in Section 4.4, so L_i is called the inductance per unit length in Section 4.4.

PROBLEMS

2.1 Derive the integral Maxwell's equations from the differential Maxwell's equations.

2.2 Derive Eq. (2.17).

2.3 Derive the boundary conditions given in Eq. (2.18).

2.4 The displacement current may be needed in analysis involving dielectrics when its amplitude cannot be neglected. Consider SiO_2 dielectric and Si substrate used in RFIC. Assume SiO_2 and Si have relative dielectric constant of 4.2 and 12, and conductivity of 0.0002 S/m and 12.5 S/m, respectively.

a) Calculate the conduction and displacement currents in SiO_2 and Si at 10, 50, and 100 GHz and comment as to whether the displacement current should be taken into account.

b) Determine the frequency at which the conduction and displacement currents in SiO_2 produced by a sinusoidal signal are equal. Repeat the calculation for Si. Provide comment concerning the results.

2.5 Consider copper, gold, and silver as the metals used in RFIC. Assume copper, gold, and silver have relative dielectric constant of 1 and conductivity of 5.8×10^7 S/m, 4.1×10^7 S/m, and 6.17×10^7 S/m, respectively.

a) Calculate the conduction and displacement currents in these conductors at 10, 50, and 100 GHz and draw a conclusion as to whether the displacement current should be taken into account.

b) Calculate the frequency at which the conduction and displacement currents are equal in each of these conductors and comment on the results.

2.6 For SiO_2 and Si considered in Problem 2.4, calculate the loss tangent at 10, 50, and 100 GHz.

2.7 For copper, silver, and gold conductors considered in Problem 2.5, calculate the following parameters at 10, 50, and 100 GHz:

a) skin depth, and

b) surface impedance.

2.8 Consider a transmission line in an RFIC with the conductors along the z direction. Assume that the electric and magnetic fields are along the x, y, and z directions, and the signal is sinusoidal.

a) Find the average power density and determine the direction of power flow.

b) Assume the transmission line is quasi-TEM (transverse electromagnetic), find the average power density and the direction of power flow.

CHAPTER 3

LUMPED ELEMENTS

Lumped elements, including inductors, capacitors, and resistors, are essential components, along with printed-circuit transmission lines, for high frequency integrated circuits (ICs), particularly radio-frequency integrated circuits (RFICs). Lumped elements have evolved significantly from their development and use for analog ICs as well as low frequency to high frequency RFICs (up to millimeter-wave regime). Besides standard on-chip lumped elements available to designers from complementary metal oxide silicon (CMOS) foundries, custom-designed lumped elements can be realized on CMOS processes up to millimeter-wave frequencies, depending on the process and structures used to realize the lumped elements. In this chapter, we will present the design, analysis, and results of on-chip lumped elements for RFICs, including inductors, capacitors, and resistors. Off-chip lumped elements are also used in subsystems and systems employing RFICs. These, however, will not be covered.

3.1 FUNDAMENTALS OF LUMPED ELEMENTS

In general, lumped elements[1] are electrical elements that behave as inductors, capacitors, or resistors. A short transmission line with very low and high characteristic impedance behaves dominantly as a capacitor or inductor, respectively, and, hence, may also be characterized as a lumped element. Specifically, lumped elements refer to any passive element whose dimensions (in any direction) are very small compared to the wavelength at a particular frequency, at which the circuit theory is valid, so that its dominant effect to an electrical signal or wave is inductive, capacitive, or resistive. Typically, a lumped element's maximum dimension is less than one-twentieth of a wavelength ($\lambda/20$).

In a rather strict view and within the RFIC domain, printed-circuit transmission lines[2] (e.g., microstrip line) or wave-guiding structures are commonly referred to as distributed elements. In general, from the electromagnetic (EM) point of view, distributed element refers to any physical structure within which the energy store is contributed dominantly by both electric and magnetic fields. That is, from the circuit theory concept, both inductive and capacitive effects are dominantly present. It should note that resistive effect is always

[1] Here, we exclude lumped-element resonators consisting of inductor, capacitor, and resistor. They are covered in Chapter 5 of Resonators.
[2] Printed-circuit transmission lines are covered in Chapter 4.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

present in any practical elements due to losses, but is not considered in this discussion without loss of generality. The existence of either electric energy, magnetic energy, or both in certain magnitude dictates the nature of an element as either lumped or distributed, and the resultant transition frequency between them. Analytically (and also intuitively), we know that a transmission line would behave as a (lumped-element) inductor at low frequencies where $\ell \ll \lambda$ with some resistive and capacitive parasitics. As the frequency is increased, the stored electric and magnetic fields increase, giving rise to the corresponding capacitance and inductance, resulting in a distributed structure for the transmission line. Of course, as the frequency is increased, resistance and conductance should also increase as well. When we determine these values through an optimization process, it could be difficult to determine the transition frequency between lumped and distributed structure through inspection of these values, since an optimization process provides values which may not be physically valid. Nevertheless, some good indications can be obtained for the transition frequency if the optimization is carried out carefully with good initial guesses.

For transmission lines, the transition frequency between lumped and distributed structure is considered the frequency at which the length of the transmission lines is a fraction of the wavelength. Although there is no strict rule for this, we typically choose this length as $\lambda/20$. To illustrate this phenomenon, we consider a narrow-width microstrip line in a 0.25-μm CMOS structure as shown in Figure 3.1. The top conducting strip is deposited on the topmost metal layer M5 and the ground plane on the lowest metal layer M1, separated from M5 by SiO_2 dielectric layers. The width and length are chosen as 2 μm and 1 mm, respectively. Using the $\lambda/20$ rule of thumb, we can approximate the transition frequency as around 8.7 GHz, corresponding to $\lambda \simeq 20$ mm. In another more analytical way to determine the transition frequency, we examine the inductance and capacitance exhibited by the microstrip line. Figure 3.2 shows a simple equivalent-circuit model of the microstrip line consisting of series resistance R and inductance L, and shunt conductance G and capacitance C. Figure 3.3 shows the reactance of the series inductance and shunt capacitance versus frequency, approximately obtained by fitting the S-parameters of the microstrip line calculated using the EM simulation program IE3D [1] to those calculated from the equivalent circuit using the Agilent Advanced Design System (ADS) [2] through an optimization process. It is noted that, due to the nature of optimization, these values, although mathematically valid, may not be physically correct. The model's inductance L is fairly constant, around 1.1–1.2 nH, up to about 10 GHz while its capacitance C is quite small. We can therefore conclude that, up to 10 GHz, the considered microstrip line behaves more like a lumped-element inductor with a

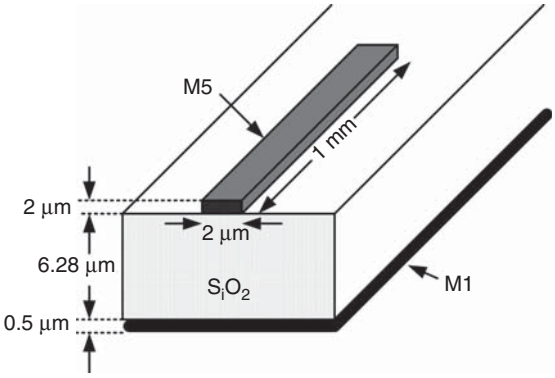


Figure 3.1. Microstrip line.

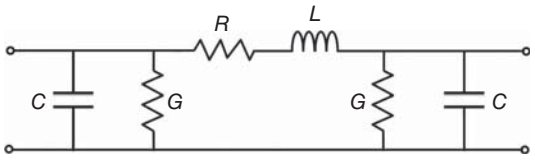


Figure 3.2. Equivalent-circuit model of a microstrip line.

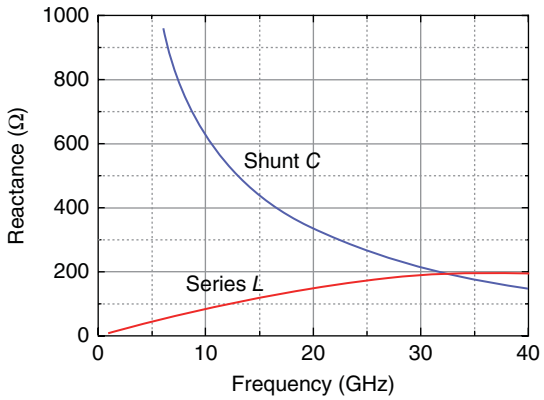


Figure 3.3. Calculated series inductive and shunt capacitive reactance of a microstrip line on a 0.25-μm CMOS process.

small capacitive effect; after that, both the inductive and capacitive effects become dominant. The transition frequency between lumped and distributed element can thus be approximated as about 10 GHz, which falls within the vicinity of $\lambda/20$. The calculated resistance R and conductance G are negligibly small due to the use of a low loss SiO_2 as the dielectric layer. It should be noted that, another figure of merit which determines the operating frequency range of this inductive element is its self-resonant frequency resulted from the resonance caused by both inductive and capacitive elements. This resonant frequency, however, is not considered here in order to clearly illustrate the transition between lumped and distributed nature. In another example demonstrating the distributed nature of lumped elements, we consider a metal–insulator–metal (MIM) capacitor in a CMOS process. MIM capacitors are the most inherent capacitors and used most often in RFICs. The considered MIM capacitor has the following parameters: size of 30-μm × 30-μm; oxide layer of 1.6-μm thick, relative dielectric constant of 3.8, and loss tangent of 0.0002; aluminum plates with top and bottom metallization of 1 and 0.6 μm, respectively. The bottom plate is separated from a grounded 300-μm silicon substrate by SiO_2 layers with total thickness of 8 μm. The MIM structure resembles a parallel-plate transmission line consisting of two conductors separated by a SiO_2 dielectric layer, and so we expect that it would behave as a distributed structure at sufficiently high frequencies, where its dimensions are no longer very small compared to the wavelengths. As on-chip MIM capacitors normally behave well up to very high frequencies, it is also expected that the transition frequency at which a lumped-element operating mode is changed to a distributed mode for such a small size capacitor (30 μm × 30 μm) may reach well into the upper end of the millimeter-wave range. We assume the MIM capacitor is modeled as an equivalent-circuit model as shown in Figure 3.42(c). In this model, C is the main capacitance while others are parasitics. Table 3.1 shows the values of the model's parameters at different frequencies, approximately determined by fitting the S-parameters of the MIM capacitor calculated using IE3D to those calculated from the equivalent circuit using ADS through an optimization process.

As can be seen, the capacitance C appears fairly constant, while the parasitics R_T, R_B, L_T, L_B, C_S and $1/G$, R_s are small and large, respectively, across 0.5–60 GHz as desired, signifying that the MIM capacitor behaves as a very good lumped element to at least 60 GHz at which the quality factor (Q)[3] is 18.6 as seen in Figure 3.4. To investigate the transition frequency, we compute the Q of the capacitor up to 400 GHz using Eq. (3.120) or (3.121) and show the results in Figure 3.4. As can be seen, the Q crosses zero at around 200 GHz, which identifies the self-resonant frequency of the capacitor or its upper operating frequency. Above 200 GHz, the Q alternates between negative and positive values, displaying the characteristics of a distributed structure. The transition frequency between the lumped- and distributed-element operations can thus be considered 200 GHz, which is extremely high as expected. For MIM capacitors with larger dimensions, the self-resonant frequency and transition frequency are lower.

[3]Quality factor is covered in Section 3.2.

TABLE 3.1. Calculated Values of the MIM Capacitor's Parameters

Frequency (GHz)	R_T (mΩ)	L_T (pH)	C (fF)	$1/G$ (KΩ)	R_s (KΩ)	C_S (fF)	R_B (mΩ)	L_B (pH)
0.5	0.482	0.92	20.76	92340.4	5421.7	2.36	0.45	0.96
1	1.134	0.99	20.79	42730.7	5110	2.33	1.096	0.99
5	0.772	1.01	20.8	6070	5253.8	2.311	0.752	0.951
10	8.15	1	20.89	2302.2	5015.4	2.31	9.21	1
15	30.6	1	20.98	1653.8	5251.6	2.32	37	1
20	64.09	1	21.11	2074.5	5120	2.33	77.9	0.99
25	85.64	1	21.29	2803.4	5130.5	2.35	104.7	0.99
30	106	0.99	21.5	3174.5	5183.5	2.36	130.6	0.99
35	148.6	1	21.81	4510.5	5893	2.4	171.8	0.904
40	207.3	0.982	22.16	3953.8	5805.5	2.43	267	0.98
50	680.3	0.861	23.3	3518.4	5116.4	2.55	838.2	0.85
60	2690.7	0.744	25.05	2996.7	5966	2.74	3252.2	0.738

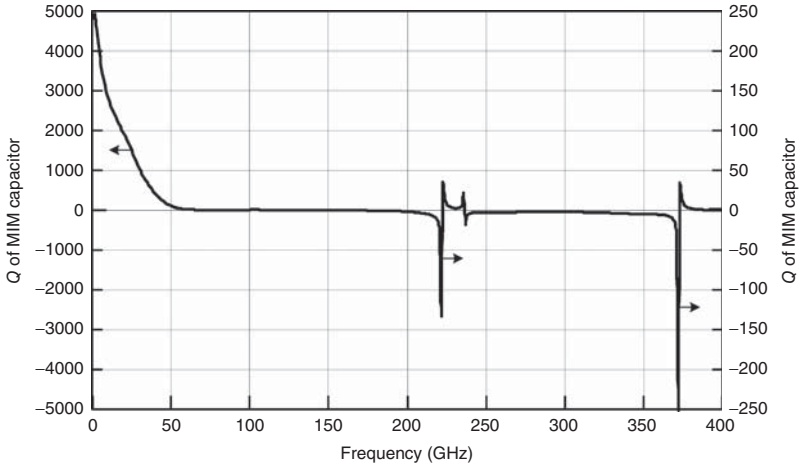


Figure 3.4. Calculated Q of the considered MIM capacitor.

3.1.1 Basic Equations

Simple equations are presented here to provide basic understanding of the lumped elements. These equations, when used together with other techniques, can also serve as ways for calculating accurately the lumped element's parameters. As will be seen later in this chapter, circuit parameters of lumped elements' equivalent-circuit models, which are commonly used in RFIC design, are normally obtained by fitting measured or calculated (from an EM simulator) S-parameters to those calculated from the equivalent-circuit models, and thus may not represent the correct values of these parameters. Using the basic equations along with an EM simulator may allow RFIC designers to calculate more accurate element values, which, when used in conjunction with a fitting program, will provide an accurate model whose element values represent closely the actual values of the elements. Here we assume all lumped elements are linear time-invariant, even though in practice they generally behave as nonlinear time-varying elements; the degree of nonlinearity and time-variation depends on the conditions upon which these lumped elements are imposed to. Only inductor and capacitor are discussed here. Resistor is not discussed considering its ready availability from foundries, its relative simplicity, and the fact that few custom-design resistors are needed.

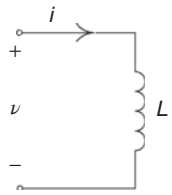


Figure 3.5. Inductor and its terminal voltage and current flow.

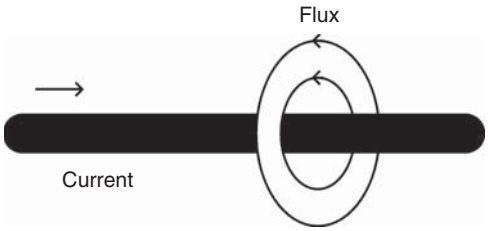


Figure 3.6. Flux produced by a current flowing in a conductor.

3.1.1.1 Inductor. Figure 3.5 shows the schematic of an inductor. The current $i(t)$ flowing in the inductor and the terminal voltage $v(t)$ are related by the expression

$$i(t) = \frac{1}{L}\int_{-\infty}^{t} v(\tau)d\tau \tag{3.1}$$

where L denotes the inductance.

Let us consider a conductor along which a current flows. The current flow produces magnetic lines of flux, which link the conductor as shown in Figure 3.6. The flux linkages $\phi(t)$ (weber-turns) produced by the current $i(t)$ flowing in the conductor are related to the induced voltage $v(t)$, from Faraday's law, as

$$v(t) = \frac{d\phi(t)}{dt} \tag{3.2}$$

where $d\phi/dt$ is known as the time rate of change of flux linkages.

The current flowing in the conductor shown in Figure 3.6 and its produced flux linkages are related via the inductance L of the conductor as

$$\phi(t) = Li(t) \tag{3.3}$$

Substituting $\phi(t)$ from (3.3) into (3.2) gives the following expression for inductance:

$$L = \frac{1}{v(t)}\frac{di(t)}{dt} \tag{3.4}$$

The instantaneous power $p(t)$ supplied to an inductor is obtained as

$$p(t) = v(t)i(t) = Li(t)\frac{di}{dt} \tag{3.5}$$

which may be positive or negative. It is possible, at any given time, that an inductor receives energy from or supply energy to the circuit, to which it is connected, which corresponds to positive or negative $p(t)$, respectively. The total energy $w(t)$ stored in an inductor is determined from

$$w(t) = \int_{-\infty}^{t} p(\tau)d\tau = \frac{Li^2(t)}{2} \tag{3.6}$$

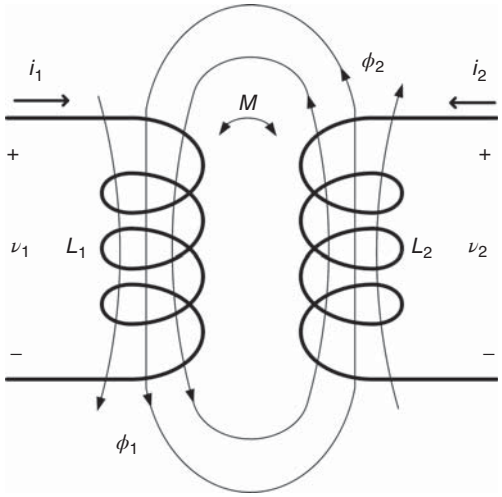


Figure 3.7. Two coupled coils and their flux linkages.

which may also be positive or negative depending on L. This energy is actually stored in the magnetic field associated with the inductor. For a passive inductor, which is commonly used and addressed in this chapter, the stored energy is always positive. Active inductors, realized by active devices such as CMOS transistors, may possess negative inductance, but negative-valued inductors have very little use in practice.

Coupled Inductor

In RFIC design, parallel- or series-connected inductors are sometimes used to realize a particular value for inductance. A common element is stacked inductor consisting of several spirals connected in series over multiple metal layers. The use of multiple inductors results in a mutual inductance between adjacent elements, caused by the magnetic coupling between these elements.

We use two simple coupled coils, as shown in Figure 3.7, which may represent two separate adjacent spirals of a spiral inductor or two individual inductors coupled to each other, to derive an analysis without loss of generality. The current flowing in one coil generates the fluxes for that coil and some of the flux lines also link the other coil due to their vicinity.[4] Thus, the flux linkages of one coil of a pair of coupled coils are determined by the currents flowing in both of the coils. The flux linkages $\phi_1(t)$ and $\phi_2(t)$ of the first and second coil, respectively, may therefore be expressed as

$$\phi_1(t) = L_1 i_1(t) + M_{12} i_2(t) \tag{3.7}$$

and

$$\phi_2(t) = M_{21} i_1(t) + L_2 i_2(t) \tag{3.8}$$

where $L_j (j = 1, 2)$ is the self inductance of coil j, $i_j(t)$ is the current flowing in coil j, and $M_{jk} (k = 1, 2)$ represents the mutual inductance of coil j with respect to coil k. It can be proved that $M_{12} = M_{21} = M$, where M is referred to as the mutual inductance of the two coupled coils. The total inductance L of the two coupled coils can be derived as

$$L = L_1 + L_2 \pm 2M \tag{3.9}$$

The plus (+) and negative (−) signs correspond to the directions of the flux lines such that the flux are added or subtracted.

[4]The direction of the flux lines follows the right-hand rule, which states that if the fingers of the right hand are closed around a coil in the positive direction of the current flow, then the thumb will point in the positive direction of the flux.

The voltage $v_1(t)$ and $v_2(t)$ appearing across the terminals of coil 1 and coil 2, respectively, are obtained, by Faraday's law, as

$$v_1(t) = \frac{d\phi_1(t)}{dt} = L_1 \frac{di_1(t)}{dt} + M_{12} \frac{di_2(t)}{dt} \tag{3.10}$$

and

$$v_2(t) = \frac{d\phi_2(t)}{dt} = M_{21} \frac{di_1(t)}{dt} + L_2 \frac{di_2(t)}{dt} \tag{3.11}$$

The total energy stored in the two coupled coils may be derived as

$$w(t) = \frac{1}{2} L_1 i_1^2(t) + M i_1(t) i_2(t) + \frac{1}{2} L_2 i_2^2(t) \tag{3.12}$$

It should be noted that the mutual inductance M may be positive or negative. M is positive or negative when the positive directions for the flux linkages of the first and second coils are such that they add to or subtract from each other, respectively, which depend on the relative directions of the voltages and currents at the terminals of the two coils. Equation (3.12) can be rewritten as

$$w(t) = \frac{1}{2} \left\{ \left[\sqrt{L_1} i_1(t) + \frac{M}{\sqrt{L_1}} i_2(t) \right]^2 + \left(L_2 - \frac{M^2}{L_1} \right) i_2^2(t) \right\} \tag{3.13}$$

Since the total energy stored in coupled coils must be positive, it is required that

$$|M| \leq \sqrt{L_1 L_2} \tag{3.14}$$

That is, the mutual inductance of a pair of two coupled coils is bounded by the geometric means of the self inductances of the two coils. Consequently, the mutual inductances in multiple coupled coils always have upper bounds determined by the coils' self inductances. The mutual coupling between coupled coils is normally described by the coupling coefficient

$$k = \frac{|M|}{\sqrt{L_1 L_2}} \tag{3.15}$$

k is small for loosely couple coils, and is equal to 1 when the coils are perfectly coupled to each other; that is, all the flux generated by one coil links the other coil. The couplings in stacked spiral inductors on RFIC structures are considered tight due to thin dielectrics between the metal layers, and are normally assumed to be unity.

Now consider a general N coupled coils. The following equations may be derived for the flux linkages $\phi_j(t)(j = 1, 2, \ldots, N)$ of the individual coils:

$$\begin{aligned}
\phi_1(t) &= L_1 i_1(t) + M_{12} i_2(t) + M_{13} i_3(t) + \cdots + M_{1N} i_N(t) \\
\phi_2(t) &= M_{21} i_1(t) + L_2 i_2(t) + M_{23} i_3(t) + \cdots + M_{2N} i_N(t) \\
&\vdots \\
\phi_N(t) &= M_{N1} i_1(t) + M_{N2} i_2(t) + \cdots + M_{N,N-1} i_{N-1}(t) + L_N i_N(t)
\end{aligned} \tag{3.16}$$

where $L_i(i = 1, 2, \ldots, N)$ is the self inductance of coil i, $M_{ij}(i, j = 1, 2, \ldots, N)$ is the mutual inductance of coil i with respect to coil j, and $M_{ij} \neq M_{ji}$ in general. The total inductance of the coupled coils can be obtained, assuming all flux are added, as

$$L = \sum_{i=1}^{N} L_i + M_{21} + \sum_{j=1}^{N-1} M_{Nj} + \sum_{j=2}^{N} M_{1j} + \sum_{j=3}^{N} M_{2j} \tag{3.17}$$

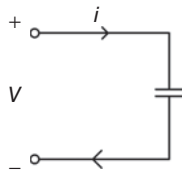


Figure 3.8. Capacitor and its terminal voltage and current flow.

3.1.1.2 Capacitor. Figure 3.8 shows the schematic of a capacitor. The capacitor's terminal voltage $v(t)$ and current flow $i(t)$ are related by the expression

$$v(t) = \frac{1}{C}\int_{-\infty}^{t} i(\tau)d\tau \tag{3.18}$$

where C represents the capacitance.

The capacitance is determined from the terminal voltage $v(t)$ and current flow $i(t)$ as

$$C = \frac{1}{i(t)}\frac{dv(t)}{dt} \tag{3.19}$$

The capacitance determines the relation between the charge q stored in the capacitor and the voltage across the capacitor as

$$q(t) = Cv(t) \tag{3.20}$$

The instantaneous power $p(t)$ supplied to a capacitor is obtained as

$$p(t) = v(t)i(t) = Cv(t)\frac{dv}{dt} \tag{3.21}$$

which may be positive or negative. It is possible, at any given time, that a capacitor receives energy from or supply energy to the circuit, to which it is connected. These correspond respectively to positive or negative $p(t)$. The total energy $w(t)$ stored in a capacitor (actually, this energy is stored in the electric field associated with the capacitor) is determined from

$$w(t) = \int_{-\infty}^{t} p(\tau)d\tau = \frac{Cv^2(t)}{2} \tag{3.22}$$

which may also be positive or negative depending on C. For a passive capacitor, which is commonly used and addressed in this chapter, the energy stored in the capacitor is always positive. It is possible to define a negative-valued capacitor, whose capacitance and energy are negative. Such a capacitor supplies energy to the circuit, to which it is connected, and is considered an active device.

3.1.1.3 Use of Lumped Elements in RFIC. In the past, at high frequencies (radio frequency, RF, regime), it was difficult to realize good lumped elements and use them in RF circuits due to various reasons such as inferior CMOS technology, lack of proper lumped element topologies and models, etc. For example, a lumped element inductor does not behave as a good inductor. This lack of good lumped elements resulted in RF circuits employing the most distributed structures (i.e., transmission lines) instead of lumped elements. Advances in CMOS technologies make feasible realization of good lumped elements at RF frequencies, up to millimeter-wave region depending on processes and geometries of lumped-element structures. This, along with good equivalent-circuit models for these elements, allows usage of many lumped elements in RFICs.

As compared to distributed elements, lumped elements have a smaller size (leading to lower cost for RFICs), wider bandwidth, smaller coupling between adjacent elements, and smaller amplitude and phase changes. The smaller coupling and amplitude and phase changes are due to the smaller size relative to wavelength of lumped elements. All these advantages allow traditional low frequency lumped-element analog circuits to be realized at RF frequencies. For instance, Gilbert mixers and Colpitts oscillators normally feasible only at low frequencies using lumped elements can now be designed at RF frequencies. Designing these components to achieve both high performance and miniaturization in the RF regime is relatively difficult using distributed elements such as a microstrip line or coplanar waveguide (CPW). Lumped elements can also be employed to provide a smaller size and wider bandwidth for RF bias chokes of RFICs as compared to those using distributed elements. Lumped elements, however, have several shortcomings. They have relatively lower quality factor mainly due to smaller dimensions; hence there is a smaller ratio between volume and surface area,[5] and possible need of multiple metal layers, particularly for realization of high inductance values. The inductor's metal layers below the top one have thinner metallization, resulting in more loss. Lumped elements have self-resonant frequencies, due to their behavior as resonators at certain high frequencies, thus having lower upper operating frequency as compared to distributed elements.[6] It should be noted that, although a lumped-elements size is much smaller than a distributed-element size, a particular lumped element may still be considered large for some RFICs. For example, lumped-element spiral inductors, whose inductance is greater than 10 nH in RFIC fabrication processes, typically occupy large space in RFICs. Therefore, miniature lumped elements, particularly inductors, are needed for RFICs.

3.2 QUALITY FACTOR OF LUMPED ELEMENTS

Practical inductors and capacitors have resistive parasitic, which can be substantial at RF, particularly those on RFICs. This resistive part gives rise to loss and degrades the inductor or capacitor's quality. Quality factor (Q) of an inductor or capacitor is a figure of merit dictating the loss of the inductor or capacitor, and hence its usefulness as a circuit element. It is a basic parameter specifying how good an inductor or capacitor is in storing the magnetic or electric energy in it, respectively, and, effectively, dictates how an inductor or capacitor would contribute to the performance of RFICs implementing it. Q of lumped elements, or, in general, of any electronic components, is conventionally defined as the ratio between the total time-average stored energy (electric and magnetic) and the time-average power loss or dissipated within the elements. Note that this definition is also used for resonators discussed in Chapter 5. Using this definition, different equations for the Q of a lumped element can be derived depending on whether it is modeled as a series circuit, a parallel circuit, or a combination of them.

Consider an inductor modeled simply as a series circuit as shown in Figure 3.9(a), where L is the inductance and R_L is the resistance accounting for the loss in the inductor. Although no parasitic capacitors are assumed in the model, they occur in practical inductors and should be considered, especially in the RF range. Taking the ratio between the stored energy, which is solely magnetic energy, and power loss, the Q of the inductor can be derived as

$$Q = \frac{\omega W_m}{P_L} = \frac{\frac{1}{2}\omega L I_{\text{rms}}^2}{\frac{1}{2} R_L I_{\text{rms}}^2} = \frac{\omega L}{R_L} \tag{3.23}$$

where I_{rms} represents the (rms) current flowing through the inductor, ω is the operating (radian) frequency, $W_m$is the time-average stored magnetic energy, and P_L represents the loss in the inductor due to its resistive part.

[5]Quality factor of an element is proportional to the element's volume-to-surface-area ratio as discussed in Chapter 5 for resonators.

[6](Distributed) transmission lines on RFICs are typically operated as quasi-TEM mode and can be used up to a frequency where loss becomes significant or higher-order propagation modes begin to appear. This frequency depends on the transmission line's geometry, but normally is in the high RF frequency range, well over typical operating frequencies of RFICs.

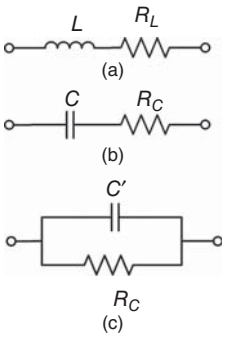


Figure 3.9. Simple series-circuit model of inductors (a) and capacitors (b), and simple parallel-circuit model of capacitors (c).

The Q of a capacitor modeled as a capacitor in series with a resistor as shown in Figure 3.9(b) can be obtained in a similar way as

$$Q = \frac{\omega W_e}{P_L} = \frac{\frac{1}{2}\omega C V_{\text{rms}}^2}{\frac{1}{2} R I_{\text{rms}}^2} = \frac{\frac{1}{2}\omega C \frac{I_{\text{rms}}^2}{\omega^2 C^2}}{\frac{1}{2} R_C I_{\text{rms}}^2} = \frac{1}{\omega R_C C} \tag{3.24}$$

where C is the capacitance, R_C is the resistance representing loss of the capacitor, V_{rms} represents the (rms) voltage across the capacitor, and I_{rms} denotes the (rms) current flowing through the capacitor. W_e is the time-average stored electric energy, and P_L represents the loss in the capacitor due to its resistive part. It is noted that parasitic inductors, which should be considered in RF, are not included in the model. Equations (3.23) and (3.24) show that Q increases as resistance decreases.

Now consider a capacitor modeled as parallel combination of a capacitor and a resistor as shown in Figure 3.9(c). The Qcan be derived as

$$Q = \frac{\omega W_e}{P_L} = \frac{\frac{1}{2}\omega C'_C V_{\text{rms}}^2}{\frac{1}{2}\frac{V_{\text{rms}}^2}{R'_C}} = \omega R'_C C' \tag{3.25}$$

where C' is the capacitance, R'_C is the resistance representing loss of the capacitor, and V_{rms} represents the (rms) voltage across the capacitor. Equation (3.25) indicates that Q of a capacitor modeled as a parallel circuit increases as the resistance increases. This result is expected as the higher the resistance the closer the modeled capacitor to a pure capacitor and the less loss due to the dielectric between the capacitor's plates.

Practical inductors and capacitors[7] also contain parasitic capacitance and inductance, respectively, as mentioned earlier, which can be substantial at RF frequencies. Therefore, Eqs. (3.23)–(3.25) are valid only at low frequencies, where parasitic capacitance and inductance are negligibly small. In the RF region, these parasitics cannot be neglected and the stored energy would consist of both magnetic and electric energies, resulting in different equations for Q, as discussed in Sections 3.4.4 and 3.5.4.

It should be noted that the forgoing equations for Q are for inductors or capacitors, which behave dominantly as inductors or capacitors, respectively, in their operating frequency range, and are used in the initial design of inductors or capacitors. These elements, however, also behave as a resonator at a certain frequency, referred to as resonant frequency, at which their Qs are for resonators and are different from those given in Eqs. (3.23)–(3.25). These Qs are not used to specify the performance of the inductor or capacitor in RFICs and are discussed in Chapter 5 for resonators.

[7]Practical inductors or capacitors do not behave as pure inductors or capacitors, except at DC.

Lumped-element inductors and capacitors are made of conductors and dielectrics on top of a silicon substrate, whose contributions in Q can be separated using the following equation borrowed from the Q of resonators discussed in Chapter 5:

$$\frac{1}{Q} = \frac{1}{Q_c} + \frac{1}{Q_d} + \frac{1}{Q_s} \tag{3.26}$$

where Q_c, Q_d, and Q_s are the conductor, dielectric, and substrate quality factors, respectively. Q_c can be calculated approximately from the loss of the conductors.

Q_d is derived in Chapter 4 of transmission lines, assuming homogeneous medium, as

$$Q_d = \frac{1}{\tan\delta} \tag{3.27}$$

where $\tan\delta$ represents the dielectric's loss tangent. Although the quality factor of an isolated substrate can be approximately calculated from Eq. (3.27) with $\tan\delta$ representing the substrate's loss tangent, this equation may not be used in calculating Q_s for lumped-element inductors due to the fact that inductors are typically placed far away or isolated (e.g., through a conductive shield) from the silicon substrate. Nevertheless, Q_s contributes significantly to the Q of inductors and so needs to be included. In CMOS processes, the dielectrics typically have relatively low loss as compared to the metals and Si substrate and may be neglected, particularly at low RF frequencies, in calculating the Q of inductors. On the other hand, the Si substrate does not affect the Q of lumped-element capacitors much due to the fact that most of the electric energy is concentrated between the conductors (or metallic electrodes) surrounding the dielectrics of capacitors, unless one of the electrodes is deposited directly on top of the substrate.

It is important to note that, the Q of a network consisting of inductors and capacitors are not determined by individual inductors or capacitors (e.g., by the lowest Q of the inductors and capacitors), but the combination of the employed inductors, capacitors, and the network topology. This is indeed expected from both the practical point of view and theoretical point of view – for instance, from the general Q expression $Q = \omega W/P_L$, where W is the total time-average stored energy in the network and P_L is the power loss of the network. Hence, different network topologies (e.g., series L/shunt C and shunt C/series L) using the same inductors and capacitors would have different Q in general.

3.3 MODELING OF LUMPED ELEMENTS

As for any circuit elements, accurate models for on-chip inductors and capacitors, either numerical or equivalent-circuit, are needed for accurate RFIC analysis and design. At very low frequency, ideally DC, a lumped-element inductor or capacitor behaves like a pure inductor or capacitor, respectively. However, at RF frequencies, a lumped element has resistive, inductive, and capacitive effects, leading to various possible equivalent circuits consisting of resistors, inductors, and capacitors. In general, for lumped elements in RFICs, accurate equivalent-circuit models should consist of multiple series and parallel elements, considering the physics of the lumped-element geometries. This is due to the fact that multiple metal layers, multiple Oxide layers, and lossy silicon substrates, are typically used in conjunction with different geometries for lumped elements, which result in complicated interactions among elements electrically constituting the lumped elements, and hence complex models. For instance, lossy silicon substrates or multiple metal layers used in inductors produce various parallel and series parasitic capacitors and resistors to inductor models. It may be quite complicated to model accurately lumped elements used in RFICs, especially for complex structures like multilayer spiral inductors or multilayer capacitors. In general, the modeling process for lumped elements can be categorized into three approaches:

1. **Lumped-element approach:** This approach employs frequency-dependent formulas for inductors, capacitors, and resistors based on circuit theory. It is useful when the dimensions of a lumped element are much smaller than the wavelength at the highest operating frequency and is suitable only at low

RF frequencies. This approach produces directly equivalent-circuit models, which are very convenient for computer-aided analysis and design of RFICs.

2. **Transmission-line approach:** In this approach, a lumped element is modeled as a set of transmission lines. As such, it can also be classified as a distributed approach. Since most circuit structures behave as distributed in the RF regime, especially in the high frequency ends, this approach is more accurate than the lumped-element approach and suitable at RF frequencies. Transmission line effects such as discontinuities, radiation, etc. may be included in this approach. The transmission line approach becomes complicated for complex structures, especially when transmission line effects such as discontinuities are taken into account. This approach is thus deemed to be good for simple structures and at frequencies not in the high RF region (e.g., below 18 GHz). The transmission-line model is also convenient for computer-aided analysis and design of RFICs.
3. **EM approach:** This approach is a truly distributed method. It is based on full-wave EM techniques, such as finite-difference time-domain (FDTD) method, through an explicit analysis or analysis via an EM simulator. It is more accurate at RF frequencies and/or for complex structures as compared to the lumped element or transmission line approach. The approach takes into account distributed effects such as discontinuities, radiation, interaction between the elements constituting lumped element, etc. Modeling this explicitly using a full-wave approach is quite complex and time consuming. Results of EM analysis of lumped elements are typically given numerically, such as values of scattering (*S*) parameters over a particular frequency range, which are not very convenient for RFIC design. For convenience in RFIC design, an equivalent-circuit model is normally derived from these numerical results, usually through a commercially available CAD (computer aided design) program, which match the *S*-parameters calculated from the EM analysis to those calculated from the equivalent-circuit model.
4. **Measurement approach:** This approach is most accurate but, from a strict point of view, applicable only to the measured lumped elements. It consists, in general, of two steps. The first step is measurement of DC parameters such as resistance and measurement of RF parameters such as *S*-parameters. The second step involves fitting the measured data to an equivalent circuit model using a commercially available CAD program. This approach is similar to the modeling of transistors based on measured *S*-parameters detailed in Chapter 9.

Equivalent-circuit models for lumped elements are very convenient and hence preferred for computer-aided analysis and design of RFICs. An accurate equivalent-circuit model can be obtained by fitting the *S*-parameters of the equivalent circuit to those measured or calculated by an accurate full-wave method, as mentioned in the EM and measurement approaches, across a frequency range of interest. Approximate equations for calculating a lumped element's parameters, such as the inductance and parasitic capacitances and resistances of an inductor, are also useful for circuit design as well as for calculating the initial values needed in the fitting process of an equivalent-circuit model. It is common practice to assess the accuracy of a particular model or equations by comparing the values of the model's parameters or values calculated from the equations with those in an accurate equivalent-circuit model obtained by a fitting process. This method of comparison, however, may not result in a very meaningful conclusion. The reason is the parameters of the accurate equivalent-circuit model are obtained through an optimization process and, as such, may not represent faithfully the actual values of the lumped elements' parameters. Nevertheless, these values may give an insight to the behavior of lumped elements qualitatively. Although the accuracy of individual parameters is difficult to assess, the accuracy of the entire equivalent-circuit model, which is essentially needed for the design of RFICs, can be evaluated through the comparison of the model's overall performance, notably its *S*-parameters.

In certain equivalent-circuit models, the fitting process may not produce converging element values at frequencies above a particular frequency because these models do not represent the actual behavior of the component at those frequencies any more. Different models need to be used to allow convergence to occur. It is noted that, for any device modeling, the model should reflect the actual physical behavior of the structure. Another note is that one of the main reasons for using a lumped-element equivalent-circuit model for any structures [inductor, capacitor, metal–oxide–semiconductor field-effect transistor (MOSFET), etc.] is to

facilitate the analysis and optimization of RFICs employing these structures without using their *S*-parameters, particularly to frequencies beyond those where the *S*-parameters are available. Therefore, it is always desirable to fit (optimize) all parameters across the interested frequency range to obtain a single set of element values working at all frequencies, instead of fitting to get different set of values for different frequencies. This, of course, is more difficult and may never converge. Yet, it should always be the first goal of RFIC designers to achieve. If that is not doable, one may need to alter the model itself or break the frequency range into several smaller ranges to fit.

3.4 INDUCTORS

Typical on-chip RFIC passive lumped-element inductors[8] have low Q, low inductance value (relative to size), and low resonant frequency, which are serious drawbacks in realizing high performance RFICs. An ideal inductor is represented by a single inductance L. Practical inductors, however, consists of inductance and parasitic capacitance and resistance, which result in resonance and finite Q. Three important electrical parameters characterizing an inductor are its inductance, resonant frequency, and Q. In RFIC design, the designers are interested not only in realizing a particular inductance for the inductor, but also in achieving a Q as high as possible across the operating frequency range and a resonant frequency well above this frequency range.

3.4.1 Inductor Configurations

On-chip inductors can be formed in virtually many ways. They can be realized in principle using a single conducting strip or multiple conducting strips connected in a particular arrangement. Except when it is narrow and relatively short compared to a wavelength, a strip used as an inductor or part of an inductor should be relatively far away from other adjacent conducting strips, with respect to the operating wavelength, to avoid effects of transmission lines possibly formed by that strip and other strips.

Figure 3.10 shows some common inductor configurations. The conducting strip, conducting loops, and conducting meander strip, shown in Figure 3.10(a)–(d) can be used for small inductances. The spiral inductors are perhaps the most commonly used structures. They are shown in Figure 3.10(e)–(g) and normally used for both small and large inductances. Inductors are typically implemented on a single metal layer (usually the topmost layer of CMOS structures, although other layers can also be used for a circuit topology to satisfy certain conditions such as layout, size or convenient interface with other circuit elements). Topmost layer has thickest metallization and farthest from the silicon substrate, resulting in less loss and hence higher Q. Multiple metal layers can also be employed for any inductor configurations to increase inductance without much increase in size. One example is stacked spiral inductors to be described later. Among these spirals, the circular spiral has the highest Q. Circular spiral, however, may not be allowed in certain CMOS processes due to its circular strips. The rectangular spiral has the lowest Q, but is also available and commonly used in every process. The polygonal spirals (hexagonal, octagonal, etc.) offer a compromise between the rectangular and circular spirals for Q, while providing convenience in layout and acceptance to certain processes. In principle, spirals with less abrupt discontinuities (e.g., right-angle bend) on the strip produce higher Q and resonant frequency.

Inductors fabricated on a single metal layer have limited inductance considering the size constraint typically imposed for RFICs. To increase the inductance without much increase in the die size, multiple spirals placed on top of each other in series can be used. These inductors are known as "stacked spirals" and are shown in Figure 3.11. The individual spirals are interconnected in such a way to allow currents flowing in the same direction in all the spirals. The total magnetic flux produced by the stacked spiral is thus equal to a summation of all the fluxes generated by the individual spirals, resulting in increased inductance.

[8]Active devices such as CMOS transistors can act as inductors, but they have high noise, distortion, intermodulation, power consumption and nonlinearity, and thus are not very desirable, unless particular circumstances exist; for example, the size of a comparable lumped-element inductor is prohibitively large for the designed RFICs.

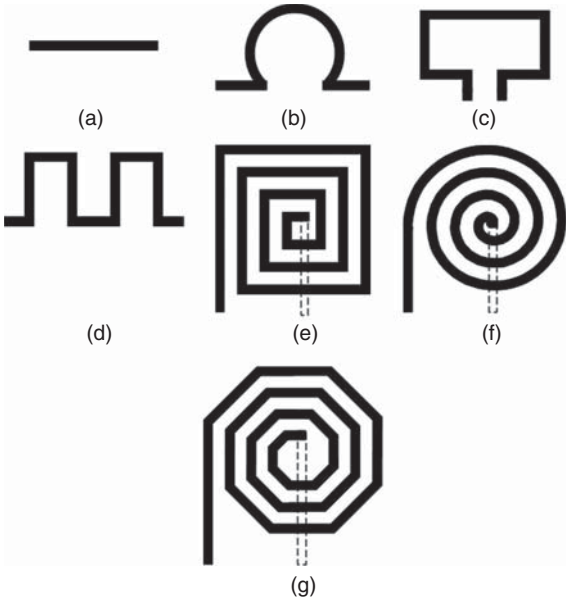


Figure 3.10. Common on-chip inductors: (a) conducting strip, (b) conducting circular loop, (c) conducting rectangular loop, (d) conducting meander strip, (e) rectangular spiral (or multi-loop rectangular), (f) circular spiral, and (g) hexagonal spiral.

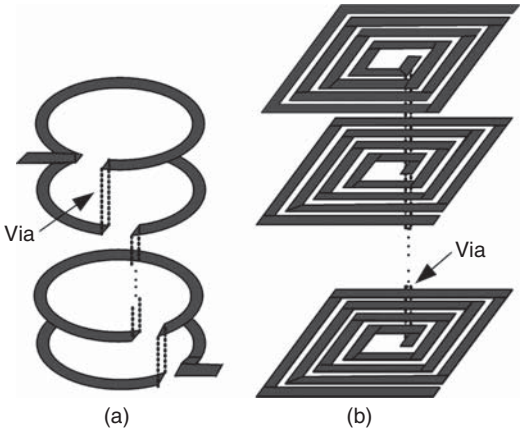


Figure 3.11. Stacked rectangular (a) and circular (b) spiral inductors. The constituent spirals are located on different metal layers and connected through via-holes.

Considering a stack of two identical spirals, the total inductance can be approximately modeled as a series of two inductances at very low frequencies, as shown in Figure 3.12, is given as

$$L_T = L_1 + L_2 + 2M \tag{3.28}$$

where $L_i (i = 1, 2)$ is the individual self inductance of each spiral and M represents the mutual inductance between the spirals, which may be obtained as

$$M = k\sqrt{L_1 L_2} \tag{3.29}$$

where k is the coefficient of coupling between the two spirals. For well-designed stacked inductors, it should be around 0.5–0.7. Assuming very tight coupling ($k \simeq 1$), the total inductance of a stacked spiral of N identical spirals is approximately given at very low frequencies as

$$L_T \simeq N^2 L \tag{3.30}$$

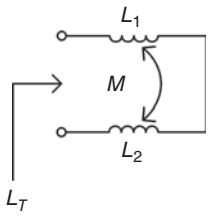


Figure 3.12. Simplified inductance model of two spirals connected in series.

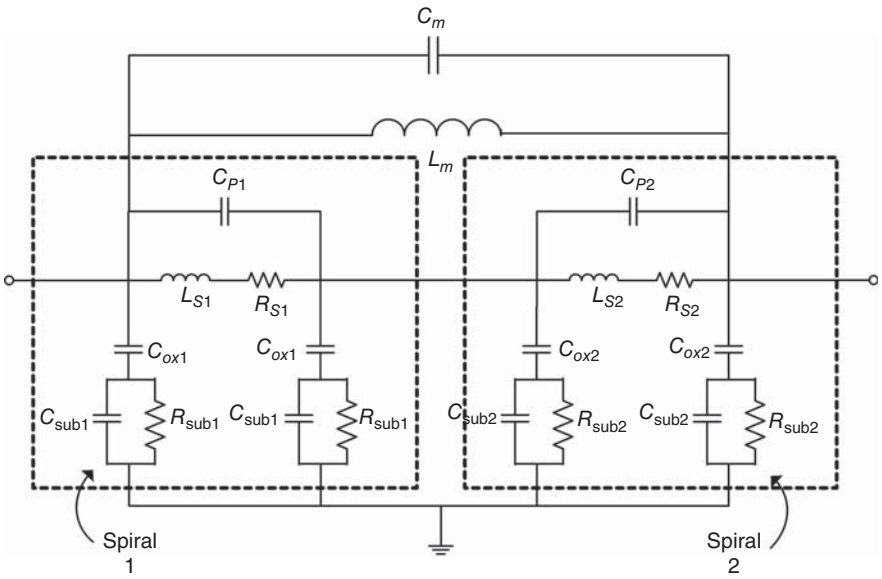


Figure 3.13. Equivalent-circuit model for a two-spiral stacked inductor. L_m and C_m represent the mutual inductance and capacitance between the two spirals; $C_{cxi}(i = 1, 2)$ are the capacitance due to Oxide layers; $C_{\text{sub}i}$ and $R_{\text{sub}i}$ are the capacitances and resistances due to silicon substrate; and L_{si} and R_{si} are the main inductances and resistances of the spirals.

where L is the inductance of each spiral. In practice, however, the total inductance L_T of RFIC stacked inductors is less than what is calculated from Eq. (3.30), primarily due to coupling between individual spirals (i.e., coupling coefficient k), especially at high frequencies. The difference depends on the number of spirals, (vertical) spacing between them (i.e., oxide layers between them), and horizontal locations of metal traces with respect to each other. Nevertheless, in a CMOS process, inductors of large inductance may be realized by stacking several spirals deposited on different metal layers.

An equivalent-circuit model for stacked spirals is much more complicated than that for a single-layer spiral due to models of individual spirals constituting the stacked inductors, coupling between these spirals, additional effects to eddy currents and parasitics, etc. Furthermore, different models should be used for stacked inductors having different numbers of spiral. Figure 3.13 shows a possible equivalent-circuit model for a two-spiral stacked inductor, demonstrating the complexity of stacked spiral's model. Models for stacked spirals are, therefore, best handled by either measurement or EM analysis, in which an equivalent-circuit model can be derived by fitting measured or calculated results (e.g., S-parameters) to those calculated from the equivalent-circuit model. Due to the fitting nature, the values of the model's parameters may not, however, represent the actual values of the stacked spirals. They are thus more useful for the design of RFICs than for the interpretation of the physics of the spirals.

Stacking of spirals, however, has several drawbacks. In CMOS processes, metal layers, separated by thin dielectric layers such as SiO_2, are close to each other, causing large (parasitic) coupling capacitances between them, effectively lower the resonant frequency and hence limiting the operating frequency range. To minimize this effect, metal layers sufficiently apart should be chosen for the spirals to reduce the coupling capacitance

(e.g., metal layer M5 and M3 instead of M5 and M4 in a CMOS process). Additionally, offset between spirals may be used to reduce the coupling effect. Moreover, Q of stacked spirals is in general smaller than that of a single-layer spiral due to additional conduction losses such as that due to finite resistance of the connecting via-holes, dielectric loss, and higher substrate loss (assuming additional spirals are on metal layers below that of the single spiral). The total equivalent coupling capacitance can be approximately estimated at low frequencies as

$$C_{\text{eq}} \simeq \frac{1}{3N^2}\left(4\sum_{i=1}^{N-1} C_i + C_N\right) \tag{3.31}$$

where N is the number of metal layers used and C_i is the coupling capacitance between adjacent spirals, with C_N denoting the coupling between the Nth spiral and silicon substrate.

The resulting resonant frequency of stacked-spiral inductor may be approximated calculated from

$$f_r \simeq \frac{1}{2\pi\sqrt{L_T C_{\text{eq}}}} \tag{3.32}$$

where L_T represents the total inductance of the stacked spiral. As can be seen in Eq. (3.31), the capacitive effect of the upper spirals is much more than that of the lowest spiral. Equation (3.32) shows that the resonant frequency of a stacked spiral, due to mutual coupling between individual spirals, may be too low for the spiral to be useful at certain frequencies. Large inductances are not typically needed for the design of RFICs, except possibly where on-chip RF chokes are needed. Therefore, in view of resultant low resonant frequencies, stacked spiral inductors having spirals directly on top of each other have limited use in RFICs.

Figure 3.14 shows a stacked octagonal spiral inductor on a 0.18-μm CMOS process, and Figure 3.15 compares the calculated Q of different stacked octagonal spiral inductors to those of an (un-stacked) single-spiral inductor. The maximum Q reduces as more layers are used, primarily due to increasing capacitive coupling.

As another example of stacked spiral inductors with offset (or asymmetrical) spirals, which are more useful for RFICs than symmetrically stacked counterparts, we consider four different asymmetrically stacked spiral inductors on a CMOS process profile as shown in Figure 3.16. The metal layers for these inductors are 6 (single layer), 6 and 5 (two layers), 6 and 2 (two layers), and 6, 4, and 2 (3 layers). The individual spirals on a stacked spiral are asymmetrically offset from each other and interconnected by vertical via-holes. Each individual spiral is octagonal with outer diameter of 100 μm, trace-width of 6 μm, trace-spacing of 2 μm, and has four turns. The trace of a spiral is offset from that of an adjacent spiral by 6 μm. Table 3.2 summarizes the calculated resonant frequencies, and maximum Q and corresponding inductances, using the EM simulator IE3D, for these spiral inductors.

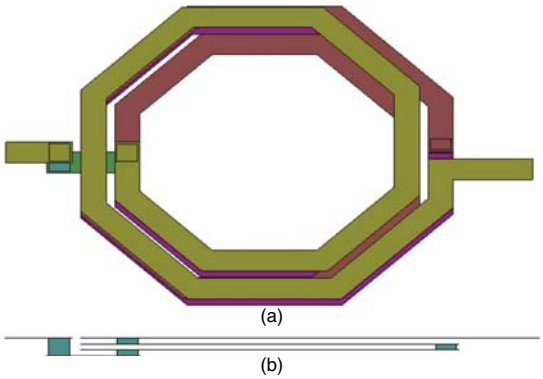


Figure 3.14. Top (a) and cross-sectional (b) of a three-layer stacked octagonal spiral inductor on a 0.18-μm CMOS process. The metal layers are M6 (top layer), M5, and M4. The spiral dimensions in each layer are the same: trace width = 6 μm, trace spacing = 2 μm, inner diameter = 60 μm, outer diameter = 88 μm. The number of turns in each layer is 1.5.

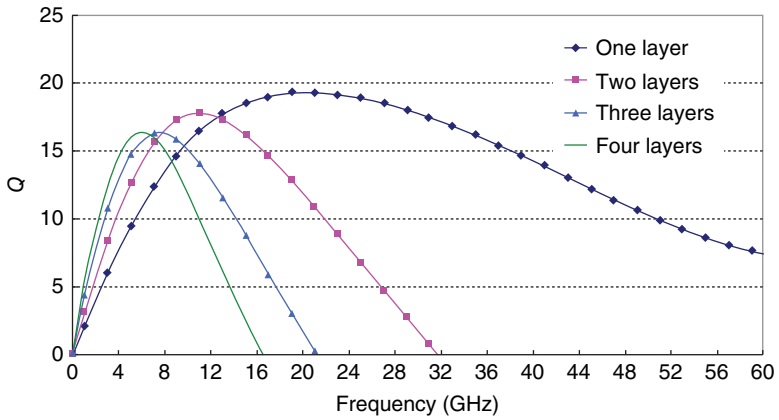


Figure 3.15. Calculated Q of a single octagonal spiral inductor (one layer on M6), two-layer stacked octagonal spiral inductor (M6 and M5), three-layer stacked octagonal spiral inductor (M6, M5, and M4), and four-layer stacked octagonal spiral inductors (M6, M5, M4, and M3) on a 0.18-μm CMOS process.

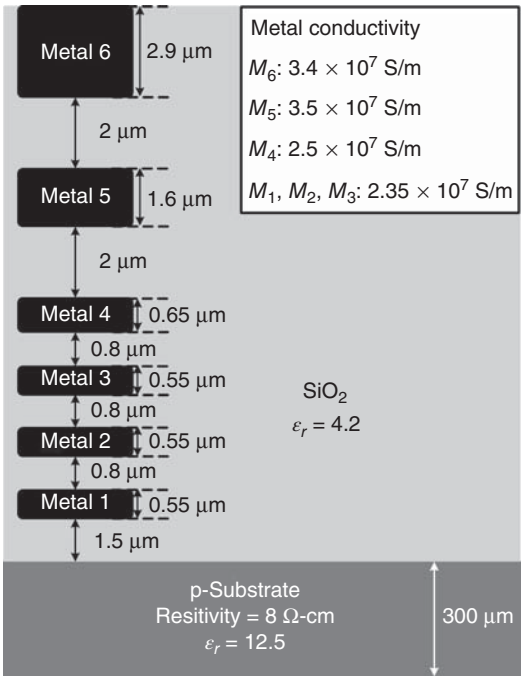


Figure 3.16. CMOS process profile used for the offset stacked spiral inductors.

3.4.2 Loss in Inductors

There are in principle two kinds of loss in inductors: one is the conduction loss due the conductors' metallization and another is the dielectric loss resulting from the dielectrics and silicon substrate beneath the conductors. Due to the complexity of CMOS metal layer structures and of some inductor geometries, it is best to evaluate these losses at RF frequencies using a full-wave EM method or software, particularly considering the fact that there are various good commercially available EM simulators. Nevertheless, closed-form equations for these losses are valuable as they allow initial estimates of the losses as well as provide some insight into the effects and contributions of these losses. While the conduction loss equations are relatively convenient to derive approximately, those due to the dielectric loss are quite complicated to derive.

TABLE 3.2. Calculated Resonant Frequencies, Maximum Q and Inductances of the Offset Stacked Spiral Inductors

Metal layers	Number of turns	Resonant frequency (GHz)	Maximum Q @ frequency (GHz)	Inductance at maximum Q (nH)
6	4	50	15.2 @ 9	1.15
6, 5	4	12.5	9.5 @ 5	4.5
6, 2	4	19.5	6.1 @ 7	4
6, 4, 2	4	9.5	4.2 @ 4	10.1

3.4.2.1 Conduction Loss. The finite metallization of a conducting trace in inductors gives rise to a series resistance, which produces conduction loss along the trace. The total resistance of a conducting trace is given exactly at DC or approximately at low frequencies by the following equation:

$$R_{\text{dc}} = \frac{\ell}{\sigma W t} \tag{3.33}$$

where σ is the trace's conductivity, t is the metallization thickness, W is the trace width, and ℓ is the total length of the trace.

At RF, the skin effect becomes more pronounced, resulting in increase of the series resistance. The series resistance becomes, considering the skin effect, at RF:

$$R_{\text{rf}} \simeq \frac{\ell}{\sigma W \delta_s (1 - e^{-t/\delta_s})} \tag{3.34}$$

where $\delta_s = \sqrt{\frac{2}{\omega\mu\sigma}}$ is the skin depth of conductors. In most RFIC fabrication processes, the metallization thickness is very small, particularly for lower metal layers near silicon substrate (typically <1 μm), hence resulting in high loss. The effect of loss for different metallization thickness is seen in Figure 3.17, which displays the calculated series resistance versus frequency for three aluminum traces of 200-μm long, 5-μm wide, and 0.6-, 1.5-, and 2.5-μm thick.

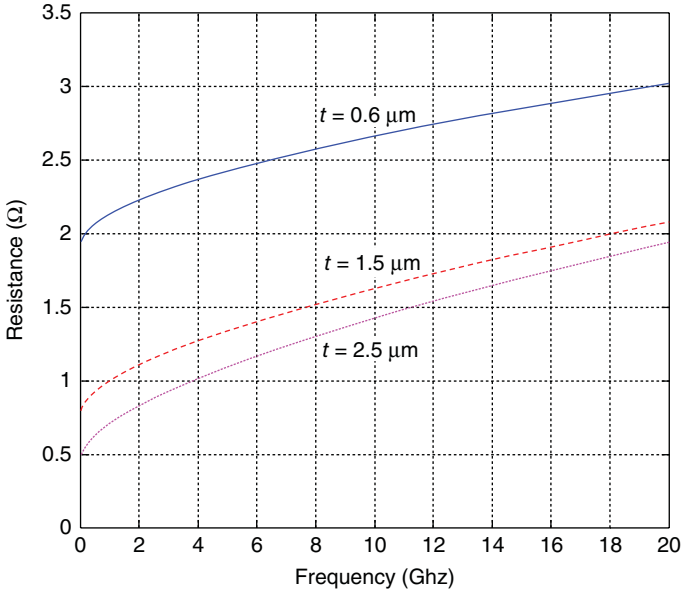


Figure 3.17. Calculated series resistances of aluminum traces.

Another contribution to the conduction loss is the eddy-current resistance. This resistance arises due to the eddy currents[9] flowing on the conducting trace. Eddy currents cause Ohmic power loss to signals traveling along the trace. The eddy-current resistance, R_e, representing the eddy-current loss is a strong function of frequency, as expected from the frequency-dependent time-varying magnetic field, and needs to be included in the total series resistance R_s of the trace as

$$R_s = R_{\text{rf}} + R_e \tag{3.35}$$

The eddy-current resistance increases with increasing frequency and, at a particular frequency, begins to increase significantly, resulting in a large resistance for the trace. This frequency is given for spiral inductors by [3]

$$f_c = \frac{3.1(W+S)}{2\pi\mu_0 W^2} R_{\text{sh}} \tag{3.36}$$

where W and S are the trace width and spacing, respectively, and

$$R_{\text{sh}} = \frac{1}{\sigma\delta_s} \tag{3.37}$$

is the surface resistivity or sheet resistance (in ohms or ohms per square as frequently used) of the trace conductor. The total series resistance of the trace in spiral inductors at RF frequencies can be approximately given, considering eddy-current effects, as [3]

$$R \simeq R_{\text{dc}} \left[1 + 0.1\left(\frac{f}{f_c}\right)^2\right] \tag{3.38}$$

It is noted that for a finite area of conductor, such as conducting trace in spirals, the surface resistivity is obtained by multiplying R_{sh} by the length and dividing by the width of the conductor, since the width elements are essentially in parallel.

3.4.2.2 Dielectric Loss. Silicon substrate contributes the most dielectric loss in inductors, while SiO_2 dielectrics produce a very small loss. To improve the yield of RFICs, heavily implanting processes are normally implemented in the silicon substrate before the oxide and metal layers above it are formed. The silicon substrate's resistivity is thus partly dependent on the density of the implanted electrons. Resistivity of silicon substrates is normally available from foundries for different processes. Most CMOS processes use silicon substrates having resistivity from 0.1 to 100 Ω·cm. These low resistive silicon substrates cause both electric and magnetic losses.

When a time-variant current flows through the inductor's trace, both the electric and magnetic fields are excited. The electric field penetrates the oxide layers beneath the trace and goes through the conductive silicon substrate, resulting in electrical loss in the substrate, which can be named "substrate electrical loss." The magnetic field upon penetrating the silicon substrate incites opposite-flowing eddy currents within the substrate, causing magnetic loss, namely "substrate magnetic loss." The eddy currents, in turn, induce a secondary magnetic field, which consequently causes reduction in the primary magnetic field generated by the current flowing along the inductor's trace. The resultant electric and magnetic energy are both dissipated in the silicon substrate, thus causing considerable substrate loss, particular at high RF frequencies. This loss mechanism is very complex to be explicitly formulated using closed-form equations.

[9]Time-varying magnetic field induces EMF according to Faraday's law, which in turn produces local currents normal to the magnetic field, called eddy currents, in the conducting trace.

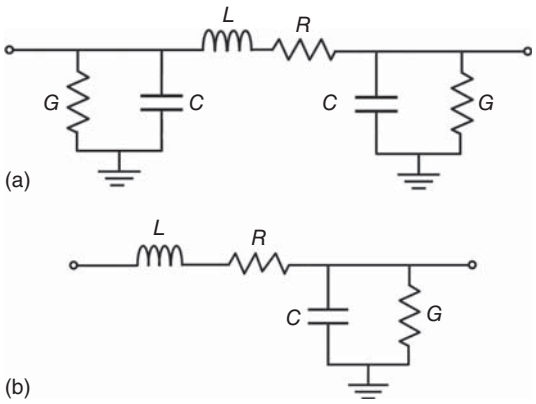


Figure 3.18. (a, b) Typical simplified equivalent-circuit models for simple inductors, as shown in Figure 3.10(a)–(d), on a single metal layer.

3.4.3 Equivalent-Circuit Models of Inductors

3.4.3.1 Strip Inductors. Simple inductors, using single conducting strips on a single metal layer, as shown in Figure 3.10(a)–(d), are relatively easy to model. They may be modeled as a combination of series resistor and inductor, and shunt conductor and capacitor as shown in Figure 3.18. The series resistor represents loss due to finite conductivity of the metal; the series inductor represents the main inductance; the shunt conductance represents loss due to SiO_2 dielectrics and silicon substrate underneath the inductor; and the shunt capacitor accounts for capacitive effects of the dielectrics and silicon substrate.

Parameters of the equivalent-circuit model shown in Figure 3.18(a) can be approximately calculated for the conducting strips shown in Figure 3.10(a)–(d) using the equations presented in [4, 5]. These equations assume the strips are located above a ground plane, which resemble microstrip lines, and are given below:

$$L(\text{nH}) \simeq 2 \times 10^{-4} \ell \left[\ln \left(\frac{\ell}{W+t} \right) + 1.193 + \frac{W+t}{3\ell} \right] K_g \tag{3.39}$$

$$R(\Omega) \simeq \frac{K_1 R_{\text{sh}} \ell}{2(W+t)} \tag{3.40}$$

$$C(\text{pF}) \simeq \frac{16.67 \times 10^{-4} \ell \sqrt{\varepsilon_{re}}}{Z_o} \tag{3.41}$$

$$G = 0 \tag{3.42}$$

$$K_g = \begin{cases} 0.57 - 0.145 \ln \dfrac{W}{h}, & \dfrac{W}{h} > 0.05 \\ 1, h \to \infty \qquad \text{(no ground plane)} \end{cases} \tag{3.43}$$

$$K_1 = 1.4 - 0.217 \ln \left(\frac{W}{5t} \right), \quad 5 < \frac{W}{t} < 100 \tag{3.44}$$

where all dimensions are in microns (μm); Z_o and ε_{re} are the characteristic impedance and effective dielectric constant of the corresponding microstrip line,[10] W is width of the microstrip line (or conducting strip); t and ℓ are the metallization thickness and total length of the strip, respectively; R_{sh} is the sheet resistance of the conductor; and h is the distance between the strip and ground plane, which may include thickness of the dielectrics and silicon substrate beneath the strip. For aluminum, which is typically used in RFIC fabrication processes, $R_{\text{sh}} = 3.26 \times 10^{-7} \sqrt{f}$.

[10] Z_o and ε_{re} of microtrip line can be determined using formulas or methods presented in Chapter 4.

Possible coupling between different parts within the circular loop, rectangular loop and meander line is neglected in the equations for the model's parameters. Coupling effects are more pronounced, making the model of the inductors become more complex, as the spaces between different parts reduce, particular at high frequencies. As the ground plane is removed, inductance increases and the factor K_g is approximately equal to 1. Losses due to the dielectrics and silicon substrate are not considered in Eqs. (3.39)–(3.44), and so the conductance G is assumed to be equal to zero. However, G for the silicon substrate is not small, especially at high RF, causing significant substrate loss. Therefore, in practice, G in the models presented in Figure 3.18 need to be evaluated and used to account for the dielectric and substrate losses. These losses, nevertheless, can be reduced substantially – and hence may be neglected, particularly at low frequencies – when some kind of isolation, such as solid/patterned metal or polysilicon shield (to be discussed later) is provided between the strips and silicon substrate. It should be noted that values obtained from Eqs. (3.39)–(3.44) can also be used as initial values in an optimization fitting process to determine more accurate values for the model by matching the S-parameters of the model to those measured or accurately calculated using a full-wave EM method or commercially available program.

3.4.3.2 Spiral Inductors. The equivalent circuit model of a spiral inductor, the model's element values, and the inductor's operating frequency range, Q and loss depend on the trace width (W), trace spacing (S), turn number (n), trace metallization thickness (t), the inner-most dimension (b_i), properties of the dielectrics and substrate beneath the spiral, and possible shield between the inductor and silicon substrate. Several of these parameters are completely controlled by RFIC designers. Other parameters depend not only on engineering judgment but also on a particular CMOS process.

Simple Models

Consider a rectangular, octagonal, and circular spiral as shown in Figure 3.19. The simplest equivalent-circuit model for spiral inductors consists of an inductor and resistor connected in series, as shown in Figure 3.20(a). Very simple equations were derived for the inductance. For a rectangular spiral inductor, the inductance can be approximately calculated by

$$L_r \simeq \mu_0 n^2 a = 4\pi \times 10^{-7} n^2 a \simeq 1.2 \times 10^{-6} n^2 a \tag{3.45}$$

where L is the inductance in henry (H), n is the number of turns, and $2a$ is the maximum dimension of the spiral in meter (m). This equation is not accurate – typically within 30% of the correct inductance. The inductance L for spirals of other shapes can be determined based on that of the rectangular spiral using the following formula:

$$L \simeq L_r \sqrt{\frac{A}{A_r}} \tag{3.46}$$

where L_r is the rectangular spiral's inductance, and A and A_r represent the areas of that spiral and the rectangular spiral. Using Eq. (3.46) for the circular and octagonal spirals, we obtain the respective inductance as

$$L_c \simeq L_r \sqrt{\frac{\pi}{4}} \tag{3.47}$$

and

$$L_o \simeq 0.91 L_r \tag{3.48}$$

These equations are derived assuming the maximum dimension of the spirals is still $2a$.

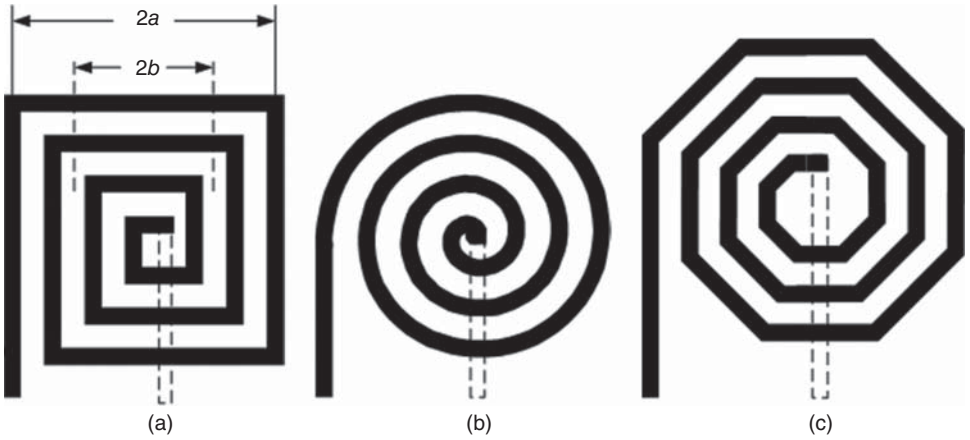


Figure 3.19. Rectangular (a), circular (b), and octagonal (c) spiral.

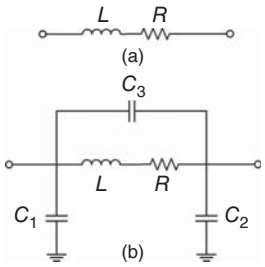


Figure 3.20. (a, b) Simple equivalent-circuit models of spiral inductors.

The number of required turns for a rectangular spiral can be approximately determined as [6]

$$n \simeq \left(\frac{PL_r}{\mu_0}\right)^{1/3} = \left(\frac{PL_r}{1.2\times 10^{-6}}\right)^{1/3} \tag{3.49}$$

where P is the winding pitch in turns per meter.

A still simple yet relatively accurate equation for calculating the inductance of a rectangular spiral is given by the Wheeler's equation [6, 7]:

$$L_r \simeq \frac{37.5\mu_0 n^2 b^2}{22a - 14b} \tag{3.50}$$

where $2a$ is the maximum dimension and $2b$ is the mean dimension of the rectangular spiral, defined as the distance between the middles of the windings, as seen in Figure 3.19(a). This equation is much more accurate than Eq. (3.43).

A more accurate model for circular spiral inductors (Figure 3.21) placed on top of a conductor-back substrate is shown in Figure 3.20(b). The loss due to the silicon substrate is neglected in this model. The model's parameters can be estimated using the following equations [5]:

$$L(\text{nH}) \simeq 0.03937\frac{a^2 n^2}{8a + 11c}K_g \tag{3.51}$$

$$R(\Omega) \simeq \frac{K_2 \pi a n R_s}{W} \tag{3.52}$$

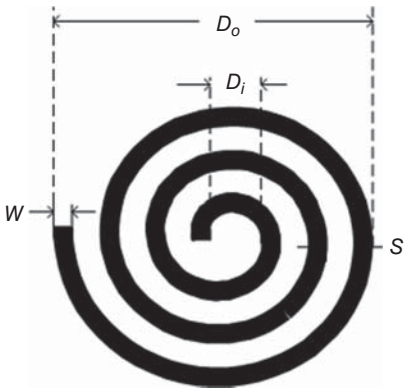


Figure 3.21. Circular spiral inductor.

$$C_1(\text{pF}) \simeq \frac{16.67 \times 10^{-4} \ell \sqrt{\varepsilon_{\text{re}}}}{Z_o} \tag{3.53}$$

$$C_2(\text{pF}) \simeq 3.5 \times 10^{-5} D_o + 0.06 \tag{3.54}$$

$$a = \frac{D_o + D_i}{4} \tag{3.55}$$

$$c = \frac{D_o - D_i}{2} \tag{3.56}$$

$$K_g = \begin{cases} 0.57 - 0.145 \ln \dfrac{W}{h}, & \dfrac{W}{h} > 0.05 \\ 1, & h \to \infty \text{ (no ground plane)} \end{cases} \tag{3.57}$$

$$K_2 = 1 + 0.333 \left(1 + \frac{S}{W}\right) \tag{3.58}$$

where all dimensions are in microns; n is number of turns; D_i and D_o are the inside and outside diameters, respectively; W is the width of the trace; h is the substrate thickness; and S is the spacing between two adjacent traces. To minimize the effect of the substrate's bottom conducting plane, which reduces the inductance value, a circular spiral should be designed to have $S < W$ and $S \ll h$. The inductance may also be estimated using the following simple equation derived by Wheeler [7]:

$$L(\text{nH}) = \frac{n^2 a^2}{8a + 11c} \tag{3.59}$$

where all dimensions are in μm. Equation (3.51) is basically a modification of (3.59), taking into account the presence of a conducting plane in the back of the silicon substrate.

The inductance value for a circular spiral, as well as square, hexagonal and octagonal spirals, can also be determined from the following equations [5, 8, 9]:

$$L = \frac{\mu_o n^2 D_{\text{av}} c_1}{2} \left[\ln \left(\frac{c_2}{\rho} \right) + c_3 \rho + c_4 \rho^2 \right] \tag{3.60}$$

$$\rho = \frac{D_o - D_i}{D_o + D_i} \tag{3.61}$$

$$D_{\text{av}} = \frac{D_o + D_i}{2} \tag{3.62}$$

where $c_i (i = 1, 2, 3)$ is given in Table 3.3 for different shapes of spirals.

Other Models

Figure 3.22 shows some other more commonly used equivalent-circuit models of rectangular spirals for RFIC design. These models are derived based on the physics of the spiral and take into account the effects of the dielectric layers beneath the spiral and substrate losses, including losses due to the substrate itself and the induced eddy currents flowing in the substrate, and thus describe the behavior of the rectangular spiral better than those presented earlier. Some analytical equations were also derived for some model parameters. These equations can be utilized to obtain more accurate equivalent models through matching between their calculated *S*-parameters and those obtained from measurement or an accurate EM simulator. Understanding

TABLE 3.3. Values of $c_i (i = 1, 2, 3)$ for Different Spiral Geometries

Spiral geometry	C_1	C_2	C_3	C_4
Square	1.27	2.07	0.18	0.13
Hexagonal	1.09	2.23	0.00	0.17
Octagonal	1.07	2.29	0.00	0.19
Circular	1.00	2.46	0.00	0.20

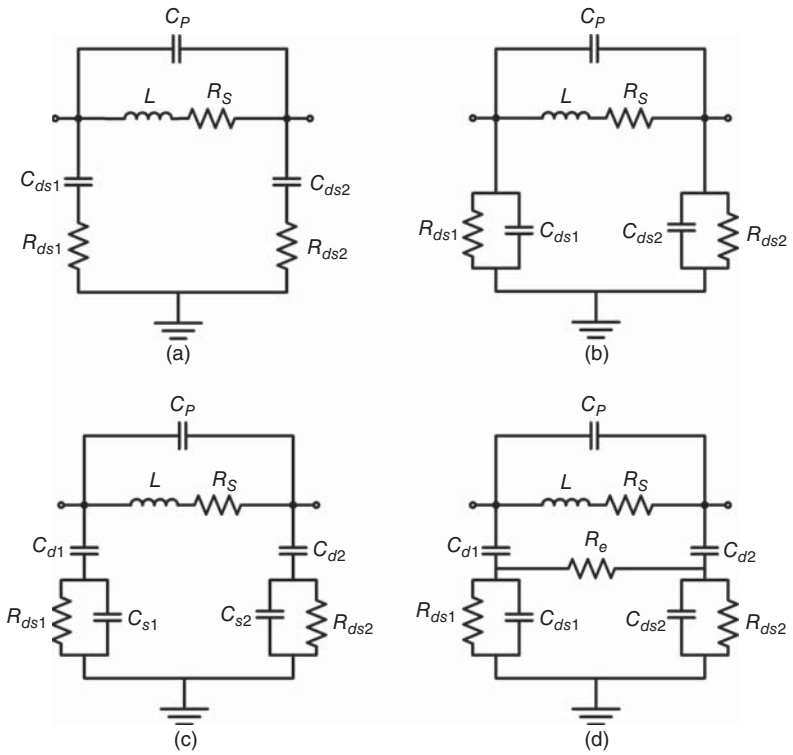


Figure 3.22. (a–d) Four commonly used models for rectangular spirals in RFICs. The ground represents the physical RF ground of the substrate (or bulk).

the physics of spiral inductors and their electrical effects allows models to be accurately derived. Parameters of these models are described as follows:

L = principal inductance of the inductor.

R_s = series resistance of the inductor's conductors, representing losses due to finite conductivity, skin effect and eddy currents.

C_p = coupling capacitance between the inductor's terminals, resulted from the couplings between the cross-under metal and the spiral (due to parallel-plate capacitances) and between the spiral's adjacent turns (due to fringing capacitances between them).

$C_{ds1,2}$ = capacitance of the oxide dielectrics and silicon substrate. $C_{ds1} \neq C_{ds2}$ in general due to the asymmetry of the spiral.

$C_{s1,2}$ = capacitance of the silicon substrate. $C_{s1} \neq C_{s2}$ in general due to the asymmetry of the spiral.

$R_{ds1,2}$ = resistance representing losses due to the oxide dielectrics and substrate (due to currents flowing into the dielectrics and substrate and eddy currents flow in the substrate magnetically induced by current on the inductor's traces). In Figure 3.22(d), the eddy-current resistance is separated from these resistances as R_e. $R_{ds1} \neq R_{ds2}$ in general.

$C_{d1,2}$ = capacitance due to the oxide dielectrics between the spiral and silicon substrate. $C_{d1} \neq C_{d2}$ in general.

R_e = resistance representing loss to the eddy currents flow in the silicon substrate. R_e is absorbed into $R_{ds1,2}$ as modeled in Figure 3.22(a)–(c).

Analytical equations for the parameters of the model shown in Figure 3.22(c) are given as [10, 11]:

$$R_s \simeq \frac{\ell}{W\sigma\delta_s(1 - e^{-t/\delta_s})} \tag{3.63}$$

where ℓ is the total length of the (winding) trace.

$$C_p \simeq mW^2\frac{\varepsilon_d}{h_{d1}} \tag{3.64}$$

where m is the number of overlap between the spiral and the cross-under metal; ε_d and h_{d1} represent the dielectric constant and thickness of the dielectric between the cross-under metal and the spiral. Here, the capacitive coupling between adjacent turns are assumed to be negligible as compared to that between the spiral and underpass metal. $C_{di}(i = 1, 2)$ is modeled as a half of the parallel-plate capacitance between the spiral and substrate, approximated as two conductors of area $W\ell$ separated by a dielectric of ε_d and h_{d2}, as

$$C_{d1} \simeq C_{d2} \simeq \frac{1}{2}W\ell\frac{\varepsilon_d}{h_{d2}} \tag{3.65}$$

where h_{d2} represents the thickness of the SiO_2 dielectric layers between the spiral and Si substrate. The resistance R_{dsi} and capacitance $C_{dsi}(i = 1, 2)$ are approximately proportional to the area occupied by the inductor and hence can be obtained as

$$R_{ds1} \simeq R_{ds2} \simeq \frac{2}{W\ell G_{\text{sub}}} \tag{3.66}$$

where G_{sub} is a fitting parameter having a dimension of conductance per unit area. It is a constant for a given substrate material and the distance of the spiral to substrate, and is typically approximated as 10^{-7} S/μm^2.

$$C_{s1} \simeq C_{s2} \simeq \frac{1}{2}W\ell C_{\text{sub}} \tag{3.67}$$

where C_{sub} is also a fitting parameter having a dimension of capacitance per unit area, typically between 10^{-3} to 10^{-2} fF/μm^2. The values for C_{ds1} and C_{ds2} can be estimated as those calculated for C_{s1} and C_{s2}, respectively. These equations may also be used to estimate values for the parameters of other models shown in Figure 3.20. The eddy-current resistance R_e included in the model shown in Figure 3.22(d) may be neglected for frequency up to around 3 GHz [12].

It should be noted that the inductance L of the spiral inductors seen in the equivalent-circuit models of Figure 3.22 is actually smaller due to the mutual inductive coupling between the spirals and silicon substrate. This is due to the fact that the eddy currents induced in the silicon substrate cause a magnetic field to weaken the primary field produced by the current flowing in the inductor, effectively reducing the inductance to [13]:

$$L' = L - M \tag{3.68}$$

where M is the resultant mutual inductance between the inductor's trace and the silicon substrate, as illustrated in Figure 3.23.

3.4.4 Resonance in Inductors

A simplified equivalent-circuit model for inductors is shown in Figure 3.24(a), in which L represents the actual inductance of the inductors, R represents loss in the inductors, and C represents parasitic capacitance, such as capacitive effects due to dielectric layers and silicon substrate. Due to the existence of resistor, inductor, and capacitor in this model, it is expected that resonance(s) would occur. To illustrate this, an equivalent parallel circuit is derived and shown in Figure 3.24(b), which comprises resistance R_e, inductance L_e, and capacitance C_e. This electrically equivalent circuit inherently possesses a self-resonant frequency, at which it behaves as

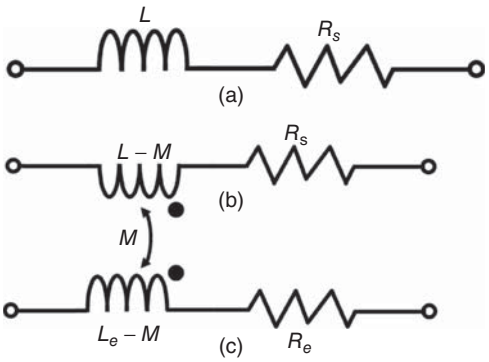


Figure 3.23. (a) Original inductance and resistance of spiral inductors and (b) modified electrical representation taking into account the mutual inductive coupling between the inductor's trace and silicon substrate. L_e and R_e represent the inductance of the silicon substrate and the eddy-current loss in the substrate.

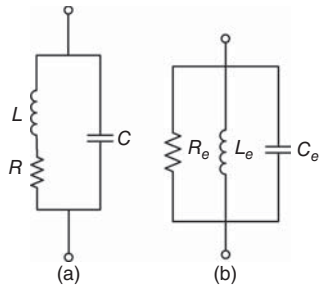


Figure 3.24. Simplified equivalent circuit of an inductor (a) and its equivalence (b).

a parallel resonator, of

$$f_r = \frac{1}{2\pi\sqrt{L_e C_e}} \tag{3.69}$$

thus limiting the operating bandwidth of the inductor to frequency $f < f_r$. Only within this operating frequency range, the inductor behaves dominantly as an inductor.[11]

It should be noted that a resonance occurs if and only if the imaginary part of the impedance of the inductor's equivalent circuit is equal to zero or the impedance is equal to a (pure) resistance. In general, an inductor behaves as a parallel resonator. At high frequencies, an inductor may also have a series resonance which, however, has no practical use. At $f < f_r$, it is an inductor. However, at $f > f_r$, it becomes a capacitor. At frequencies much higher than the inductor's resonant frequency, the inductor's dimensions may be comparable to the operating wavelength, which effectively prohibits the inductor to be classified as a lumped element. At these frequencies, it is considered a distributed element and has multiple resonant frequencies, which are inherent for distributed structures.

3.4.5 Quality Factor of Inductors

Below its resonant frequency and within certain low frequencies, where the inductor is assumed to have a dominantly inductive value, the inductor's quality factor (Q) is approximately given by the usual formula

$$Q = \frac{\omega W_m}{P_L} \tag{3.70}$$

In practice, however, inductors have parasitic capacitances. These parasitic capacitors do not store energy – instead, working against the inductors to reduce the total energy stored within the inductors to[12] ($W_m - W_e$). Therefore, the Q of an inductor at frequencies lower than the resonant frequency is more accurately given by

$$Q = \frac{\omega(W_m - W_e)}{P_L} \tag{3.71}$$

where W_m denotes the time-average magnetic energy stored in the inductor, W_e denotes the time-average electric energy corresponding to all the parasitic capacitors, and P_L represents all the power loss or dissipated in the inductor and parasitic capacitors including those resulting from the conductor and substrate losses. Equation (3.71), however, suggests that Q would equal to zero at the self-resonant frequency of the inductor, where the total energy stored by the inductor vanishes. This frequency actually implies the transition frequency at which the inductor becomes a pure resistor and above which capacitive effect becomes more pronounced. The inductor then becomes more of a capacitor at frequencies greater than the resonant frequency and Eq. (3.71) is no longer valid.[13] Therefore, Eq. (3.71) can be used only at frequencies below the inductor's resonant frequency. The zero quality factor at resonance has no significance and usefulness in circuit design and performance since, at the resonant frequency, the inductor ceases to operate inductively and is no longer useful for circuit design. However, if the inductor is used as a resonator at that frequency, then the total energy stored in that resonator would be positively contributed from both the magnetic energy and electric energy stored within the inductor and (parasitic) capacitor, hence leading to a non-zero Q.

As for other RF structures, the Q of inductors fabricated in a silicon process is frequency dependent. Figure 3.25 shows typical variation of the Q of spiral inductors in CMOS processes versus frequency. At very low frequencies, the electric energy stored within the parasitic capacitors is negligible and the total stored energy is dominated by the magnetic energy. The inductor at these frequencies can thus be described simply

[11] The actual inductance of this inductor is not the equivalent inductance L_e.

[12] In contrast, the total energy stored in a resonator is constructively contributed by both electric and magnetic energies stored within the constituting capacitor and inductor, respectively, and is thus equal to the sum of these energies.

[13] Above the resonant frequency, where the inductor behaves dominantly as a capacitor, the Q is given as $Q = \omega(W_e - W_m)/P_L$.

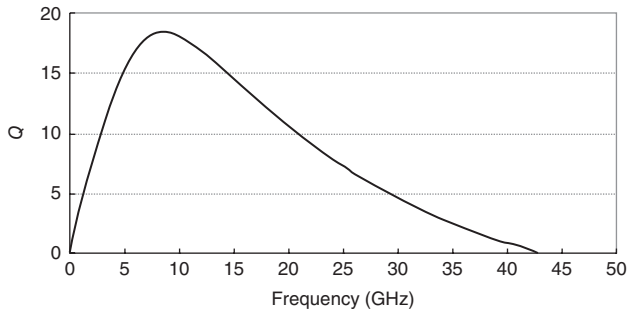


Figure 3.25. Variation of Q of spiral inductors. The results are calculated using the EM simulator IE3D for a 0.68-nH octagonal spiral inductor, having 2 turns, 144-μm^2 area, 15-μm trace width, and 2-μm trace spacing, on the topmost metal layer of a 0.18-μm CMOS process.

as a series of a pure inductor and resistor, as mentioned earlier. As frequency is increased, the stored electric energy increases accordingly, hence reducing Q, which can also be seen from Eq. (3.71). More power loss also results with increasing frequency, further reducing Q.

Another useful equation for calculating Q of inductors, or any passive elements including capacitors, can be derived based on the impedance of the elements and the Q formula generally defined as the ratio between the total time-average stored energy and the time-average power loss or dissipated within the elements. We consider a passive element, which is assumed to be modeled by any equivalent circuit. The impedance[14] of that element can be expressed as $Z = R + jX$ regardless of its equivalent circuit model. The power loss in the element is caused by the resistance R and the energy supplied by an external source is stored by the reactance X. It is noted that X is essentially an overall equivalent reactance of the element, and so the energy stored by X is actually resulted from the energy stored in the magnetic and electric fields associated with the internal inductive and capacitive elements existing in the inductor, respectively. The maximum energy stored in these individual inductive or capacitive elements may be larger than that stored by X. Moreover, there might be some energy exchange between these internal L and C, which are not accounted for in the energy stored by X. The positive and negative values of the instantaneous power of an external source exciting the element represent the transfer of energy back and forth between the source and the element. Without loss of generality, we assume a periodic sinusoidal excitation. The element stores and releases energy, via its reactance, alternately every quarter of a cycle during which the instantaneous power is positive and negative, respectively. The total energy stored in the element can be expressed as

$$W = \int_{t_0}^{t_0+\frac{T}{4}} v(t)i(t)dt = \int_{t_0}^{t_0+\frac{T}{4}} jXI_0 e^{j\omega t} \cdot I_0 e^{j\omega t} dt$$
$$= jXI_0^2 \int_{t_0}^{t_0+\frac{T}{4}} e^{j2\omega t} dt \tag{3.72}$$

where T is the period, t_0 denotes the starting time of the energy storage, and I_0 is the maximum current amplitude.

For inductors, X is positive and the energy storing occurs during the time intervals from $T/4$ to $T/2$ and $3T/4$ to T. Therefore, the stored energy can be written as

$$W = \int_{T/4}^{T/2} v(t)i(t)dt = jXI_0^2 \int_{T/4}^{T/2} e^{j2\omega t} dt$$

$$= \frac{XI_0{}^2}{2\omega}(e^{j\omega T} - e^{j\omega T/2})$$

[14]This is the impedance looking into the two terminals of the element.

$$= \frac{XI_0^2}{2\omega} e^{j\omega T/2}(e^{j\omega T/2} - 1)$$

$$= \frac{jXI_0^2}{\omega} e^{j3\omega T/4}\left(\frac{e^{j\omega T/4} - e^{-j\omega T/4}}{2}\right)$$

$$= \frac{jXI_0^2}{\omega} e^{j3\omega T/4}\sin(\omega T/4)$$

$$= \frac{jXI_0^2}{\omega} e^{j3\pi/2}\sin(\pi/2)$$

$$= \frac{XI_0^2}{\omega} \tag{3.73}$$

The average power loss is given by

$$P_L = RI_0^2 \tag{3.74}$$

Q of inductors can then be obtained from Eqs. (3.73) and (3.74) as

$$Q = \frac{\omega W}{P_L} = \frac{X}{R} \tag{3.75}$$

or

$$Q = \frac{\text{Im}(Z)}{\text{Re}(Z)} \tag{3.76}$$

which can also be written as

$$Q = -\frac{\text{Im}(Y)}{\text{Re}(Y)} \tag{3.77}$$

where $Y = 1/Z$ is the admittance of the element. The Q of inductors can therefore be exclusively calculated from their impedance or admittance, which can be determined from their equivalent-circuit models typically used in RFIC design. As an example, we consider a two-port π-network representing an inductor, as shown in Figure 3.26(a), the impedance and admittance of the inductor are determined as the respective impedance and admittance looking into the inductor's terminals seen in Figure 3.26(b) as

$$Y = \frac{1}{Z} = \frac{Y_1Y_2 + Y_1Y_3 + Y_2Y_3}{Y_1 + Y_3} \tag{3.78}$$

from which, the inductor's Q can be obtained according to (3.76) or (3.77).

The Q of an inductor (or in fact any passive elements such as capacitor) can also be determined using the impedance (Z_{11}) or admittance (Y_{11}) matrix parameter of its two-port representation. We assume an inductor is described by its impedance Z or admittance Y, which may be represented by an equivalent circuit consisting of a single or multiple elements [e.g., a three-element π network shown in Figure 3.26(a)], in two-port networks as shown in Figure 3.27. For the series connection of Figure 3.27(a), the admittance-matrix parameter Y_{11} of the two-port network can be obtained as

$$Y_{11} = \left.\frac{I_1}{V_1}\right|_{V_2=0} \tag{3.79}$$

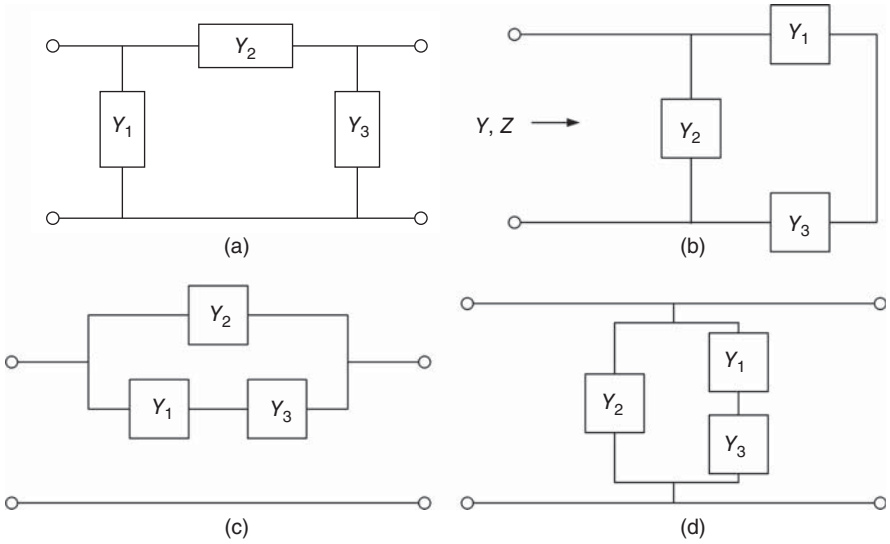


Figure 3.26. Two-port π-network representing an inductor (a), its equivalent one port used for determining the inductor's impedance and admittance (b), and its series (c) and shunt (d) connection in a two-port network. Y_1 and Y_3 normally consist of a resistor and capacitor connected in parallel; $Y_1 = Y_3$ for symmetrical structures and $Y_1 \neq Y_3$ for asymmetrical structures such as stacked spiral inductors with constituent spirals located on different metal layers.

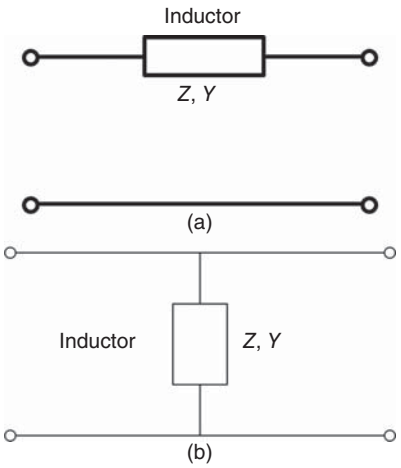


Figure 3.27. Inductor represented by impedance Z or admittance Y connected in series (a) and shunt (b) in a two-port network.

which is essentially the admittance Y of the inductor, and hence, upon using (3.77), we get

$$Q = -\frac{\mathrm{Im}(Y_{11})}{\mathrm{Re}(Y_{11})} \tag{3.80}$$

For the shunt connection seen in Figure 3.27(b), the impedance-matrix parameter Z_{11} of this network is

$$Z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0} \tag{3.81}$$

This impedance is equal to the impedance Z of the inductor. Therefore, the Q of the inductor can be obtained from (3.76) as

$$Q = \frac{\mathrm{Im}(Z_{11})}{\mathrm{Re}(Z_{11})} \tag{3.82}$$

The Q of the inductor in the series [Figure 3.27(a)] or shunt [Figure 3.27(b)] configuration cannot be calculated using Z_{11} or Y_{11}, respectively, since Z_{11} or Y_{11} obtained by open- or short-circuiting the corresponding two-port network does not represent the impedance or admittance of the inductor, respectively.

The use of (3.80) or (3.82) to calculate the Q of inductors (or other passive elements such as capacitors) depends on the way their equivalent-circuit representations connected in the two-port network. To illustrate this, we consider again the two-port equivalent circuit of an inductor, as shown in Figure 3.26(a). In order to calculate the Q of such inductor based on the two-port element Y_{11} or Z_{11}, we connect the equivalent circuit in series or shunt in a two-port network as shown in Figure 3.26(c) and (d), which electrically resemble Figure 3.27(a) and (b), respectively. The Q of the inductor can therefore be calculated using (3.80) and corresponding Y_{11} or (3.82) and corresponding Z_{11}.

It is particularly noted that the Q of inductors (or any passive elements) that are represented by a two-port equivalent circuit, such as that in Figure 3.26(a), should not be determined using Eq. (3.80) or (3.82) with Y_{11} or Z_{11} being the element of the admittance or impedance matrix of the two-port equivalent circuit with Port 2 short-circuited, as normally done in practice, since doing so would remove the Q contribution from Y_3. While this practice of short-circuiting Port 2 for calculating Q is fine in practical inductors with the resistance and capacitance of Y_3 being such large and small that can be neglected, respectively, it is not theoretically correct and not practically accurate when these resistance and capacitance cannot be neglected – for instance, in spiral inductors located close to the silicon substrate possibly encountered in multistack inductors. To illustrate this point, we consider an inductor modeled by an equivalent circuit as shown in Figure 3.28(a) and calculate the Q of the inductor versus frequency using (3.76) and its impedance Z, and using (3.82) and Z_{11} with and without Port 2 short-circuited. As expected and seen in Figure 3.28(b), the results from (3.76) and (3.82) with Port 2 not short-circuited to ground are identical and more accurate than that using (3.82) with Port 2 short-circuited.

At the self-resonant frequency of the inductor, X is equal to zero, leading to zero quality factor by Eqs. (3.75)–(3.77), (3.80), and (3.82). This result is the same as that obtained from the formula for Q given in Eq. (3.71). Eqs. (3.75)–(3.77), (3.80), and (3.82) can thus be used only for frequencies below the resonant frequency. Another point worth mentioning is that the Q formulas (3.75)–(3.77), (3.80), and (3.82) are derived using the "equivalent" resistance and reactance (or susceptance) represented by an inductor, which are, in general, a combination of the actual inductance, capacitance, and resistance exhibited by the inductor. These

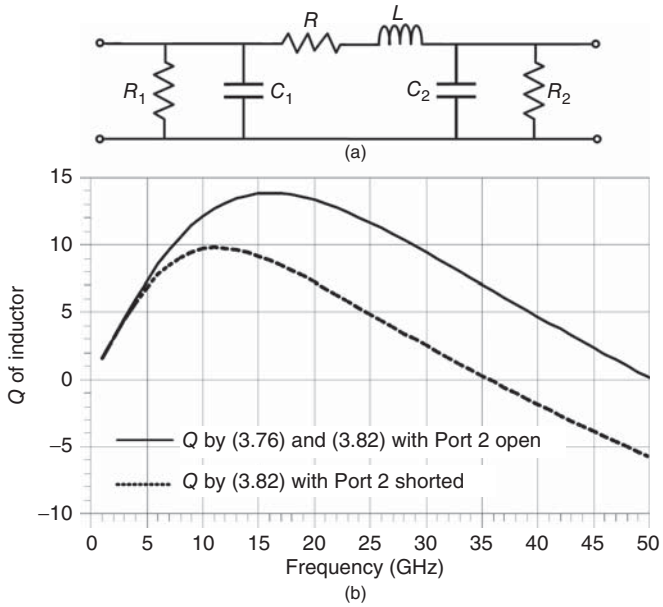


Figure 3.28. Equivalent-circuit model of an inductor with $R = 4\ \Omega$, $L = 1\,\text{nH}$, $R_1 = R_2 = 2\ \text{k}\Omega$, $C_1 = C_2 = 0.02\ \text{pF}$ (a) and calculated inductor Q (b). For simplicity, we assume the model's parameters are independent of frequency.

resistance and reactance (or susceptance) are different from the actual ones produced by the inductor itself, particularly at high frequencies. Therefore, Eqs. (3.75)–(3.77), (3.80), and (3.82) should be used with care and their accuracy needs to be judged depending on particular inductors and operating frequencies. Nevertheless, at frequencies much lower than the resonant frequency, the inductor should behave dominantly as an inductor and so these equations should be relatively accurate. It is especially reminded that the Q discussed in this section represents the intrinsic or unload Q of an inductor itself, regardless of the surrounding elements and/or circuits. Therefore, no external elements should be included in the evaluation of this Q – for instance, Y_{11} of a two-port inductor model as shown in Figure 3.26(c) should not be determined with a 50-Ω termination at Port 2 and used in calculating Q from Eq. (3.80).

It is noted that, while equivalent-circuit models of inductors are useful for circuit design and resonance calculations, they may not provide accurate results for Q.[15] For calculating accurate Q, the actual inductance, capacitance, and resistance of inductors need to be determined accurately.

Q of inductors plays a critical role in the performance of RFICs, such as oscillator's phase noise or amplifier's noise figure. It is therefore very crucial that on-chip inductors are designed to have Q as large as possible. Several techniques can be employed to enhance the Q as described in the following section.

3.4.6 High Q Inductor Design Considerations

The Q of an inductor along with its inductance and physical size are the most important design parameters. Q of an inductor depends on its principal inductance as well as its resistive, inductive and capacitive parasitics. It is inversely proportional to the inductor's loss. For a given value of inductance, to minimize loss and hence maximize Q of inductors, a number of design considerations must be carefully taken by RFIC designers. In order to achieve this, it is important to identify sources of contributions to the loss and hence Q of inductors, based upon which various methods to increase Q can be devised. The dominant factor affecting Q of an inductor, as inferred from Figure 3.23, is different for different frequency ranges. For example, referring to any model in Figures 3.20 and 3.21, the dominant loss at low frequencies is due to the series resistance R_s of the conducting trace, and hence reducing this resistance would increase Q. Reduction of this resistance can be obtained by various means such as increasing the trace width or stacking multiple via-connected conducting traces. On the other hand, at higher frequencies, the mutual inductive coupling between the inductor's trace and silicon substrate, represented by L_e, M, and R_e, becomes more dominant, which needs to be reduced in order to obtain an increased Q. Therefore, enhanced-Q methods should consider specific operating frequency regions of inductors employed in RFICs. It should be noted that, as the resonant frequency also depends on the main inductance and parasitic capacitances, attention must be paid to the resonant frequency of an inductor as it is optimized for high Q, in addition to the inductor size.

Another important remark concerning the design of inductors in general is that careful consideration needs to be exercised in designing to achieve a particular inductance. For a given inductance, there are several different inductor topologies that can be used and, for a chosen topology such as spiral inductor, there are different combinations of physical dimensions and/or structures that can be considered. As an example, we assume that the designed inductor is modeled as an equivalent circuit shown in Figure 3.20(b). Different designs for an inductor, while can give a desired inductance "L," may produce different values for the parasitic elements (C_p, R_s, R_{ds1}, R_{ds2}, C_{ds1}, C_{ds2}). These parasitics affect circuits or elements that are connected directly to the inductor – for instance, C_{ds1} and C_{ds2} with proper values may inadvertently resonate with inductors (or inductive elements of circuits) connected directly to Port 1 and Port 2 of the designed inductor, respectively. Therefore, the design of an inductor requires consideration of both the inductor itself and the surrounding circuits; that is both inductance and parasitic values need to be considered in the inductor design so that proper values for parasitics can be obtained, rather than focusing only on achieving a particular inductance. In general, all parasitic elements should be kept as negligible as possible. The same consideration should also be used for the design of other elements.

[15]For instance, $Q = \frac{\omega W_m}{P_L} \neq \frac{\frac{1}{2}\omega L_e I_{\text{rms}}^2}{\frac{1}{2}R_e I_{\text{rms}}^2} = \frac{\omega L_e}{R_e}$

3.4.6.1 Q Contributed by Physical Volume and Area of Inductors. The fundamental expression for Q is

$$Q = \frac{\omega W}{P_L} \tag{3.83}$$

where W represents the total time-average energy stored in the inductor and P_L is the power loss, or energy loss per second, in the inductor. Note that W and P_L may be contributed by both magnetic and electric energy. At low frequencies, the electric energy stored in the parasitic capacitors of an inductor is relative small and may be neglected. In general, the Q of an inductor can be separated into two terms according to

$$\frac{1}{Q} = \frac{1}{Q_c} + \frac{1}{Q_d} \tag{3.84}$$

where Q_c and Q_d are the conductor Q, contributed by the conductor loss, and dielectric Q, contributed by the dielectrics and silicon substrate beneath the inductor, respectively. Since on-chip inductors such as spirals may be considered as distributed structures in RF regime, for given dielectrics and silicon substrate, Q_c and hence Q of an inductor is approximately proportional to the ratio between the volume occupied by the inductor and the conducting surface area in the inductor, borrowing from the discussion of distributed resonators, as

$$Q \propto \frac{\text{Inductor volume}}{\text{Conducting surface area}} \tag{3.85}$$

So the Q of inductors can be increased by designing the inductors such that they occupy a large volume while containing less metal area – for instance, removing the innermost turns of a spiral while keeping the same total area or volume enhances the Q. It should be noted that the Q given in Eq. (3.85) is frequency dependent due to the fact that the electrical dimensions of inductors and the energy and loss implicitly associated with volume and surface area are the function of frequency.

3.4.6.2 Q Contributed by Conductor and Dielectric-Substrate Losses. The Q of inductors is given by Eq. (3.23), assuming the inductors are modeled as an inductance L in series with a resistance R_L accounting for the losses due to conductor (R_c) and dielectric and substrate (R_{sub}) losses. Low loss SiO_2 dielectrics are typically used in RFIC fabrication processes, so the dominant dielectric-substrate loss is due to silicon substrate. For a given inductance and frequency, Q is inversely proportional to the series resistance R_L. Reducing R_L, or specifically R_c and/or R_{sub}, hence increases Q.

Conductor Resistance R_c

Conduction loss is mainly contributed by two different kinds of current flowing in conductors. One contribution is due to currents penetrating or flowing into the conductor, which relates to the conductor's skin depth, and another one is due to the magnetic-induced eddy currents flowing in the conductor. Both result in an increase of resistance.

The conduction loss due to skin depth effects can be reduced by proper design of the conductor itself. Using thick metallization helps reduce the series resistance and hence conduction loss. In the CMOS process, the topmost metal layer typically has the thickest metallization. Advanced CMOS processes may have the topmost layer of several microns thick. Connecting multiple metal layers through via-holes also overcomes the inherent thin metallization problem in these processes. Use of conductors with high conductivity (e.g., copper versus aluminum) reduces the loss as well.

For loss due to the flow of eddy currents on conductors, one can reduce the loss at RF frequencies by reducing the magnitude of eddy currents, particularly at high frequencies where eddy currents are strong. It should be mentioned that these eddy currents occur due to the time-varying magnetic fields generated not only by the time-varying or AC currents flow in other adjacent conductors but also by the time-varying currents flow in the conductor itself. Moreover, eddy currents are perpendicular to magnetic fields. As such, the magnitude of eddy currents depends on the geometry of the conductor itself and its relative position and

Figure 3.29. Rectangular (a) and circular (b) spiral inductor with variable trace width.

orientation with respect to adjacent conductors. Eddy currents in spiral inductors can be reduced by implementing techniques to reduce or disturb their flow on the traces of inductors. For instance, the eddy-current magnitude can be reduced by making the widths of the inner turns narrower than those of the outer turns [14], as shown in Figure 3.29. This technique is based on the fact that the inner turns are exposed to a stronger magnetic field, which results in more magnetically induced eddy currents on the inner conducting traces and hence loss, as compared to the outer turns. The eddy-current loss on the inner traces can thus be a substantial portion of the total conduction loss of spiral inductors. On the other hand, the inner turns have small magnetic flux due to their small area, thus contributing less to the total inductance value as compared to the outer turns. Making the trace's width in the inner turns narrower than that in the outer turns thus results in increased Q with possibly only a small sacrifice in the inductance value. A narrow trace has large DC resistance as seen in Eq. (3.33). However, this can be compensated for by using a wider trace in the outer turns. The Q enhancement is more pronounced at high RF frequencies since the eddy current effect is more at these frequencies.

As a demonstration of Q improvement through optimized layout, three types of octagonal spirals are designed using a 0.18-μm CMOS process: one with a constant trace width of 6 μm, one with a constant trace width of 15 μm, and another one having varying trace width from 6 μm at the innermost turn to 15 μm at the outermost turn – all in the topmost metal layer. These inductors all have five turns and trace spacing of 2 μm. The 15-μm-width inductor has an inner diameter of 60 μm, while the diameters of the 6-μm-width and varying-width inductors are around 70 μm. These dimensions are chosen so that all the inductors have the same inductance of around 3.5 nH. Figure 3.30 shows the layouts of these inductors and Figure 3.31 shows their calculated inductance and Q using IE3D. At frequencies lower than 1.5 GHz, the inductor with a wider trace (15 μm) has the largest Q, which is primarily due to the fact that the conduction loss is the main contribution to the total loss at these frequencies. At frequencies higher than 3.5 GHz, however, the inductor with a narrower trace of 6 μm has the highest Q. At these frequencies, the substrate loss becomes the dominant contributor. The varying width inductor, on the other hand, has the highest Q from 1.5 to 3.5 GHz. These results demonstrate the feasibility of improving Q over certain frequency ranges through optimization of the layout. It is noted that the calculated results in Figure 3.31(b) do not show significant improvement in Q for different cases. This is due to the fact the designed inductors' traces are purposely placed not close to the spiral's center, where the magnetic field reaches its maximum, as imposed by the employed CMOS process design kit (PDK).

Substrate Resistance R_{sub}

Any practical substrate has loss, typically referred to as (shunt or leakage) substrate loss, resulting in an increase to the overall loss of any passive structures deposited above the substrate. In CMOS processes, multiple dielectric layers, typically oxide, have very small loss, causing negligible loss. However, the silicon substrate, either intrinsic or heavily doped, has low resistivity and is physically located relatively close to metal layers – depending on a process and/or metal layers used, this distance can be only a few microns. This causes significant (substrate) loss effect on inductors, resulting in substantially reduced Q. Inductors therefore should not be deposited directly onto silicon substrate or onto metal layers adjacent to the substrate.

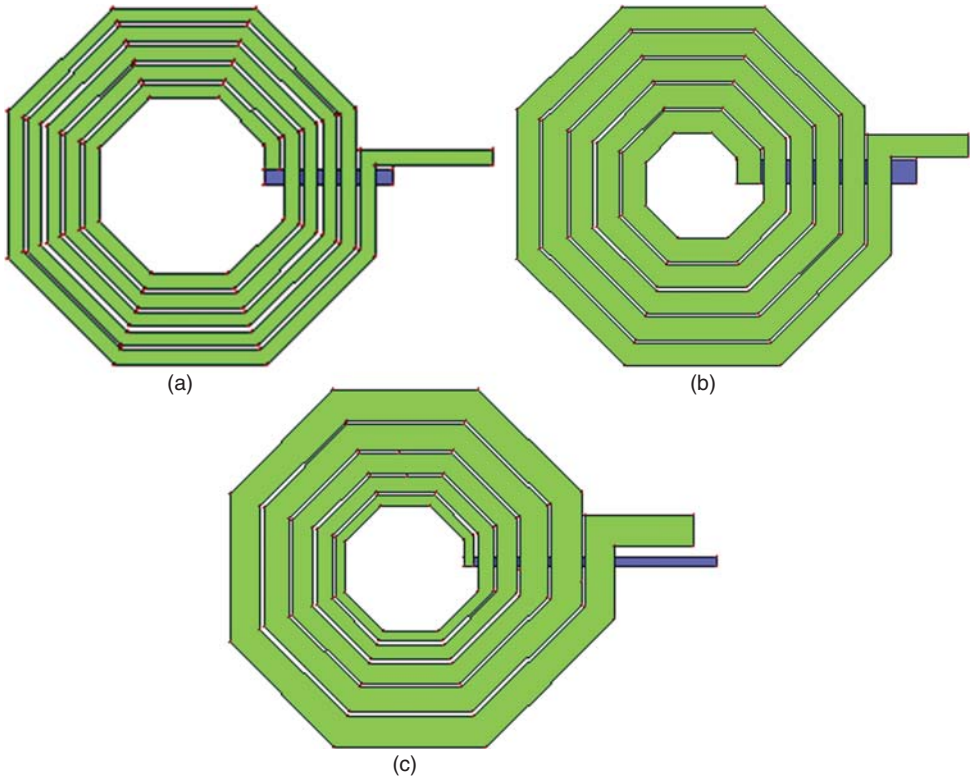


Figure 3.30. Layouts of octagonal spirals on a 0.18-μm CMOS process with (a) 6-μm constant trace width, (b) 15-μm constant trace width, and (c) 6-to-15 μm varying trace width.

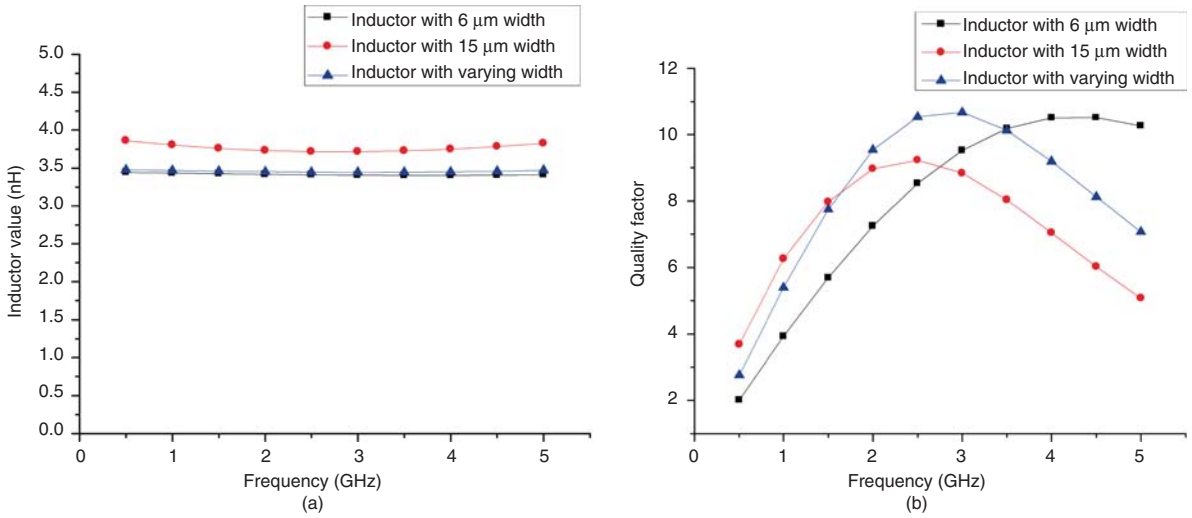


Figure 3.31. Calculated inductance (a) and Q (b) of the three inductors in Figure 3.30.

Substrate loss is contributed by two factors. One is the shunt resistance due to conduction and displacement currents flowing within the substrate, resulting in loss which is usually classified as dielectric loss. This loss is the substrate electrical loss, mentioned earlier. The other is the eddy-current resistance from magnetically induced eddy currents in the substrate, which is due to the conductive nature of silicon. The resultant loss is the substrate magnetic loss, stated previously. These substrate losses are more severe at high RF – typically above 2 GHz depending on silicon substrates. Having identified these sources of loss, the substrate loss can

then be minimized using several ways. One way is to reduce or prevent the electric and magnetic fields of the signal, and hence the signal itself, from propagating into the lossy silicon substrate, which will result in reduction or elimination of loss due to the silicon substrate. For instance, solid conductive layers beneath the inductor may be used as a shield to prevent electric field from penetrating into the silicon substrate. These solid conductive shields can be implemented using a metal layer (typically the bottom-most metal layer) or the polysilicon (or poly) layer. A poly shield is a low conductivity layer and thus has less eddy-current loss than a metal shield. A conductive shield, however, increases the shunt capacitance of the inductor significantly, which is actually the capacitance between the inductor and the shield. These effects lead to reduced resonant frequency for the inductor and thus limiting its operating bandwidth. It may also increase the magnitude of the eddy currents due to additional magnetic fields generated by the induced eddy currents on the shield. The effects of the shield on the inductor's parasitic capacitance and eddy currents are less if the inductor's metal layers are far from the shield. A solid polysilicon shield causes less effect on the shunt capacitance and eddy currents than a metal shield and is preferred. Due to increased parasitic shunt capacitance, solid conductive shields are not very attractive for inductors used in RFICs.

A better shielding technique is using patterned ground shields (PGSs), which can also be implemented by either a metal layer or a polysilicon layer [15]. Figure 3.32 shows layouts of a patterned metal ground shield and the spiral inductor above it. In the PGSs, slots perpendicular to the spiral's segments are formed in the shield. These slots essentially cut across the flow of eddy currents, blocking or disturbing the eddy currents and hence reducing the corresponding eddy-current loss. It is noted that these eddy currents are not the eddy currents stimulated by the magnetic fields penetrating into the silicon substrate. Ideally, the slots should be sufficiently small to prevent the electric field from penetrating into the silicon substrate. In practical circuits, especially at very high RF, the PGSs allow part of the electric field penetrating into the silicon substrate, causing some substrate electrical loss.

In general, the silicon substrate is grounded at the bottom, and so grounding the shield helps reduce the substrate electrical resistance [R_{ds1} and R_{ds2} shown in Figure 3.22(d)] and further minimizing the substrate electrical loss. This grounding technique should also be applied to the solid conductive shield mentioned earlier. The PGSs, however, have significantly less (parasitic) shunt capacitance as compared to the solid ground shields, and are thus preferred. As for solid ground shields, a patterned shield implemented on a poly layer gives higher Q than that using a metal layer due to its low conductivity which results in less eddy currents. It should be noted that silicon substrate is partly conductive and is located relatively near the inductor's layer, thus also producing a capacitor in parallel with the inductor, resulting in increased resonant effect and effectively lowers the operating frequency further. Note that this capacitance is not addressed in the equivalent-circuit models shown in Figure 3.22. Another effect is the eddy currents in the silicon substrate

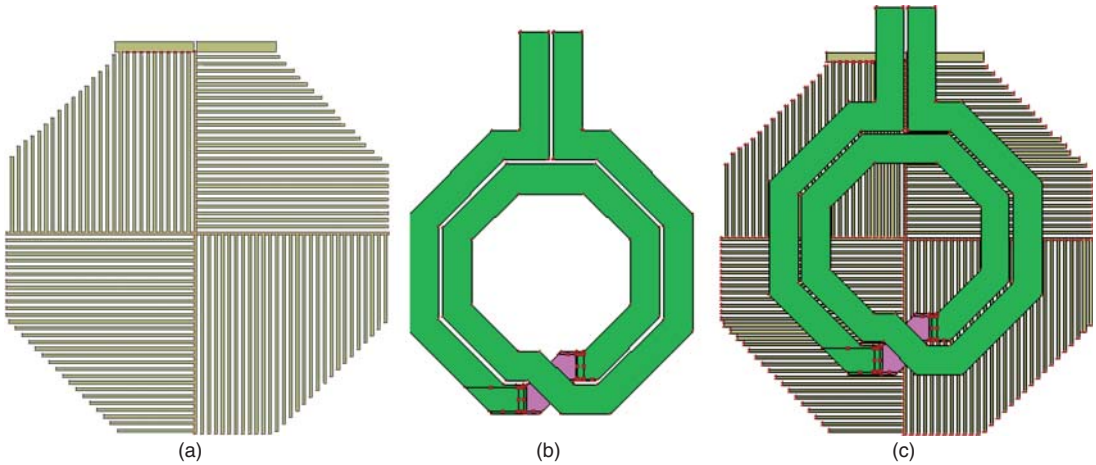


Figure 3.32. (a) Layout of a patterned metal ground shield (slots are the white areas). (b) Layout of a spiral inductor. (c) Layout of the spiral inductor above the patterned metal ground shield.

flows in the opposite direction to the currents on the inductor's trace, hence partly canceling the inductance produced by the spiral inductor. A well-designed patterned metal shield can achieve negligible eddy currents flowing on the shield and, thus, also may give comparable Q to a patterned poly shield. It should be noted that, due to the shield, the substrate capacitance (C_{sub}) is also reduced. This reduction effectively increases the parasitic capacitance formed by C_{ox} in series with C_{sub}, resulting in an undesired lower resonance frequency. PGSs, especially those using a polysilicon layer, are used in RFICs as an effective means for achieving high Q for on-chip inductors.

It should be noted that ground conductive shields, whether solid or patterned, can only prevent or partly prevent electric fields from penetrating into the silicon substrate. A solid or patterned conductive plane, on the other hand, does not stop magnetic fields from existing within the conductive silicon substrate, which consequently excite eddy currents causing magnetic loss in the substrate. This phenomenon basically results from the primary time-varying magnetic fields, which subsequently induce other magnetic fields and eddy currents in conductive materials, as dictated in Faraday's law. Therefore, while the substrate electrical loss due to penetrating electric fields may be significantly reduced through the ground conductive shields, the substrate magnetic loss may not. In order to reduce the magnetic loss, some kinds of mechanism need to be employed to reduce the magnetic fields induced by the current flowing in the inductor, which consequently results in reduced magnetic fields in the substrate. One such way is driving inductors differentially to be discussed later. Another interesting point is, as the frequency is increased, the distance between a spiral inductors and its conductive shield is increased with respect to the operating wavelength, thus lessening the effect of the shield upon the Q of the inductor. At sufficiently high frequencies such as those in the millimeter-wave region, the inductor Q with and without a conductive shield may be comparable, making the shield become unnecessary.

Figure 3.33 shows a symmetric spiral inductor on a 0.18-μm CMOS process. Figure 3.34, showing calculated results for Q of the same spiral inductor (Figure 3.34) under three different designs, illustrates the substrate losses resulting from magnetic loss and from both magnetic and electrical losses. The first curve (spiral on lossless substrate) assumes a perfect silicon substrate ($\sigma = 0$), which does not have either substrate electrical or magnetic loss. In the second curve (spiral on PGS), a patterned metal ground shield on the lowest metal layer M1 is used, which significantly alleviates the substrate electrical loss. Accordingly, the substrate loss is approximately due to magnetic loss. The third curve (spiral on Si) uses the actual conductive silicon substrate, and thus includes both the substrate electrical and magnetic losses. These results demonstrate the significant effects of substrate losses on the Q of on-chip inductors and hence on the performance of RFICs. As can be

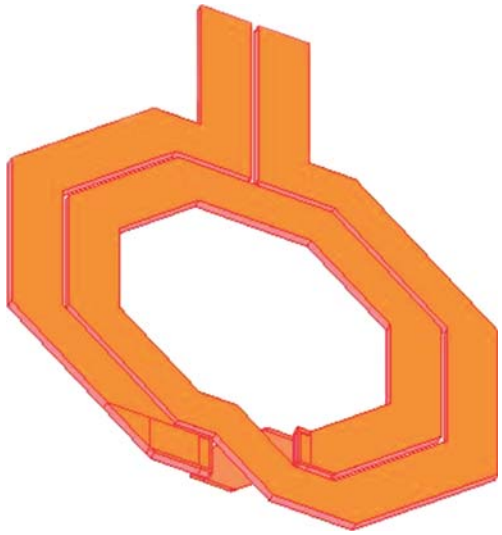


Figure 3.33. Layout of a two-turn symmetric spiral inductor on a 0.18-μm CMOS process. The trace is on the topmost metal layer M6 and the cross-under is on the metal layer M5. The trace with and spacing are 15 and 2 μm, respectively. The inner and outer diameters are 80 and 144 μm, respectively. The inductance is approximately 0.68 nH.

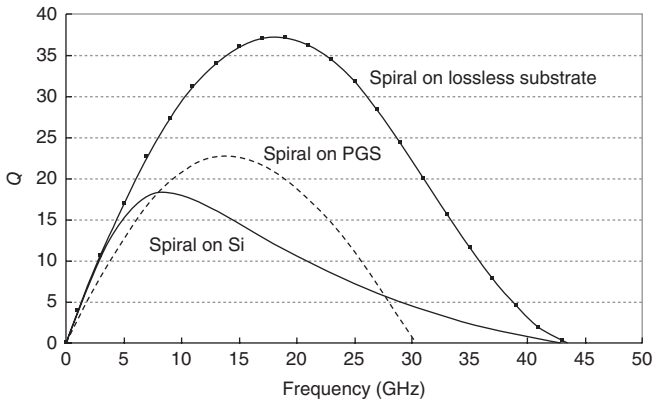


Figure 3.34. Calculated Q of the spiral inductor in Figure 3.33 for three different designs using the EM simulator IE3D.

inferred, a patterned conductive shield cannot prevent substrate loss due to the magnetic fields induced within silicon substrates. The first and third curves show that the resonant frequencies, corresponding to zero-Q, for perfect and actual silicon substrates are equal; this is not truly accurate as the parasitic capacitance for the actual silicon substrate should be higher than that for the perfect silicon substrate, which should result in lower resonant frequency.

3.4.6.3 Q Enhanced by Physical and/or Electrically Induced Changes. Q of spiral inductors can be enhanced by physically modifying the spiral and/or its surrounding physical structures, such as using varying trace width and/or patterned slots on a beneath conductive shield, as discussed earlier. The basic principle behind this is to achieve a spiral structure that produces an overall electrical effect of a high Q spiral. This electrical effect can be described by an equivalent circuit that gives high Q. For instance, using a patterned conductive shield underneath a spiral effectively creates a structure that has an electrical effect of reduced substrate resistance and hence increasing Q. In the same principle, altering the performance of a spiral by an electrical means also produces an electrical effect that may result in high Q. Differentially exciting a spiral structure is one such way of producing an electrical effect that may enhance the Q.

The Q of spiral inductors can be significantly increased by driving them differentially, instead of singly ended as normally done, in which 180° out-of-phase voltages or currents are applied to the two ports of the inductors. Differentially driven spiral inductors were developed and demonstrated with a 50% increase in Q and a wider operating frequency range as compared to singly ended driven asymmetric inductor [16]. Figure 3.35 shows layout of a differentially driven spiral inductor. The configuration is symmetric as opposed to asymmetric spirals typically driven single-ended. The inductor's traces can be on the same metal layer or different layers, but they are not continuous. For inductors on a single metal layer, the connections between traces are formed using cross-over and cross-under metals, as seen in Figure 3.35. Apart from increased Q, differentially driven symmetric spirals operate with a wider frequency range, and are physically small and easier to connect with other elements in RFICs (due to their symmetry).

Differentially driven spiral inductors are attractive for RFICs due to several reasons. Various differential circuits, such as differential amplifiers, oscillator, mixers, are often used in RFICs due to their inherent advantages such as common-mode rejection including rejection of noise from power supply, less harmonic products, and large output voltage swing. A differential circuit also facilitates direct connection with other balanced circuits without the need of baluns or transformers. A single symmetric spiral inductor may be considered electrically equivalent to two separate conventional spiral inductors, whose ports are both electrically and physically asymmetric. Figure 3.36(a) shows two asymmetric spiral inductors, with the inner ends tied together, to form a symmetrical structure. Figure 3.36(b) shows the equivalent symmetric inductor. Ports 1 and 2 of the inductors in Figure 3.36(a) and (b) are driven differentially, as shown. In these two symmetric inductors, currents i_1 and i_2 flow in opposite directions due to the differential excitation at ports 1 and 2. However, i_1 and i_2 each maintains the same direction along the trace of the corresponding inductor due to

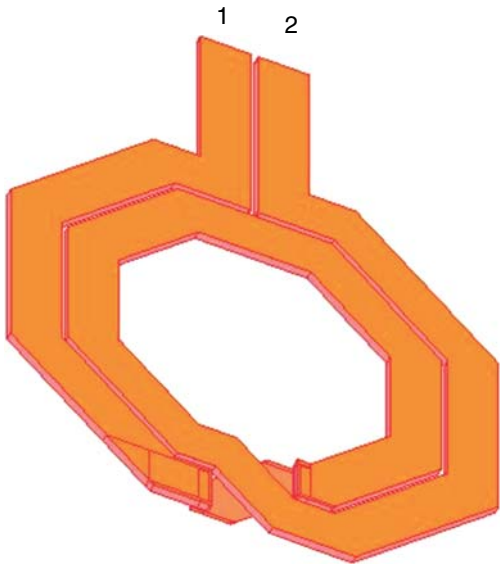


Figure 3.35. Symmetric spiral inductor. Ports 1 and 2 may be driven differentially (e.g., with ±1 V).

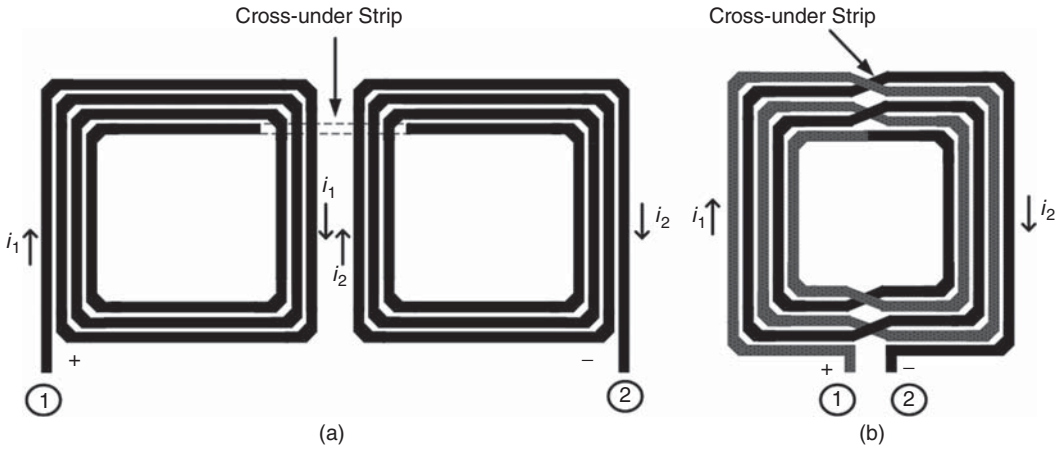


Figure 3.36. Two asymmetric spiral inductors with the inner ends connected via a cross-under metal strip (a) and their equivalent symmetric spiral inductor (b). In these inductors, ports 1 and 2 are driven differentially.

the fact that each individual inductor is singly ended driven independent of the other. Here we assume these two inductors are sufficiently apart to produce negligible mutual electric and magnetic couplings between them. When Ports 1 and 2 are driven differentially with opposite voltages (e.g., ±1 V), the initial currents i_1 and i_2 have opposite directions but those flowing along any two adjacent traces have the same direction. This results in supposition of individual magnetic field produced by each trace, hence causing an increase in the overall inductance. Compared to the differentially driven asymmetric inductor pair, which requires a sufficient space between the inductors, the symmetric inductor has a smaller size. Furthermore, its connection with other passive and/or active devices is more convenient due to the proximity of the inductor's two terminals.

The improvement in the Q of a differentially driven inductor over that of a singly ended driven inductor can be estimated as follows. Using the equivalent circuit for spiral inductors shown in Figure 3.22(c), simplified equivalent circuits for singly ended and differentially driven spiral inductors can be derived as shown in Figure 3.37(a) and (b), respectively [16]. As can be seen in these figures, the differential excitation produces

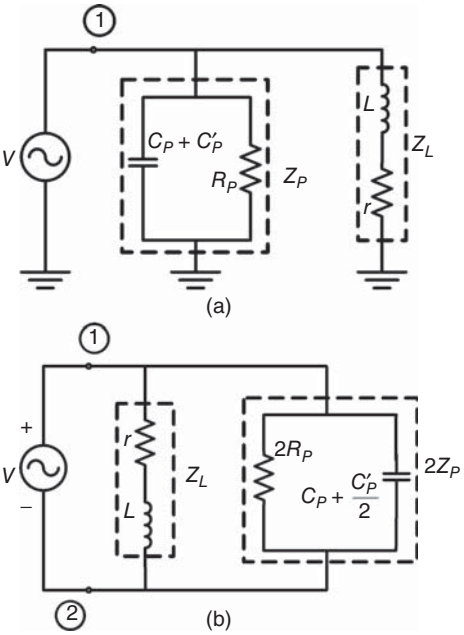


Figure 3.37. Equivalent circuits for singly ended (a) and differentially (b) driven spiral inductors derived from the equivalent-circuit model of spiral inductors shown in Figure 3.22(c). Port 2 in (a) is grounded.

an electrical effect of increasing the parallel resistance to $2R_P$ for the differentially driven spiral as compared to only R_P for the singly ended driven spiral. The ratio of the Q's between the differential and single-ended drives can be derived as

$$\frac{Q_d}{Q_{\text{se}}} = \frac{2(R_P + R_L)}{2R_P + R_L} \tag{3.86}$$

where Q_d and Q_{se} are the differentially and singly ended driven quality factors, respectively, and

$$R_L = r(1 + Q_L^2) \tag{3.87}$$

with

$$Q_L = \frac{\omega L}{r} \tag{3.88}$$

At low frequencies, $R_P \gg R_L$ and the two Q's are approximately equal. Within a low frequency range, Q_L dominates and both the Q's increase with increasing frequency. In the high frequency range, when the frequency is increased, R_L increases while R_P decreases, resulting in a larger Q for the differential drive. At certain high frequencies, the effects of substrate loss on Q becomes minimum, thus both the single-ended and differential Q's reach their respective maximum. However, as the frequency keeps increasing, R_P dominates the loss and the Q's reduce accordingly. Equation (3.88) shows that theoretically the differential Q can be twice as large as the single-ended Q at certain frequency where $R_P \ll R_L$.

Figure 3.38 compares the Q of the same inductor under single-ended and differential excitations. It is noted that, in practical differentially driven inductors, the induced magnetic fields are never equal to zero but at a much reduced level as compared to those of single-ended drive. This reduced magnetic field is another key factor, besides reduction of the shunt capacitive parasitics, that helps improve the Q of differential inductors.

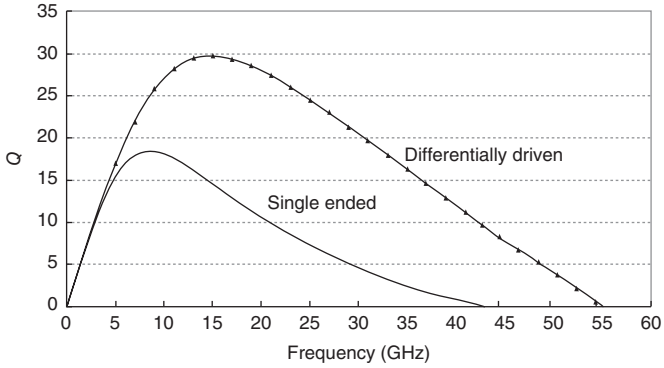


Figure 3.38. Calculated Q of the spiral inductor in Figure 3.33 (without a patterned metal ground shield) for single-ended and differential drives using the EM simulator IE3D. The single-ended curve is the same as the Spiral-on-Si curve in Figure 3.34.

3.5 LUMPED-ELEMENT CAPACITORS

Ideal capacitors consist of only capacitors, represented by a capacitance C. Practical on-chip capacitors, however, have capacitance, inductance, and resistance, which result in resonant frequency and finite Q. Three important electrical parameters characterizing a capacitor are its capacitance, resonant frequency, and Q. In RFIC design, not only the designers are interested in realizing a particular capacitance, but also in achieving a Q as high as possible for the capacitor at the interested frequencies and a resonant frequency well above the desired operating frequency range. As compared to the commonly used on-chip spiral inductors, the most widely used on-chip MIM capacitors have a very high Q and a resonant frequency much higher than the operating frequencies.

3.5.1 Capacitor Configurations

Various configurations can be employed to realize on-chip capacitors using combinations of conductors and dielectrics. They are basically based on the principle of EM couplings between conductors resulting from electric fields produced by these conductors within a particular dielectric environment. Figure 3.39 illustrates some commonly used capacitor configurations.

The first three capacitor configurations, as shown in Figure 3.39(a)–(c), involve only a single metal layer and primarily produce capacitances from the coupling of electric fields between the edges of the gaps via the dielectrics and substrate underneath and the dielectrics or air above them. The produced capacitances are thus relatively small and hence suitable for circuits requiring small capacitances, such as high impedance matching, or for those that can use small capacitances, such as DC block at very high RF (e.g., above 100 GHz). The end- and edge-coupled capacitors, however, have virtually no or very little associated inductances, and hence can be used over an ultra-wide bandwidth (UWB) and at a very high RF. Accurate capacitances for the end- and edge-coupled capacitances can be determined from the even- and odd-mode capacitances [17] or by an EM simulator. The interdigital capacitor, as shown in Figure 3.39(c), particularly can be modeled as multiconductor parallel-coupled transmission lines. It occupies a relatively large area, as compared to the other capacitors shown in Figure 3.39, and is thus not suitable for RFICs even though it requires only a single metal layer. In addition, it also contains inductance resulting from the multiple-conductor transmission lines, thus possibly producing multiple resonances, further making it not very suitable for high RF, particularly for broadband circuits. The presence of inductance is more pronounced in the interdigital capacitor as compared to the broadside-coupled capacitor [Figure 3.39(f)], which is also based on coupled transmission lines, due to the relative small capacitance in interdigital capacitors. In these three capacitors, the capacitive coupling occurs primarily at the edges where the current distribution is maximum. This current concentration limits the Q of these capacitors.

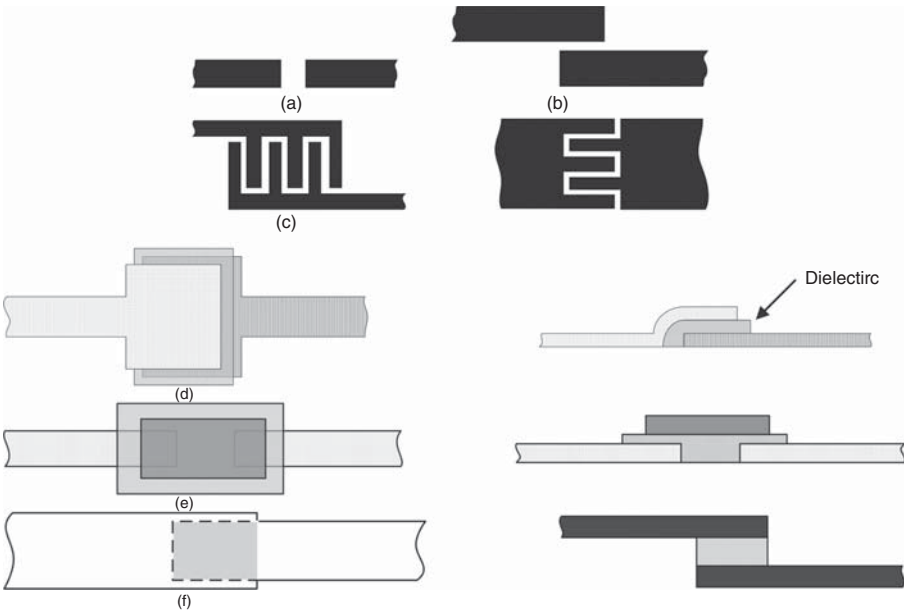


Figure 3.39. Common on-chip capacitors: Gap capacitor (a), edge-coupled capacitor (b), interdigital capacitor (c), MIM capacitor (d), end-coupled MIM capacitor (e), and broadside-coupled capacitor (f). The conductors for (e) and (f) can be placed off-set to allow partial overlaps between the conductors.

The MIM capacitor, broadside-coupled capacitor, end-coupled MIM capacitors, as shown in Figure 3.39(d)–(f), use multiple metal and dielectric layers, typically two metal and one dielectric layers, and require relatively small areas. These capacitors can be designed to have small, moderate or large capacitances. They are suitable for both impedance-matching, bypass and blocking capacitors. The capacitances are produced by direct coupling between the conducting patches or metal plates via the in-between dielectric(s). In these capacitors, the capacitive coupling occurs across the plates and the current distributes more uniformly between these plates. Therefore, the Q's of these capacitors are expected to be higher than those for the capacitors in Figure 3.39(a)–(c). These capacitors are basically like "parallel-plate capacitors" with the MIM capacitor, shown in Figure 3.39(d), particularly resembling most closely the parallel-plate capacitor. The conducting strips or patches used in these capacitors are implemented on metals sandwiched within an RFIC structure. The topmost layer in RFIC fabrication processes has the thickest metallization and farthest from the silicon substrate, resulting in less loss and hence higher Q for top-metal circuit structures. It is normally employed for the gap and interdigital capacitors and for the top plate of the MIM capacitors. The MIM capacitors [Figure 3.39(d)] are the most commonly used capacitor in RFICs. They are available with different capacitance values in the design kits or libraries of CMOS processes, where they can be used directly in RFICs. The end-coupled MIM capacitors [Figure 3.39(e)] are not typically available in RFIC fabrication processes and need to be custom-designed. The produced capacitance can be considered two MIM capacitance in parallel and the design of each capacitance is similar to that for the MIM capacitor Both the MIM and end-coupled MIM capacitors have inductance associated with them, which results in resonant frequencies and hence limited operating bandwidths. These resonant frequencies, however, are usually very high, depending on the capacitors' dimension, and are typically well above the desired operating frequencies for the capacitors.

The broadside-coupled capacitors [Figure 3.39(f)] provide an alternative for the MIM capacitors available in CMOS processes for small to large capacitances through custom design. The capacitance of these capacitors is produced based on the same principle of the MIM capacitor. This structure provides a convenient way of realizing a desired capacitance, instead of being restricted by available MIM capacitors in a particular CMOS process' library. RFIC designers can design various broadside-coupled capacitors to achieve desired capacitance values and configurations suitable for particular circuit layouts by exploiting multiple

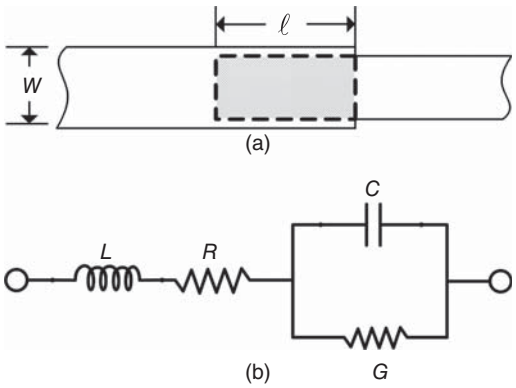


Figure 3.40. Broadside-coupled capacitor (a) and its simplified equivalent circuit model (b). $R = R'\ell$, $L = L'\ell$, $G = G'\ell$, $C = C'\ell$, where R', L', G', and C' represent the resistance, inductance, conductance, and capacitance per unit length of the equivalent transmission line.

metal and dielectric layers. Broadside-coupled capacitors are particularly attractive for simple DC blocks in ultra wideband circuits. The design and analysis of broadside-coupled capacitors can be performed using the broadside-coupled transmission lines [18]. In CMOS structures, the dielectric layers are usually very thin, around few microns. As such, the energy between the two plates is concentrated mostly within the plates, unless a ground plane, such as that used for a transmission line connecting directly to the broadside-coupled structure, is relatively large and located very close to the plates. In practice, this ground plane's effect on the broadside structure is negligibly small.[16] Therefore, the broadside-coupled structure can be considered simply as a single two-conductor transmission line, consisting of two metal strips with thin dielectrics between them. For thin dielectrics, the electric fields tend to be concentrated between the conductors, making the other nearby conductors become less dominant. Consider the conventional transmission line equivalent circuit and the fact that the top and bottom plates are connected to external elements (e.g., transmission line) at one end with the other end open; we can approximately represent the broadside-coupled capacitor structure, as shown in Figure 3.40(a), by an equivalent-circuit model shown in Figure 3.40(b). The per-unit-length parameters and hence the model parameters can be determined using procedures for two-conductor transmission lines or from an EM simulator. Assume the dielectrics above the top plate, between the plates, and below the bottom plate are the same, and the silicon substrate is sufficiently far or isolated (e.g., by a conductive shield) from the bottom plate, the plates can be considered embedded within a homogeneous medium and can therefore support transmission electron microscope (TEM) propagation mode (see Chapter 4 of transmission lines). Under this condition, the per-unit-length shunt conductance G' and capacitance C' are related by

$$\frac{G'}{C'} = \omega \tan \delta \tag{3.89}$$

where $\tan \delta$ is the dielectric's loss tangent. The per-unit-length inductance L' and C' are related by

$$L'C' = \mu\varepsilon = \frac{1}{c^2}\varepsilon_r \tag{3.90}$$

where $c = 3 \times 10^8$ m/s is the speed of light in free space and ε_r is the dielectric's relative dielectric constant (for SiO_2, it is around 3.8). The per-unit-length resistance R' is determined by the metal's surface resistance and skin depth. In typical CMOS processes, the metal thickness is relatively thin (less than a few micrometer), making skin effect significant, particularly at high RF frequencies.

[16]For instance, if a microstrip line is used for connection, the ground plane would be located several layers away from the broadside-coupled structure, and thus sufficiently far from structure, in order to have a sufficiently wide microstrip line. For CPW, the ground strips are on the same metal layer with the top plate and do not affect the broadside structure even nearby due to the off-set locations between them.

At frequencies where both the width W and length ℓ of the capacitor are relatively small compared with the wavelength of the corresponding broadside transmission line, $\lambda = c/f\sqrt{\varepsilon_r}$, the capacitance C can be approximately derived as

$$C = C_p + 2C_f W \tag{3.91}$$

where C_f is the fringing capacitance per unit length at the end of the strip and C_p is the parallel-plate capacitance of the plates given by

$$C_p = \varepsilon_o \varepsilon_r \frac{W\ell}{d} \tag{3.92}$$

with d being the spacing between the plates, assuming W and ℓ are relatively large compared to d (e.g., W/d and $\ell/d > 10$).

Due to the availability of multiple metal and dielectric layers, and via-holes in CMOS processes, other configurations for capacitors can be derived by utilizing properly the metals, via-holes, and dielectrics to create physical structures that produce capacitances. Besides lumped-element capacitors created by physical layouts, other capacitors in CMOS processes can also be realized. One of them is the metal oxide semiconductor (MOS) capacitor utilizing the gate capacitance of CMOS transistors. Another one is the varactor diode producing capacitance from its junction. The varactor's capacitance can be changed with applied bias voltage, and is thus normally used for electronically tuned circuits such as voltage-controlled oscillator (VCO). The Q of a varactor is in general inversely proportional to the varactor's capacitance tuning range.

3.5.2 Equivalent-Circuit Models of Capacitors

MIM capacitors are the most widely used on-chip capacitors for RFICs. Figure 3.41 shows a general overall configuration of MIM capacitors, consisting of two plates separated by a single or multiple dielectric layers (typically SiO_2). The top plate is usually on the topmost metal layer of a CMOS structure. The bottom plate can be on any metal layer and can sit directly on the silicon substrate or on a SiO_2 layer deposited on top of the silicon substrate. The dominant capacitance of an MIM capacitor is the parallel-plate capacitance between the two metal layers forming the plates. This capacitance can thus be large due to small spacing between the two plates. Other capacitances are fringing capacitances along the edges of the plates, capacitances between the bottom plate and other metal layers and the silicon substrate below the bottom plate, capacitances between the top plate and other metal layers above the top plate (if the top plate is not the topmost metal layer).

The simplest capacitance model for MIM capacitors is an ideal capacitor between the two plates. This capacitance is given from the well-known parallel-plate capacitance formula:

$$C = \varepsilon_o \varepsilon_r \frac{W\ell}{d} \tag{3.93}$$

where W and ℓ are the width and length of the plates, respectively; d is the thickness of the dielectric between the plates (or separation of the plates); $\varepsilon_o = 8.854 \times 10^{-12}$ F/m (farad/meter) is the permittivity or dielectric

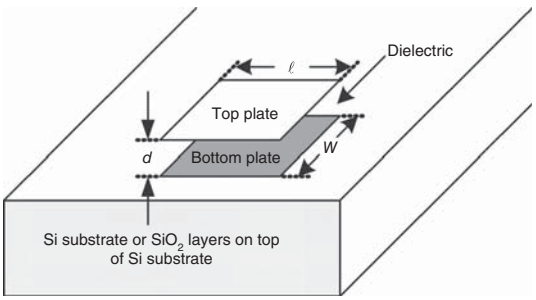


Figure 3.41. General MIM capacitor configuration.

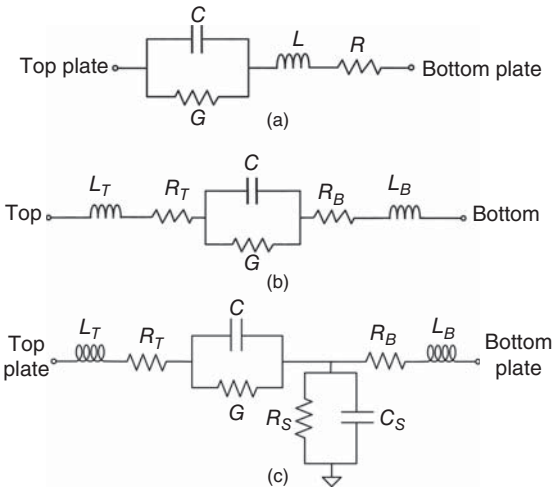


Figure 3.42. (a–c) Simple equivalent-circuit models for MIM capacitors.

constant of the vacuum (in practice, free space is normally considered a vacuum); and ε_r represents the relative dielectric constant or relative permittivity of the dielectric between the plates. Equation (3.93) gives a reasonably accurate capacitance value when the plate dimensions (W and ℓ) are relatively large compared to the distance between the plates, typically greater than 10 for W/d and ℓ/d. As the plate dimensions are reduced, the parallel-plate capacitance becomes smaller and the fringing capacitances along the edges could be a significant portion of the parallel-place capacitance and hence cannot be neglected. The fringing capacitances effectively increase the plate dimensions; this increase can be approximately determined. To account for these fringing capacitances, the dimensions of the capacitor plates can be reduced accordingly.

More accurate, yet still simple, models for MIM capacitors are given in Figure 3.42.

Model in Figure 3.42(a)

The (main) capacitance C produced by the parallel plates is approximately given by

$$C = \begin{cases} \varepsilon_o \varepsilon_r \dfrac{W\ell}{d}, & \dfrac{W}{d} \text{ and } \dfrac{\ell}{d} \gg 1 \quad \text{(a)} \\ \varepsilon_o \varepsilon_r \left[\dfrac{W\ell}{d} + 2(W + \ell + 2d)\right], & \text{otherwise} \quad \text{(b)} \end{cases} \tag{3.94}$$

Equation 3.94(b) approximately includes effects of fringing capacitances along the edges of the plates by assuming each side of the plates is increased by a value equal to the thickness of the dielectric between the plates. This assumption of increasing by "d" is indeed not very accurate. The resistance R accounting for loss due to finite conductivity of the top and bottom plates is obtained as

$$R = \frac{2}{3}\frac{R_{\text{sh}}\ell}{W} \tag{3.95}$$

where $R_{\text{sh}} = 1/\sigma\delta_s$ is the surface resistivity or sheet resistance (in ohms or ohms per square) of the bottom plate with σ and δ_s being the conductivity and skin depth of the plate, respectively. The conductance G represents loss due to the dielectric layer between the plates and is given by

$$G = \omega C \tan\delta \tag{3.96}$$

where $\tan\delta$ is the dielectric's loss tangent. The parasitic inductance L represents the inductance of the conducting plates and can be obtained as

$$L = 2\times 10^{-4}\ell\left[\ln\left(\frac{\ell}{W+t}\right)+1.93+\frac{W+t}{3\ell}\right]K \tag{3.97}$$

$$K = 0.57 - 0.145\ln\left(\frac{W}{d}\right), \quad \frac{W}{d} > 0.05 \tag{3.98}$$

where t is the metallization thickness of the plates. In Eqs. (3.97) and (3.98), the inductance L is in nano-henry (nH), $\omega = 2\pi f$ with f in Hz, and the dimensions are in micrometer (μm).

Model in Figure 3.42(b)

The model in Figure 3.42(b) is essentially the same as that in Figure 3.42(a) with the parasitic parameters L and R decomposed into those for the top and bottom plates. In this model, the main capacitance C and its associated conductance G are the same as those in Figure 3.42(a). R_T and L_T represent the loss (due to finite conductivity) and the inductance of the top plate, respectively. R_B and L_B represent the loss (due to finite conductivity) and the inductance of the bottom plate, respectively.

Model in Figure 3.42(c)

The model in Figure 3.42(c) is the most comprehensive among those in Figure 3.42. It is the next-level model from that of Figure 3.42(b) with added elements accounting for the effects of the SiO_2 layers and silicon substrate beneath the bottom plate. This model, as can be expected, leads to better convergence than those in Figure 3.42(a) and (b) when its model parameters are determined through fitting using optimization, provided that good initial guesses are used. This model employs the same elements as those in Figure 3.42(b), except R_s and C_S which are used to account for the loss and capacitance of both the SiO_2 layers and Si substrate

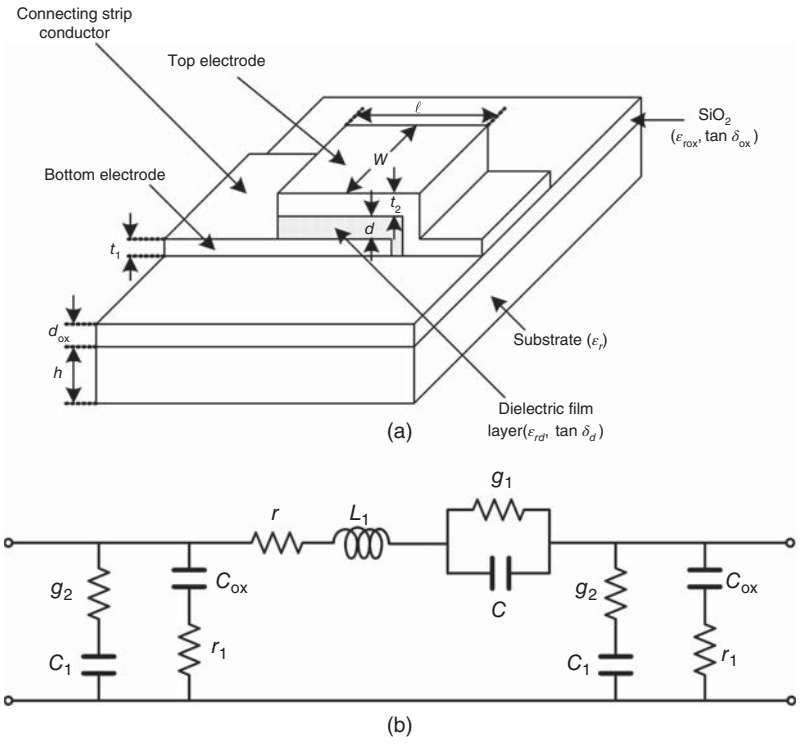


Figure 3.43. MIM capacitor on CMOS structures (a) and its equivalent-circuit model (b).

underneath the bottom plate, respectively. It is noted that contributions of R_s and C_S to the top plate are negligibly small due to its farther distance from the SiO_2 layers and Si substrate and, particularly, due to the fact that the top plate is effectively shielded by the bottom plate from those underneath it.

Simulation for a typical CMOS 30-μm × 30-μm on-chip MIM capacitor is presented in Section 3.1, showing a self-resonant frequency of 210 GHz and Q higher than 1000 up to 30 GHz and around 19 at 60 GHz. Although these high Q values are unlikely to be achieved in practice, they demonstrate possible use of on-chip MIM capacitors as a lumped element up to millimeter-wave frequencies, depending on their dimensions.

Figure 3.43 shows a more detailed two-port equivalent-circuit model for MIM capacitors on CMOS structures along with the capacitor configuration [19]. The main capacitance C can be determined using Eq. 3.94 for parallel-plate capacitances. Other model parameters are considered parasitics and are given below.

$$r = \frac{2}{3}\frac{\rho\ell}{W}\left[\frac{1}{\delta_1\left(1-e^{-t_1/\delta}\right)} + \frac{1}{\delta_2(1-e^{-t_2/\delta})}\right] \tag{3.99}$$

$$\delta_1 = \begin{cases} t_1, & \delta \geq t_1 \\ \delta, & t_1 \geq \delta \end{cases} \tag{3.100}$$

$$\delta_2 = \begin{cases} t_2, & \delta \geq t_2 \\ \delta, & t_2 \geq \delta \end{cases} \tag{3.101}$$

$$\delta = \sqrt{\frac{2\rho}{2\times 10^{-4}\pi^2 f}} \tag{3.102}$$

$$L_1 = \frac{2}{3}(L' + L'') \tag{3.103}$$

$$L' = 2\times 10^{-4}\ell\left(\ln\frac{2\ell}{W+t_1} + 0.5 + \frac{W+t_1}{3\ell}\right) \tag{3.104}$$

$$L'' = 2\times 10^{-4}\ell\left(\ln\frac{2\ell}{W+t_2} + 0.5 + \frac{W+t_2}{3\ell}\right) \tag{3.105}$$

$$g_1 = \frac{1}{\omega C\tan\delta_d} \tag{3.106}$$

$$g_2 = \frac{1}{\omega C_{\text{ox}}\tan\delta_{\text{ox}}} \tag{3.107}$$

$$C_{ox} = \frac{0.5\times 10^{-3} W\ell\varepsilon_{\text{rox}}}{36\pi d_{\text{ox}}} \tag{3.108}$$

$$C_1 = \frac{1}{2}\frac{\varepsilon_r}{h}\ell W \tag{3.109}$$

$$r_1 = \frac{2}{\ell W G_{\text{sub}}} \tag{3.110}$$

where all dimensions (W, ℓ, t_1, t_2, h, and d_{ox}) are in micrometers (μm), capacitance is in pico-farad (pF), inductance is in nH; G_{sub} is the substrate conductance per unit area; and $\rho = 1/\sigma$ is the conductor's resistivity.

MIM capacitors employing multiple dielectric and metal layers can be configured to obtain increased capacitances. Figure 3.44 shows a capacitor using three copper and two dielectric layers (metal–insulator–metal–insulator–metal or MIMIM) fabricated on a Si substrate and its simplified equivalent-circuit model [20]. The configuration is similar to that of an MIM capacitor with an additional metal-dielectric structure. It can be considered two MIM capacitors connected in parallel. Table 3.4 shows the parameters of the model given in Figure 3.44(b) for this MIMIM capacitor and an MIM capacitor extracted from *S*-parameters measured from 0.5 to 10 GHz [20]. Both of these capacitors have the same size of 25 μm × 50 μm and dielectric with the MIM capacitor's dielectric thickness being the same as that of one of the two constituent MIM capacitors of the MIMIM capacitor. It is noted that while the main capacitance of the MIMIM capacitor is about twice of that for the MIM counterpart as expected from the structures of these capacitors, the other model elements have comparable values considering uncertainty in the extraction process. The MIMIM capacitors can achieve increased capacitance without using very thin dielectric layers as compared to the conventional MIM capacitors, hence overcoming the limit of dielectric thickness encountered in the latter. Figure 3.45 shows a possible structure for an MIM capacitor realized using six layers of SiO_2 and seven layers of metal on a CMOS process. It is noted that MIM capacitors employing multiple metal and dielectric layers are in fact a version of the interdigital capacitor shown in Figure 3.39(c) configured vertically using broadside-coupled structures instead of being configured horizontally. The resultant capacitance is thus much larger with smaller areas as compared to that of the interdigital capacitors.

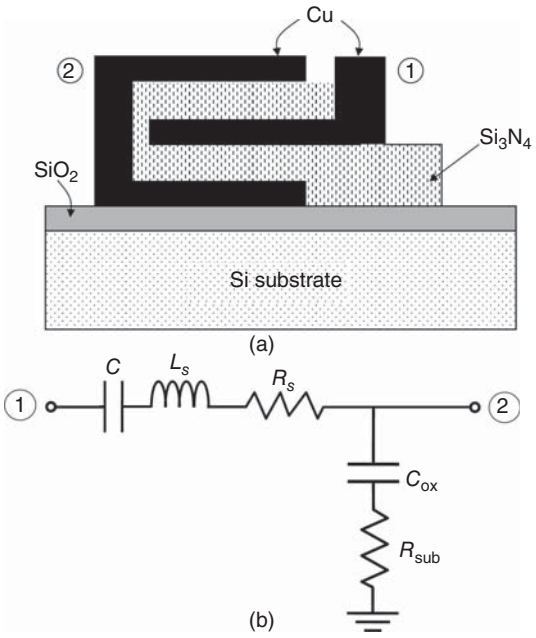


Figure 3.44. Cross section of an MIMIM capacitor using three metal and two dielectric layers on Si substrate (a) and its simplified equivalent-circuit model (b). C is the main capacitance. L_s, R_s, C_{ox}, and R_{sub} are considered parasitics with L_s representing the series inductance due to the copper strips, R_s representing the loss due to the copper and dielectric, C_{ox} representing the capacitive effect due to the SiO_2 layer, and R_{sub} representing the substrate loss.

TABLE 3.4. Measured Model Parameters of 25 μm × 50 μm MIMIM and MIM Capacitors

Capacitor	C (pF)	L_s (nH)	R_s (Ω)	C_{ox} (fF)	R_{sub} (kΩ)
MIM	1.02	0.128	0.61	7.01	4.50
MIMIM	1.91	0.124	0.36	8.49	4.03

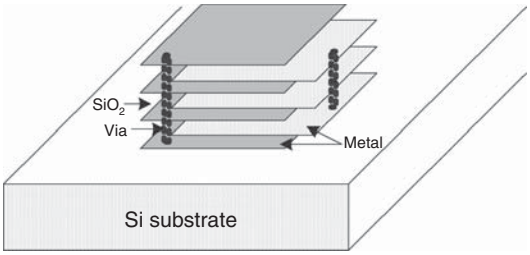


Figure 3.45. MIM capacitor consisting of six SiO_2 layers and seven metal layers on a CMOS process.

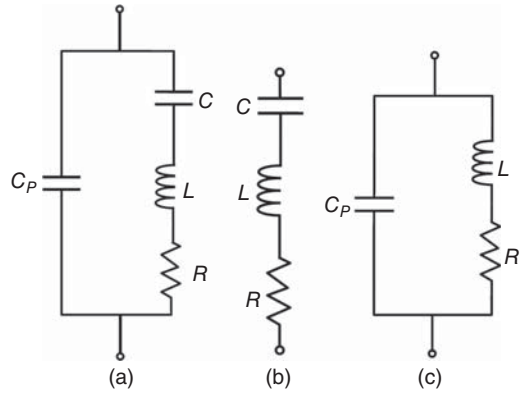


Figure 3.46. Simplified equivalent circuits of capacitors.

3.5.3 Resonance

In contrast with inductors, which typically have parallel resonance,[17] capacitors may have both parallel and series resonances [5]. Without loss of generality, consider a simplified equivalent-circuit model for capacitors shown in Figure 3.46(a), in which C represents the capacitor's self (main) capacitance, R represents the total resistance, and hence loss, across the capacitor's terminals (e.g., resistances due to contacts, metal plates and dielectrics in MIM capacitors), L is the inductance due to metal parts, and C_P represents parasitic capacitance across the capacitor's terminals (e.g., fringing capacitances between the edges of an MIM capacitor). The existence of both series and parallel capacitors warrant possibilities of both series and parallel resonances. As the main capacitance C is, in general, large compared to the parasitic capacitance C_P, it is expected that the dominant resonance is the series resonance occurred at low frequencies due to C. The series resonance due to the parasitic capacitance C_P occurs at high frequencies and considered a parasitic resonance.

By inspection of Figure 3.46(a), it is seen that at low frequencies, C_P may be neglected due to its small value, and the equivalent-circuit model can be simplified as shown in Figure 3.46(b). This circuit inherently possesses a series resonant frequency of

$$f_S = \frac{1}{2\pi\sqrt{LC}} \tag{3.111}$$

which limits the operating bandwidth of the capacitor to frequencies $f < f_S$.

At high frequencies, the main capacitance C, which is relatively much larger than C_P, may be neglected, and the resultant equivalent-circuit model is shown in Figure 3.46(c). This circuit forms a parallel resonator at a (parallel) resonant frequency given by

$$f_P = \frac{1}{2\pi\sqrt{LC_P}} \tag{3.112}$$

[17] Inductors may also have series resonance at sufficiently high frequencies.

The series resonant frequency f_S is in general much larger than its parallel counterpart f_P, and is the principal resonant frequency of capacitors. At series resonance, the capacitor becomes a resistor having resistance R, which is typically small for well-designed capacitors. Therefore, this capacitor exhibits a small loss and large reflection with a phase approximately equal to 180° at resonance when connected in series and parallel, respectively.

At parallel resonance, the capacitor can be treated as an equivalent resistor whose value is

$$R_P = \frac{1}{4\pi^2 R(f_P C_P)^2} \tag{3.113}$$

As R and C_P are normally small, the equivalent resistance R_P is large. This capacitor then causes a large reflection with a phase approximately equal to zero or a small loss when connected in series or parallel, respectively.

Table 3.5 lists the behavior of capacitors in relation to the resonant frequencies. Both the series and parallel resonances need to be assessed in the design of capacitors.

As for inductors, a resonance occurs if and only if the imaginary part of the impedance of the capacitor's equivalent circuit is equal to zero. At $f < f_S$, it is a capacitor. However, at $f_S < f < f_P$, it acts as an inductor, and at $f > f_P$, it becomes a capacitor again. At frequencies much higher than the capacitor's parallel resonant frequency, the capacitor's dimensions may be comparable to the operating wavelength, which effectively prohibits the capacitor to be classified as a lumped element. At these frequencies, it is considered distributed element and has multiple resonant frequencies, which are an inherence of distributed structures.

3.5.4 Quality Factor

Below the series resonant frequency, the behavior of capacitors is dominated by capacitance C and, hence, its Q is given by the usual formula

$$Q = \frac{\omega W_e}{P_L} = \frac{\frac{1}{2}\omega C V_{\text{rms}}^2}{\frac{1}{2}\frac{V_{\text{rms}}^2}{R}} = \frac{1}{\omega RC} \tag{3.114}$$

where ω is the operating frequency of the capacitor.

In practice, however, capacitors have parasitic inductances. These parasitic inductors, however, do not store energy – instead, working against the capacitors to reduce the total energy stored within the capacitors to $(W_e - W_m)$. Therefore, the Q of a capacitor at frequencies lower than the series resonant frequency is more accurately given by

$$Q = \frac{\omega(W_e - W_m)}{P_L} \tag{3.115}$$

where W_e denotes the time-average electric energy stored in the capacitor, W_m denotes the time-average magnetic energy corresponding to all the parasitic inductors, and P_L represents all the power loss or dissipated in the capacitor and parasitic inductors including those resulting from the conductor and substrate losses. Equation (3.115), however, suggests that Q would equal to zero at the self-resonant frequency of the capacitor, where the total energy stored by the capacitor vanishes. This frequency actually implies the transition frequency at which the capacitor becomes a pure resistor and above which inductive effect becomes

TABLE 3.5. Behavior of Capacitors with Respect to Resonant Frequencies

$f < f_S$	Capacitor
$f_S < f < f_P$	Inductor
$f > f_P$	Capacitor

more pronounced. The capacitor then becomes more of an inductor at frequencies greater than the resonant frequency and Eq. (3.115) is no longer valid.[18] Therefore, Eq. (3.115) can be used only at frequencies below the capacitor's series resonant frequency. The zero quality factor at resonance has no significance and usefulness in circuit design and performance since, at the resonant frequency, the capacitor ceases to operate capacitively and is no longer useful for circuit design. However, if the capacitor is used as a resonator at that frequency, then the total energy stored in that resonator would be positively contributed from both the electric energy and magnetic energy stored within the capacitor and (parasitic) inductor, hence leading to a nonzero Q.

Another useful equation for calculating Q of capacitors is based on the impedance of the capacitors and the corresponding Q formula is defined as the ratio between the total time-average stored energy and the time-average power loss or dissipated within the capacitors, as discussed in Section 3.4.4.

The total energy stored in capacitors can be expressed as Eq. (3.72). For capacitors, the parameter X in that equation is negative and the energy storing occurs during $0–T/4$ and $T/2–3\,T/4$. Accordingly, the stored energy can be derived as

$$W = \int_0^{T/4} v(t)i(t)dt = -\frac{XI_0^2}{\omega} \tag{3.116}$$

from which, Q can be derived, using Eq. (3.74) for the power dissipated, as

$$Q = \frac{\omega W}{P_L} = -\frac{X}{R} \tag{3.117}$$

or

$$Q = -\frac{\mathrm{Im}(Z)}{\mathrm{Re}(Z)} \tag{3.118}$$

which can also be written as

$$Q = \frac{\mathrm{Im}(Y)}{\mathrm{Re}(Y)} \tag{3.119}$$

where $Y = 1/Z$ is admittance of the element. Similar to the Q of inductors, the Q of capacitors can also be derived using Z_{11} or Y_{11} of one-port equivalent-circuit of capacitors, as shown in Figure 3.47(a), as

$$Q = -\frac{\mathrm{Im}(Z_{11})}{\mathrm{Re}(Z_{11})} = \frac{\mathrm{Im}(Y_{11})}{\mathrm{Re}(Y_{11})} \tag{3.120}$$

or

$$Q = \frac{\mathrm{Im}(Y_{11})}{\mathrm{Re}(Y_{11})} \tag{3.121}$$

using a two-port equivalent circuit, as shown in Figure 3.47(b), where Y_{11} is a parameter of the two-port admittance matrix. As for the case of inductors, care needs to be exercised using proper equation for Q. For instance, assuming a capacitor is represented by a π-equivalent network such as that in Figure 3.26(a), we can derive another expression for Q, making use of (3.118), as

$$Q = -\frac{\mathrm{Im}(Z_{11})}{\mathrm{Re}(Z_{11})} \tag{3.122}$$

where Z_{11} is a parameter of the two-port admittance matrix.

Similar to the inductor Q formulas (3.75)–(3.77) and (3.80), (3.82), at the self-resonant frequency of the capacitor, X is equal to zero, leading to zero quality factor by Eq. (3.117)–(3.121). Equations (3.117)–(3.121) can only be used for frequencies below the resonant frequency. Furthermore, these equations are derived

[18] Above the resonant frequency, where the inductor behaves dominantly as a capacitor, the Q is given as $Q = \omega(W_e - W_m)/P_L$.

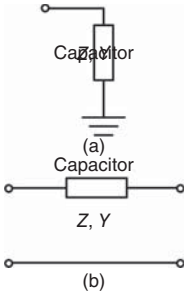


Figure 3.47. Capacitor represented by impedance Z or admittance Y connected as a one-port network (a) and in series as a two-port network (b).

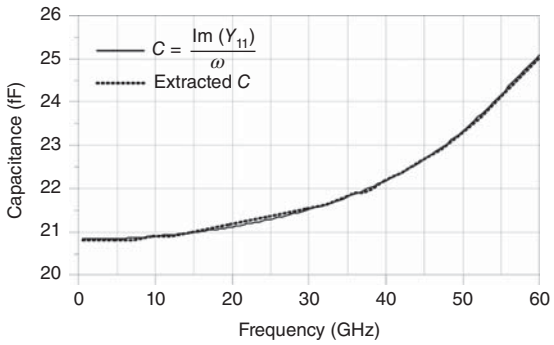


Figure 3.48. Capacitance of a CMOS on-chip 30-μm × 30-μm MIM capacitor.

using the total equivalent reactance and susceptance represented by a capacitor, respectively, which are different from those produced by a pure capacitor. Therefore, they should be used with care and their accuracy needs to be judged depending on particular capacitors and operating frequencies. At frequencies well below the resonant frequency, however, the capacitor exhibits more as a capacitive element, hence making (3.112)–(3.121) more accurate.

A useful remark is made here concerning the (main) capacitance C of capacitors. This capacitance, regardless of models used, can be approximately determined from the admittance matrix obtained from the calculated or measured S-parameters, based on the above discussion for Q, as

$$C = \frac{\text{Im}(Y_{11})}{\omega} \tag{3.123}$$

where, again, Y_{11} represents the admittance of the one-port equivalent-circuit of capacitors, as shown in Figure 3.47(a), or of the two-port equivalent circuit, as shown in Figure 3.47(b). This equation is particularly accurate and hence useful when a capacitor has small parasitic elements. As an example, we plot in Figure 3.48 the capacitance determined from (3.123) and that obtained from a fitting process for the 30-μm × 30-μm MIM capacitor discussed in Section 3.1 up to 60 GHz. The results show that these different capacitance calculations agree very well. It is recalled, from the earlier discussion, that this MIM capacitor behaves well as a capacitor to at least 60 GHz at which its Q is 18.6.

3.5.5 High Q Capacitor Design Considerations

On-chip capacitors on RFIC fabrication processes, like any capacitors, have losses and thus finite Q. The losses in capacitors are typically due to both metal and dielectrics used in capacitors with the earlier playing a more significant role. The Q of capacitors can be enhanced by proper design including use of metal layers having high conductivity, such as copper or gold, and thick metallization and use of low loss dielectrics.

Removal of the silicon area underneath the metal of a capacitor reduces the substrate loss, and hence increasing Q, and parasitic substrate capacitance. For interdigital capacitors, use of a dielectric having low relative dielectric constant beneath the conductor reduces the overall dielectric loss due to increased electric field in the air. The Q and resonant frequency of MIM capacitors are improved with increasing W/ℓ ratio due to the fact that, as this ratio is increased, both the dc resistance and parasitic inductance reduce.

3.6 LUMPED-ELEMENT RESISTORS

Resistors used in RFICs are planar resistors [21] and used mainly as matching elements and terminations. Typically, RFIC designers use resistors available from a CMOS process' library directly, rather than designing their own resistors. In general, the most important parameters of on-chip resistors for circuit design are their resistances, operating frequencies, and power-handling capabilities. Other considerations include tolerance, repeatability (from wafer to wafer or lot to lot), and temperature variation. Due to the existence of undesired parasitic capacitances, inductances, and resistances, particularly in a high RF range, resistors may resonate at certain high frequencies, rendering their ineffectiveness as resistors in RFICs. Careful simulations and/or measurements of on-chip resistors are therefore needed to assess not only their resistances and associated parasitics, but also their resonances, to be used for accurate circuit design.

3.6.1 Resistor Configurations

In RFICs, resistors are typically formed with a thin resistive material such as doped polysilicon deposited on a silicon substrate. Figure 3.49 shows a simplified configuration of typical polysilicon resistors. The contacts at the ends of poly resistors can be formed by directly connecting the doped polysilicon with a metal layer or using a silicide layer, with the latter resulting in much lower contact resistance and hence is preferred. Another type of resistors is n-well resistors. Figure 3.50 shows a simplified configuration of n-well resistors.

3.6.2 Basic Resistor Equations

Consider a typical chip resistor used in hybrid circuits or for off-chip applications, having a rectangular slab of resistive material terminated at two ends with metallic contacts, as shown in Figure 3.51. The resistance can be calculated exactly at DC or approximately at low frequencies using Eq. (3.33), assuming uniform current flow. This equation can be rewritten as

$$R = R_{\text{sh}} \left(\frac{\ell}{W} \right) \tag{3.124}$$

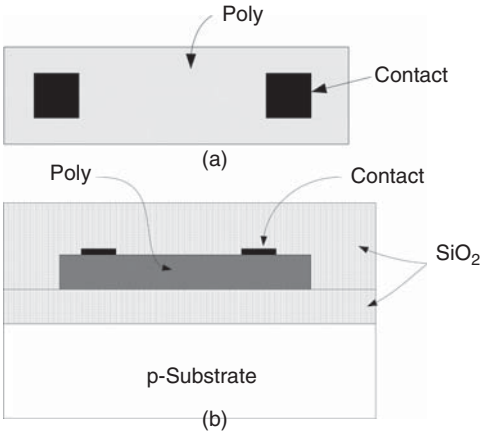


Figure 3.49. Top (a) and side (b) view of a poly resistor.

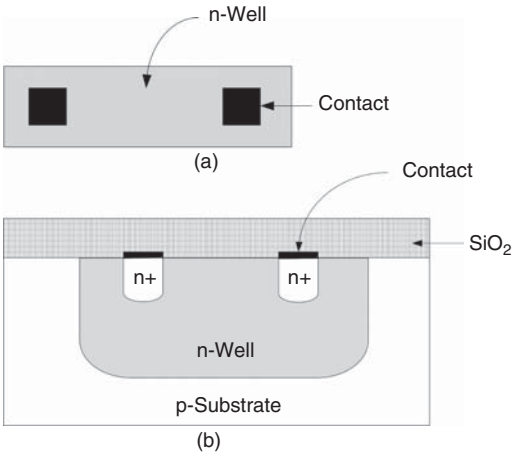


Figure 3.50. Top (a) and side (b) view of an n-well resistor.

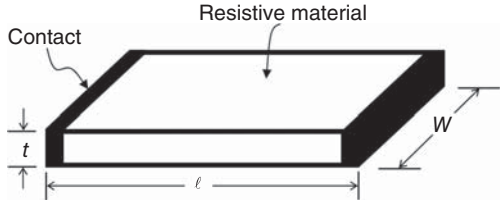


Figure 3.51. Typical chip resistor.

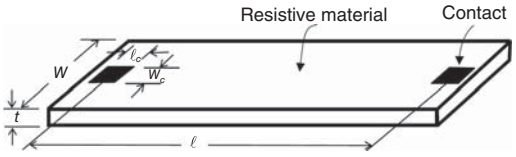


Figure 3.52. Simple on-chip resistor.

where ℓ and W are the length (between contacts) and width of the resistor, and R_{sh} is the sheet resistance, measured in *ohms* (Ω) *per square*, given as

$$R_{\text{sh}} = \frac{\rho}{t} \tag{3.125}$$

with ρ being the resistivity of the resistive material, assuming it is homogeneous material, and t being the material's thickness. As can be seen from Eq. (3.125), the length and width of planar resistors are not as important as their ratio, which is specified in *squares*. For instance, a resistor with equal length and width has one square and that with length as twice as width has two squares. The approximate resistance of on-chip resistors can thus be simply obtained as the product of the number of squares and sheet resistance of the resistors. In designing on-chip resistors, typically, a sufficiently large width is first determined to handle expected DC current, and the length is calculated accordingly based on the sheet resistance.

Figure 3.52 shows the top view of a simple on-chip resistor. Practical on-chip resistors have length and width different from the (designed) physical dimensions, resulting in resistance different from the designed value. Although the change in dimensions is small, it may be significant, particularly for physically small resistors in which effect of the change is more pronounced. Furthermore, current crowding at the contacts causes current flow nonlinearly at the contact edges, resulting in increased resistance. In typical on-chip resistors, the contacts do not extend across the resistor's width, making the current flow into and out of the resistor

nonuniform. The current must crowd inward as it approaches the contacts, causing an increase in the overall resistance. Contact resistance produced by the contacts also increases the overall resistance.

The changes in dimensions of practical on-chip resistors due to processing can be taken into account in the resistance calculation using their effective lengths and widths, which can be estimated using information provided in some CMOS processes. Since most resistors have a width much smaller than the length, the dominant factor affecting the resistance due to variation in dimensions is the change in width.

The basic equation (3.124) for resistance calculation can be modified for more accurate calculation of on-chip resistors as

$$R = R_{\text{sh}}\frac{L_e}{W_e} + \Delta R_d + \Delta R_w + \Delta R_c \tag{3.126}$$

where L_e and W_e are the effective length and width of the resistor, respectively; ΔR_d is the increase in resistance due to current crowding at the contacts; ΔR_w represents the change in resistance due to the difference between the widths of the contact and resistor; and ΔR_c stands for the contact resistance resulting from the potential difference between the resistive material and the contact. ΔR_d may be estimated from [22] as

$$\Delta R_d = 2\ln 4\frac{tR_{\text{sh}}}{W_e} \tag{3.127}$$

and ΔR_w is approximately given by [22]

$$\Delta R_w = \frac{R_{\text{sh}}}{\pi}\left[\frac{1}{k}\ln\left(\frac{k+1}{k-1}\right) + \ln\left(\frac{k^2-1}{k^2}\right)\right] \tag{3.128}$$

for $W_c \gg W_e - W_c$, where W_c is the width of the contact, respectively, and

$$k = \frac{W_e}{W_e - W_c} \tag{3.129}$$

The resistance increase ΔR_w is negligibly small for most practical on-chip resistors provided that the length is sufficiently large as compared to the width. The contact resistance ΔR_c can vary significantly from lot to lot and the value for two contacts can be approximately obtained from [21–23] as

$$\Delta R_c = \frac{2\sqrt{R_{\text{sh}}\rho_c}}{W_c}\coth\left(\sqrt{\frac{R_c}{\rho_c}}\right) \tag{3.130}$$

where ρ_c is the specific resistance of the contact interface ($\Omega\cdot\mu m^2$), which is dependent on the materials in contact and the processing conditions. Equation (3.130) also takes into account the resistance increase due to current crowding at the contacts resulting from the current flowing nonuniformly when entering or leaving the contact vertically.

Resistance change due to temperature variation is an important consideration in circuit design. This change can be estimated using

$$R(T) = R(T')[1 + 10^{-6}C_T(T - T')] \tag{3.131}$$

where $R(T)$ and $R(T')$ are the resistances at temperatures T and T' (°C), respectively; C_T is the temperature coefficient of the resistive material in parts per million per degree Celsius (ppm/°C).

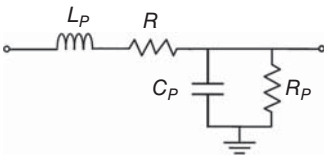


Figure 3.53. Simple equivalent-circuit model for on-chip resistors. R represents the resistor's total resistance, L_P represents the parasitic inductance associated with the contacts and connection, C_P represents the parasitic capacitance due to substrate, oxide layers and fringing fields, and R_P stands for the loss due to oxide layers and substrate.

3.6.3 Equivalent-Circuit Models of Resistors

An ideal resistor is represented by a pure resistance. Practical on-chip resistors, however, have undesired parasitics such as inductance due to metallic contacts and capacitances due to surrounding oxide layers, silicon substrate, and fringing fields of signals traversing along the resistors. Figure 3.53 shows a simple equivalent-circuit model for on-chip resistors. In the RF regime, on-chip resistors are actually distributed structures, just like any other passive elements. As such, more complicated distributed models for these resistors, such as those consisting of cascade of multiple circuits shown in Figure 3.51, may be needed for accurate circuit design.

REFERENCES

1. IE3D Program, Version 10.0, *Zeland Software Inc.*, Fremont, CA, USA.
2. Advanced Design System (ADS), Agilent, Palo Alto, California, USA.
3. W. B. Kuhn and N. M. Ibrahim, "Analysis of Current Crowding Effects in Multiturn Spiral Inductors," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-49, pp. 31–38, Jan. 2001.
4. R. E. Chaddock, "The Application of Lumped Element Techniques to High Frequency Hybrid Integrated Circuits," *Radio and Electronics Eng.*, Vol. 44, p. 414–420, 1974.
5. I. Bahl, *Lumped Elements for RF and Microwave Circuits*, Artech House, Norwood, Massachusetts, 2003, Chapters 2 and 5.
6. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, Cambridge, UK, pp. 48–49, 1998.
7. H. A. Wheeler, "Simple Inductance Formulas for Radio Coils," *Proc. IRE*, Vol. 16, pp. 1398–1400, Oct. 1928.
8. S. S. Mohan, et al., "Simple Accurate Expressions for Planar Spiral Inductances," *IEEE J. Solid States Circuits*, Vol. 34, pp. 1419–1424, Oct. 1999.
9. R. L. Bunch, D. I. Sanderson, and S. Raman, "Quality Factor and Inductance in Differential IC Implementation," *IEEE Microw. Mag.*, Vol. 3, pp. 82–92, Jun. 2002.
10. C. P. Yue and S. S. Wong, "Physical Modeling of Spiral Inductors on Silicon," *IEEE Trans. Electron Devices*, Vol. ED-47, pp. 560–568, Mar. 2000.
11. C. P. Yue et al., "A Physical Model for Planar Spiral Inductors on Silicon," *IEDM Tech. Digest*, pp. 155–158, Dec. 1996.
12. C. P. Yue and S. S. Wong, "A Study on Substarte Effects of Silicon-Based RF Passive Components," *1999 IEEE MTT-S Int. Microwave Symp. Digest*, pp. 1625–1628, Jun. 1999.
13. J. N. Burghartz and B. Rejaci, "On the Design of RF Spiral Inductors on Silicon," *IEEE Trans. Electron Devices*, Vol. ED-50, pp. 718–729, Mar. 2003.
14. J. M. Lopez-Villegas, et al., "Improvement of the Quality Factors of RF Integrated Inductors by Layout Optimization," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-48, pp. 76–83, Jan. 2000.
15. C. P. Yue 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, pp. 743–752, May 1998.
16. M. Danesh and J. R. Long, "Differentially Driven Symmetric Microstrip Inductors," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-50, pp. 332–341, Jan. 2002.

17. C. Nguyen and K. Chang, "Millimeter-Wave Planar Integrated Circuit Filters," in *Topics in Millimeter-Wave Guided Propagation*, (K. Button, ed., Academic Press, pp. 99–202, 1988.
18. M. Tran and C. Nguyen, "Modified Broadside-Coupled Microstrip Lines Suitable for MIC and MMIC Applications and a New Class of Broadside-Coupled Band-Pass Filters," *IEEE Trans. on Microwave Theory Tech.*, Vol. MTT-41, No. 8, pp. 1336–1342, Aug. 1993.
19. X. Z. Xiong and V. F. Fusco, "A comparison Study of EM and Physical Equivalent Circuit Modeling for MIM CMOS Capacitors," *Microw. Opt. Technol. Lett.*, Vol. 34, pp. 177–181, Aug. 2002.
20. C. Zhen, G. Lihui, Y. Mingbin, and Z. Yi, "A Study of MIMIM On-Chip Capacitor Using Cu/SiO2 Interconnect Technology," *IEEE Microwave Wireless Compo. Lett.*, Vol. 12, pp. 246–248, Jul. 2002.
21. A. Hastings, *The Art of Analog Layout*, pp. 158–159, Englewood Cliffs, New Jersey, Prentice Hall, 2006.
22. C. Y. Ting and C. Y. Chen, "A Study of the Contacts of a Diffused Resistor," *Solid-State Electronics*, Pergamon Press, Vol. 14, pp. 433–438, 1971.
23. H. Murrmann and D. Widmann, "Current Crowding on Metal Contacts to Planar Devices," *IEEE Trans. Electron. Dev.*, Vol. ED-16, No. 12, pp. 1022–1024, Dec. 1969.

PROBLEMS

3.1 A physical element may behave as both lumped and distributed structures and the transition frequency between these states is important for the RFIC design. Consider a microstrip line with the top conducting strip deposited on the topmost metal layer M7 and the ground plane on the metal layer M2 of a CMOS structure. Assume the ground plane is 10 times as wide as the top strip and the total thickness of the SiO_2 dielectric layers between M7 and M2 is 8 μm. The SiO_2 dielectric is assumed to have a relative dielectric constant of 3.8 and loss tangent of 0.004. The top and bottom conductors have metallization thickness and surface resistivity of 2 μm, 0.5 Ω/sq, and 0.5 μm, 0.8 Ω/sq, respectively. The width and length of the strip are 10 μm and 0.8 mm, respectively. The equivalent circuit of this microstrip line is given in Figure 3.2.

a) Calculate R, L, G, and C using an EM simulator.

b) Determine R, L, G, and C using the S-parameters of the microstrip line obtained from an EM simulator and those of the equivalent circuit using a circuit simulator. Describe the procedure and include all pertinent data that you use in arriving at the values for these parameters. Plot R, L, G, and C obtained from Part (a) and Part (b) versus frequency and wavelength of the microstrip line from 0 to 40 GHz in 5-GHz steps [plot corresponding results from Part (a) and Part (b) in the same chart]. Discuss the results and draw a conclusion, particularly the transition frequency between the lumped and distributed regions. You may use $R = 20\,\Omega$ and $G = 10^{-4}$ mhos as the initial guesses. The initial guesses for L and C may be calculated from $L \simeq Z_o \sin(\omega\ell/v)/\omega$ and $C \simeq \tan(\omega\ell/2v)/\omega Z_o$, where Z_o and v are the characteristic impedance and phase velocity, respectively, and ℓ is the length of the transmission line. You may also use the EM results in Part (a) for the initial guesses.

3.2 In RFICs operating in the microwave frequency range, inductors are sometimes realized using narrow-width, or high characteristic-impedance, transmission lines. These inductors, however, have a relatively large parasitic shunt capacitance, which degrades the circuit performance over a broad bandwidth. Lumped-element inductors, if can be designed in the same frequency range, have less parasitic capacitance and are more useful for wideband circuits. To demonstrate this, design a 1-nH inductor using:

a) A microstrip line having 5-μm width on the topmost metal layer and the ground plane on the metal layer M1 of the CMOS process profile described in Figure P3.1. Assume the ground plane is 10 times as wide as the top strip. Use an EM simulator to calculate the inductance and parasitic capacitance at 1 and 5 GHz.

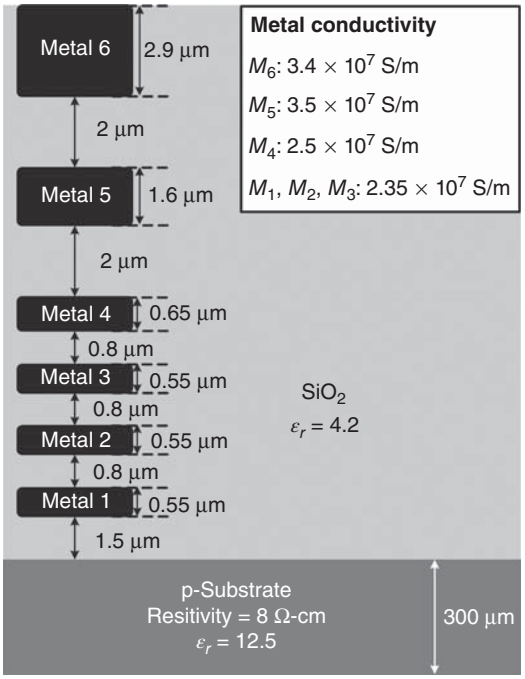


Figure P3.1.

b) An octagonal spiral inductor having a trace width and spacing of 10 and 2 μm, respectively, on the topmost metal layer of the CMOS process profile described in Figure P3.1. Use an EM simulator to calculate the inductance and parasitic capacitance at 1 and 5 GHz.

c) Compare the results in (a) and (b) and draw a conclusion.

3.3 On-chip MIM capacitors in CMOS processes may behave as both lumped and distributed elements (at extremely high frequencies). Use any submicron CMOS process of your choice, design or choose any MIM capacitor structures and/or dimensions (you may use an existing MIM capacitor in the design kit provided in your selected CMOS process) and demonstrate this phenomenon by calculating the capacitor model parameters using both an EM simulator and a circuit simulator. If you do not select MIM capacitors from some CMOS processes, then a possible structure that you can use is an MIM capacitor with the following parameters: size of 120-μm × 120-μum; oxide layer of 3-μm thick, relative dielectric constant of 3.8, and loss tangent of 0.0002; aluminum plates with top and bottom metallization of 1 and 0.6 μm, respectively. The bottom plate is separated from a grounded 300-μm silicon substrate by SiO_2 layers with a total thickness of 10 μm. Plot the model parameters versus frequency from 1 to 100 GHz. Plot Q versus frequency from 1 to 220 GHz. Describe the procedure and include all pertinent data that you use in arriving at the values for these parameters. Discuss the results and make a conclusion, particularly concerning the transition frequency at which a lumped-element operating mode is changed to a distributed mode. The equivalent-circuit models as shown in Figure 3.42(c), or any other model of your choice, can be used for the MIM capacitor.

3.4 Consider a simple equivalent-circuit model for inductors as shown in Figure P3.2. Derive an expression for the quality factor of these inductors using this model.

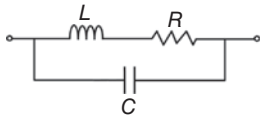


Figure P3.2.

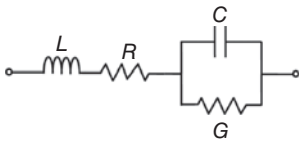


Figure P3.3.

3.5 Consider a simple equivalent-circuit model for capacitors as shown in Figure P3.3. Derive the quality factor for these capacitors using this model.

3.6 Consider the stacked inductors included in Table 3.2. Use any available submicron CMOS process with at least five metal layers (or the process profile shown in Figure P3.1), a circuit simulator, and an EM simulator, compute the main inductance, resonant frequency and Q of these inductors and of single-metal spirals having the same line with and spacing, and number of turns. Plot these calculated parameters versus frequencies from 0.5 to 10 GHz. Compare and discuss the results among these stacked spirals and to those calculated for the single-metal spirals on metal layer 5, particularly with respect to the inductance values, resonant frequencies and Q. Describe the procedure and include all pertinent data that you use.

3.7 Stacked spiral inductors with offset (or asymmetrical) spirals reduce the coupling effect and hence are more useful for RFICs than symmetrically stacked counterparts. Consider asymmetrically stacked spiral inductors on the CMOS process profile as shown in Figure P3.1 (or any available submicron CMOS process having at least six metal layers). The individual spirals on a stacked spiral are asymmetrically offset from each other and interconnected by vertical via-holes. Each individual spiral is octagonal with an outer diameter of 100 μm, trace width of 6 μm, and trace spacing of 2 μm. The trace of a spiral is offset from that of an adjacent spiral by 6 μm.

Metal layers	Number of turns for each spiral
6 (single layer)	4
6, 5	4
6, 2	4
6, 4, 2	4

a) Use an EM simulator to calculate and plot S_{11}, S_{21}, and S_{22} (dB), and main inductance and Q of these inductors versus frequency from 0.5 to 20 GHz (for multilayer spirals) and from 5 to 60 GHz (for single-layer spiral) in 0.5-GHz steps. List the following parameters of the inductors (corresponding to metal layers and number of turns) in a table: resonant frequency, maximum Q and corresponding frequency and inductance, highest inductance and corresponding frequency and Q, and suggested operating frequency range. Compare and discuss the results among these stacked spirals and to those calculated for the single-metal spiral on the topmost metal layer 6.

b) Assume the inductors are represented by an equivalent-circuit model as shown in Figure 3.22(c). Use a circuit simulator and the EM-simulation S-parameters obtained in Part (a) to determine the model's elements at 1 and 2 GHz for the inductors. Describe the procedure and include all pertinent data that you used. Compare and comment on the obtained values with those calculated in Part (a) if the equivalent-circuit models used in Part (a) (i.e., used by the EM simulator) and (b) are comparable.

3.8 Consider the same CMOS process and a single-layer octagonal spiral with the same outer diameter, trace width and trace spacing as described in Problem 3.7. Assume the spiral is on the topmost metal

layer 6. Use an EM simulator to calculate and plot Q and the inductance of this inductor versus frequency from 0.5 to 50 GHz in 0.5-GHz steps, and the resonant frequency with and without a solid ground shield on metal 1. Compare and discuss the results of Q, inductance and resonant frequency between these two cases. Are the Qs between the two cases significantly different at some frequencies? Provide your rationale as to why or why not the Qs are so different, if any.

3.9 Consider two rectangular spiral inductors made by copper. The dielectric surrounding the spirals is air. Each spiral has an inner radius of 50 μm, trace width of 10 μm, and trace spacing of 2 μm. The numbers of turns for these spirals are 3 and 5. Calculate and plot the inductances of these inductors versus frequencies from 0.5 to 20 GHz using Eqs. (3.45), (3.50), and an EM simulator. Compare and assess the accuracy of these equations with respect to the EM simulator.

3.10 Assume an on-chip inductor is modeled as a parallel circuit shown in Figure P3.4.

a) Prove that the inductor represented by that model behaves as a parallel resonator over a narrow frequency range around the resonator's resonant frequency. Derive the resistance, inductance, and capacitance of this equivalent parallel resonator in terms of R and L of the model and Q of the equivalent resonator.

b) Derive Q of the inductor represented by the model in Figure P3.4 in terms of R, L, and C.

3.11 Consider a π-network representing an on-chip inductor in a CMOS process, as shown in Figure P3.5. Assume the network has the following parameters at 10 GHz: $R = 2.81\,\Omega$, $L = 0.63$ nH, $R_1 = 68,579\ \Omega$, $C_1 = 0.0125$ pF, $R_2 = 58,778\,\Omega$, and $C_2 = 0.0102$ pF. The Q of the inductor is calculated using either Eq. (3.77), with Y being the admittance of the inductor (i.e., the admittance looking into the terminals of the π equivalent circuit), or Eq. (3.80) with Y_{11} being the element of a two-port network containing the inductor's equivalent circuit.

a) Compute Q at 10 GHz using Eq. (3.77).

b) Compute Q at 10 GHz using Eq. (3.80). Compare the result to that calculated in Part (a). Comment on these results.

c) Plot the admittance or Q versus frequency and, from which, determine the inductor's self resonant frequency. We assume herein the inductor's parameters are constant versus frequency.

3.12 Repeat Problem 3.11 assuming $R = 4\,\Omega$, $L = 1$ nH, $R_1 = R_2 = 2$ kΩ, $C_1 = C_2 = 0.02$ pF at 15 GHz. For Part (c), we assume the inductor's parameters are constant versus frequencies.

3.13 Design a 10-nH square spiral inductor for your RFIC using two different outer diameters of 300 and 400 μm. Assume that the trace width and spacing (edge-to-edge) are 10 and 2 μm, respectively. The spiral is on the topmost metal layer as shown in Figure P3.6. The cross-under metal strip has a thickness of 0.5 μm separated from the topmost metal by a 1.5-μm SiO_2 layer. The metal is copper having σ of

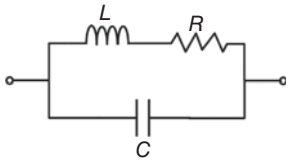


Figure P3.4.

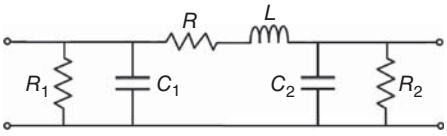


Figure P3.5.

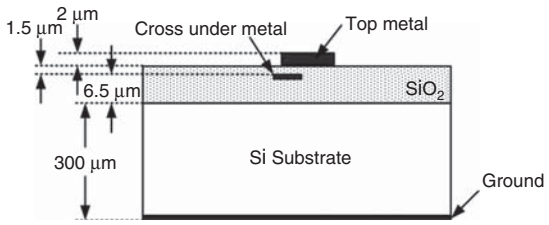


Figure P3.6.

5.8×10^7 S/m. The SiO_2 dielectric and grounded silicon substrate have relative dielectric constant (ε_r), loss tangent (tan δ) of 3.8, 0.0002 and 11.9, 0.02, respectively.

a) Calculate the inductance using Eqs. (3.45) and (3.50). Comment on the results.
b) Calculate the inductance using an EM simulator at 0.5 and 1 GHz and compare the result to those in Part (a). The EM simulation may produce the inductance along with other element values in a specific equivalent-circuit model used in the EM simulator. List all these element values as well and comment on the calculated results of these elements.
c) Calculate Q of this inductor using the EM simulator in Part (b) from 0.5 to 10 GHz. Comment on the results.
d) Calculate and plot the inductor's impedance versus frequency from 0.5 to 10 GHz.
e) Determine the inductor's resonant frequency.

3.14 Design a 5-nH rectangular spiral inductor to be used as an RF choke for an RFIC over the UWB of 3.1–10.6 GHz. The design frequency is the UWB center frequency of 6.85 GHz. Assume the spiral's metal structure is the same as that in Problem 3.13. The trace's width and spacing, outer diameter, and inner diameter (or number of turns) are the design parameters to be determined. As part of the design process, use an EM simulator to determine the main inductance value and quality factor at the design frequency, and the resonant frequency. Use the EM simulator to calculate and plot all the element values of the inductor's equivalent circuit from 0.5 to 20 GHz in 0.5-GHz steps. Show the layout of the inductor. You would need to make sure that the inductor's resonant frequency is well above the highest operating frequency of 10.6 GHz. Would you prefer to use this inductor or a quarter-wavelength transmission line for an RF choke over 3.1–10.6 GHz? Provide your rationale. Is the Q of this inductor important for RF chokes?

3.15 An on-chip square spiral inductor can be represented by the equivalent-circuit model shown in Figure 3.22(c). The spiral has trace width of 10 μm, trace spacing of 2 μm, maximum (outer) dimension of 170 μm, and five turns. The metal-dielectric structure for the spiral is the same as that of Problem 3.13. The model's parameters are: $R_{ds1} = R_{ds2} = 390\ \Omega$, $C_{s1} = C_{s2} = 67$ fF, $C_{d1} = 161$ fF, $C_{d2} = 150$ fF, $L = 7.2$ nH, $C_p = 26.5$ fF, and $R_s = 6.8\ \Omega$.

a) Calculate and plot the S-parameters of the model using the equivalent circuit with any circuit simulator (not EM simulator) from 1 to 10 GHz.
b) Calculate and plot the S-parameters of the actual inductor using an EM simulator from 1 to 10 GHz. Compare and discuss the results with respect to those in Part (a).
c) Determine new values for the model's parameters using the calculated S-parameters in Part (b). Describe the procedure and discuss the results. Ideally, the model should have constant parameter values across an interested frequency range. This, however, may not be possible across wide frequency ranges. Breaking up the interested frequency band into narrow subbands, or even using single frequencies, may be necessary.

3.16 Consider an on-chip square spiral inductor having a trace width of 10 μm, trace spacing of 2 μm, outer diameter of 210 μm, and six turns. The metal-dielectric structure for the spiral is the same as

that of Problem 3.13. Assume the inductor is represented by the equivalent-circuit model shown in Figure 3.22(c).

a) Calculate and plot the *S*-parameters of the inductor using an EM simulator from 1 to 15 GHz. Using the model included in the EM simulator, determine the self-resonant frequency, and maximum Q and corresponding frequency.

b) Use a circuit simulator along with the *S*-parameters obtained in Part (a) to determine the model's parameters, resonant frequency, and Q. Ideally, each element of the model should have a constant value across the interested frequency range (e.g., 1–15 GHz in this case) instead of different values at different frequencies. This, however, may not be possible across wide frequency ranges. Breaking up the interested frequency band into narrow sub-bands, or even using single frequencies, may be necessary. Possible initial guesses for the model are: $R_{ds1} = R_{ds2} = 100\ \Omega$, $C_{s1} = C_{s2} = 80$ fF, $C_{d1} = 45$ fF, $C_{d2} = 30$ fF, $L = 5$ nH, $C_p = 50$ fF, and $R_s = 3\ \Omega$. Describe the procedure and discuss the results. If you are not able to obtain (converged) element values above a certain frequency, then please provide the reasons. Compare the obtained resonant frequency and maximum Q with those calculated in Part a.

3.17 Assume sinusoidal excitations with period T and initial current phase of 0°, prove that:

a) An inductive element stores and releases energy during the intervals $T/4$–$T/2$, $3\,T/4$–T and 0–$T/4$, $T/2$–$3\,T/4$, respectively.

b) A capacitor stores and releases energy during the intervals 0–$T/4$, $T/2$–$3\,T/4$, and $T/4$–$T/2$, $3\,T/4$–T, respectively.

3.18 Assume sinusoidal excitations with period T and initial current phase of θ_o, determine the intervals during which the energy is stored and released for inductors and capacitors.

3.19 Derive Eq. (3.86).

3.20 Consider an inductor modeled as an equivalent circuit consisting of parallel R, L, and C. Derive the following equation for Q of the inductor:

$$Q = \frac{R}{\omega L}\left[1 - \left(\frac{\omega}{\omega_r}\right)^2\right]$$

where ω_r is the inductor's resonant frequency. Determine the frequency range under which this equation is valid.

3.21 Consider two 4-turn stacked octagonal spiral inductors on metal layers M6, M5, and M6, M2 of a submicron CMOS process given in Problem 3.7. Using an EM simulator to calculate the following parameters from 0.5 to 20 GHz in 0.5 GHz steps:

a) The mutual capacitance C56 between the spirals on M5 and M6, and C5S between M5 and silicon substrate. Discuss the results.

b) The mutual capacitance C26 between the spirals on M2 and M6, and C2S between M2 and silicon substrate. Discuss the results. Compare the results with those in Part (a) and draw some conclusion with respect to the different metal layers used for stacked inductors and frequencies.

3.22 Derive Eq. (3.34).

3.23 Provide detailed descriptions concerning the existence of eddy currents on conductors and within conductive substrates and their effects.

3.24 Derive Eq. (3.48).

3.25 Can the commonly used equivalent-circuit models of on-chip spiral inductors, as shown in Figure 3.22, be significantly improved? Describe a possible model that can represent these on-chip spiral inductors more accurately. Provide your rationale.

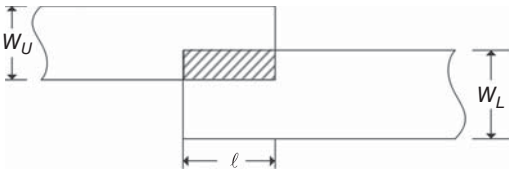


Figure P3.7.

3.26 Derive R, L, G, and C in the equivalent circuit shown in Figure 3.2 in terms of the parameters of the Z (impedance) or Y (admittance) matrix.

3.27 An offset broadside-coupled structure, as shown in Figure P3.7, provides an alternative for MIM-like capacitors in CMOS processes. Its usefulness lies in the fact that the offset can be chosen to achieve various capacitance values. Its equivalent circuit model, from the point of view of two-conductor transmission lines, can be represented by Figure 3.40(b). Derive an approximate expression for the capacitance C assuming W_U, W_L, and ℓ are very small compared to a wavelength of the corresponding broadside-coupled transmission line. It is assumed that W_U, W_L, and ℓ are relatively large compared to the distance d between the plates (or conductors).

3.28 Consider an on-chip MIM capacitor and its equivalent-circuit model as shown in Figure P3.8. The top (T) and bottom (B) plates have the same size. The relative dielectric constant and loss tangent of the SiO_2 layer between the plates and the Si substrate are 3.9 and 0.0001 and 11.7 and 0.005, respectively. In the model, C represents the main capacitance, modeled as a parallel-plate capacitance, R_s represents the parasitic resistance of contacts and metal plates, L_s is the parasitic inductance due to metal plates, and C_p denotes the capacitance between the bottom plate and the bottom of the Si substrate, and R_p represents the substrate resistance.

a) Determine the model's parameters by matching the S-parameters of the MIM capacitor calculated from its physical structure and the equivalent-circuit model described in Figure P3.8 using EM and circuit simulators from 0.5 MHz to 15 GHz, respectively. Initial values used in the fitting process may be determined using appropriate equations in Section 3.5.2. Plot and compare the model's S-parameters (magnitude and phase) and those from the EM simulation on the same Smith charts. It is noted that different (final) values for the model's parameters may be obtained using different initial values. As far as the overall circuit is concern, different values for the model's parameters

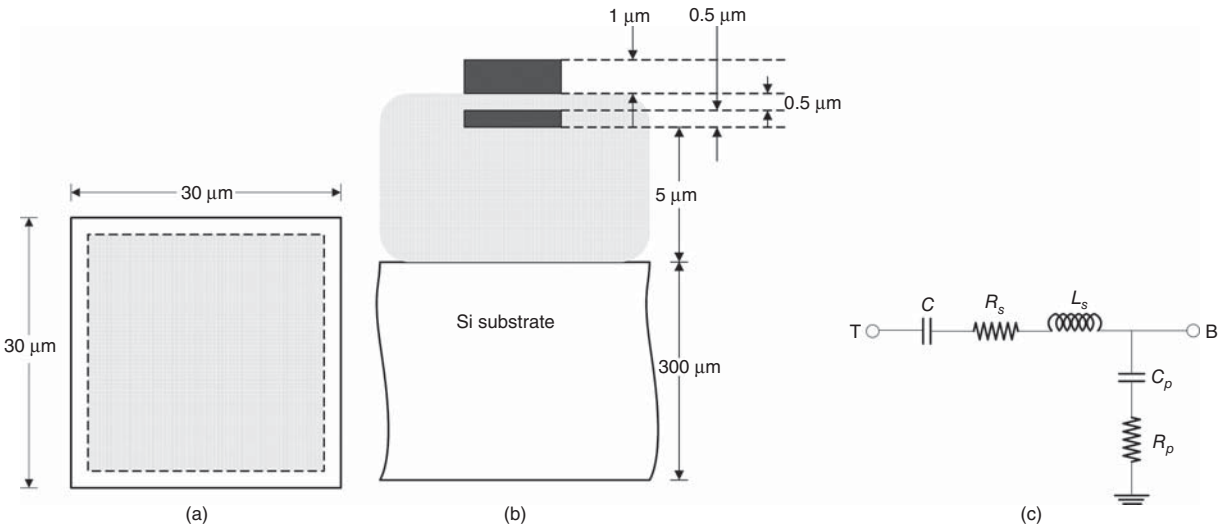


Figure P3.8.

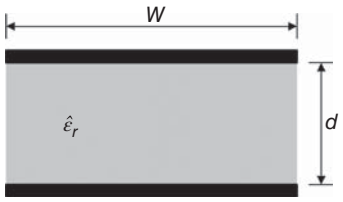


Figure P3.9.

(for the same *S*-parameters) do not affect the design of circuits employing the capacitor. However, they may produce different results for the capacitor's Q and resonant frequency.

b) Calculate Q at 1, 5, and 10 GHz using the quality-factor equation along with the equivalent-circuit model and an EM simulator. Compare and comment on the results.

c) Calculate the resonant frequency (GHz) using equation along with the equivalent-circuit model and EM simulator. Compare and comment on the results.

d) Calculate the impedance represented by the equivalent-circuit model and use it to determine the series and parallel resonant frequencies. Compare the results to those calculated using equations and comment.

3.29 Consider an MIM capacitor with length ℓ as a parallel-plate transmission line as shown in Figure P3.9. Assume that this capacitor can be modeled as the usual transmission-line equivalent circuit, but with R, L, G, and C representing the "total" (not per-unit-length) resistance, inductance, conductance, and capacitance. Assume also that the dielectric between the plates is given as $\hat{\varepsilon} = \varepsilon' - \varepsilon''$, the conducting plates have conductivity σ, and the width W is very large compared to the dielectric thickness d. Derive the "total" R, L, G, and C of the capacitor's model. (See also Problem 4.4 in Chapter 4).

3.30 Consider the spiral inductor with 300-μm outer diameter in Problem 3.13 and the MIM capacitor in Problem 3.28, whose peak Q is already determined using an EM simulator in these problems. Instead of these inductor and capacitor, you can also choose a spiral inductor and an MIM capacitor available in a sub-micron CMOS PDK. Consider three networks employing these inductor and capacitor: one consisting of series inductor and shunt capacitor, one consisting of series capacitor and shunt inductor, and one comprising series inductor and series capacitor. Use an EM simulator to calculate the Q of these three networks. Compare and discuss the results with respect to the EM-simulated Q of the individual inductor and capacitor.

3.31 Typical interdigital capacitors are realized on one metal layer as shown in Figure 3.39(c). These configurations, while uniplanar and simple to fabricate, occupy relative large size and do not offer much flexibility in circuit design. Using multiple metal layers along with via-holes inherent in CMOS processes may overcome these drawbacks. Suggest a different configuration for the interdigital capacitor using multiple metal layers in CMOS processes that is suitable for small, moderate and large capacitances. Provide details of the structure and its operation.

3.32 Variation of capacitance versus temperature due to dielectric, etc., can be a problem in circuit design. A small rate of change of capacitance value with temperature is desirable for good circuit stability versus temperature. Consider a CMOS capacitor having a capacitance of 10 pF and a rate of change of 100 ppm/°C, where ppm/°C stands for part-per-million per degree Celsius, calculate the variation of capacitance over a temperature change of 20 °C.

3.33 Besides capacitors realized using conductors and dielectrics, other forms of capacitors are also available in CMOS processes, including MOS capacitor realized using the gate capacitance of transistors and varactor providing variable capacitances. Describe these capacitors in details: their configurations, operations, possible capacitance range, and Q. Particularly, provide possible use of varactors for RFICs.

3.34 Consider a simple on-chip resistor as shown in Figure 3.53. Its resistance can be approximately calculated using Eq. (3.124). The length of the resistor, as noted in Figure 3.52, is measured between the centers of the contact pads. Would it more accurate to consider the length as between the inner edges or outer edges of the pads? Provide your rationale.

3.35 Consult an available design rule from a CMOS foundry. Describe all types of resistors used in CMOS processes of that particular foundry in details.

3.36 Consider a 50-Ω resistor in the library of an available CMOS process. Compute the parameters of the equivalent-circuit model, shown in Figure 3.53, for that resistor from 0.5 to 10 GHz using proper EM and circuit simulators. Compare and comment on the obtained resistance to 50 Ω and that calculated using the closed-form equations in Section 3.6.2.

CHAPTER 4

TRANSMISSION LINES

Transmission lines are important elements in radio-frequency integrated circuits (RFICs). The design of RFICs, especially at extremely high frequencies such as those in the upper end of the millimeter-wave range, would not be complete, or even optimum, without considering transmission lines. While the use of transmission lines in RFICs is, in general, not appealing due to their relatively large size as compared to on-chip lumped elements, they are inevitable at extremely high frequencies where lumped elements do not behave properly and/or have quality (Q) factors so low that render their effectiveness. Even at low radio-frequency (RF) ranges, where on-chip elements are preferred and used, RFIC designers should make use of the theory of transmission lines to properly implement lumped elements in circuits to achieve certain features. Therefore, it is crucial that RFIC engineers have sufficient knowledge in transmission lines and consider their possible use in RFICs. The basics of transmission lines can be found in various textbooks in electromagnetics, for example, Reference [1]. In this chapter, we will address the fundamentals of transmission lines for both single and multiconductor transmission lines including transmission-line equations and important transmission-line parameters such as characteristic impedance, propagation constant, phase velocity, effective relative dielectric constant, dispersion, loss, distortion, impedance, reflection coefficient, etc. This chapter also discusses synthetic transmission lines and commonly used printed-circuit transmission lines, particularly their use for RFICs.

4.1 ESSENTIALS OF TRANSMISSION LINES

To illustrate the possible need of considering transmission lines in RFICs, we consider a very simple interconnect consisting of two conductors, or one conductor and a ground plane, with the input and output of the two conductors connected to a voltage source and a load, respectively, as shown in Figure 4.1. We assume the source produces a sinusoidal signal $v_S(t) = V_S \cos(2\pi ft)$ at $f = 200$ MHz, the substrate supporting the interconnect has a relative dielectric constant $\varepsilon_r = 4.7$, and the length of the interconnect is $\ell = 0.1$ m. As

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

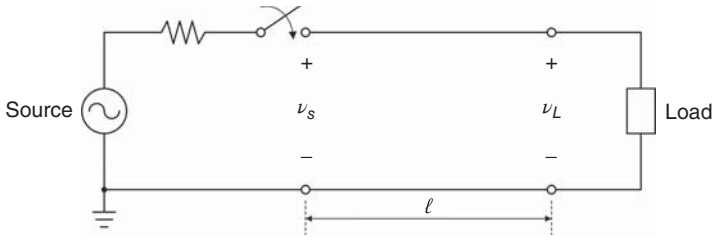


Figure 4.1. A simple interconnect acting as a transmission line.

the signal, or specifically the voltage wave, propagates along the interconnect with a velocity $v = c/\sqrt{\varepsilon_r} = 1.384 \times 10^8$ m/s, we have, with respect to time, the following results:

Time	Source end	Load end
$t = 0$	$v_S = V_S$	$v_L = 0$
$t = \frac{\ell}{v} = 0.72\,\text{ns}$	$v_S = 0.62\,V_S$	$v_L = V_S$

As can be seen, there is a voltage difference between the voltages at the source and load ends, which implies that the interconnect should be considered a transmission line and not simply an (interconnecting) conductor.

The source voltage at the instant $t_o = \ell/v$ can be written as

$$v_S(t_o) = V_S \cos(2\pi f t_o) = V_S \cos\left(2\pi \frac{v}{\lambda} \cdot \frac{\ell}{v}\right)$$
$$= V_S \cos\left(2\pi \frac{\ell}{\lambda}\right) \tag{4.1}$$

where λ is the wavelength. Equation (4.1) shows that when ℓ/λ is very small, an interconnect can be simply treated as "conductor" and circuit analysis can be employed. However, when ℓ/λ is sufficiently large, an interconnect should be considered a transmission line, or distributed element.

4.2 TRANSMISSION-LINE EQUATIONS

Transmission lines, like other elements in RFICs, possess certain electrical properties that are needed for circuit design. The characteristics of any transmission line are governed by a set of equations commonly known as transmission-line equations. To facilitate the analysis of transmission lines under different excitations and operating conditions, we classify these equations into two kinds: general transmission-line equations, which are valid for excitations of any time variation, and sinusoidal steady-state or time-harmonic transmission-line equations, which are applicable only when the excitation is a sinusoidal waveform and the transmission line is under a steady-state operation. The transmission-line parameters such as characteristic impedance, however, are the properties of a transmission line itself regardless of the excitation, provided that the magnitude of the excitation is within a certain limit.

4.2.1 General Transmission-Line Equations

A (single) transmission line exists only when it has two conductors. Consequently, a transmission line is normally represented as two parallel lines as shown in Figure 4.2(a). The transmission line is assumed to be uniform; that is, its geometry, including the conductor dimensions, spacing between them, and dielectric media around the conductors remain constant along the transmission-line length or its longitudinal (or axial)

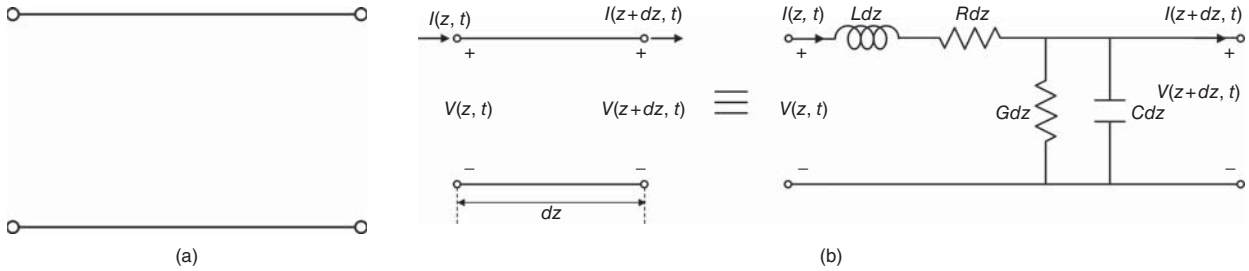


Figure 4.2. (a) A transmission line represented by two parallel conductors and (b) an equivalent circuit of an infinitesimal transmission line; $V(z,t)$, $V(z+dz,t)$ and $I(z,t)$, $I(z+dz,t)$ are the instantaneous voltages and currents at the input (z) and output ($z+dz$), respectively.

direction. From electromagnetic theory, it is known that a signal, or wave, propagating along the line is due to the interchange of the electric and magnetic energy. As a signal traverses the line, it carries with it the electric and magnetic fields. It is also known that as a signal propagates along the line, it suffers losses due to imperfect conductors and dielectrics constituting the line. Using circuit theory, we can then say that a transmission line can be characterized electrically as a combination of inductors, representing the magnetic energy, capacitors, representing the electric energy, resistors, representing the loss due to conductors, and other resistors representing the loss due to dielectrics. In view of this and the fact that circuit theory is valid only at frequencies where dimensions are very small compared to the wavelengths at those frequencies, we divide the transmission line into many infinitesimal elements, each having a length of dz much shorter than the operating wavelength. We can now represent each tiny element by an equivalent circuit consisting of a series inductance Ldz, a series resistance Rdz, a shunt capacitance Cdz, and a shunt conductance Gdz, as shown in Figure 4.2(b). L, R, C, and G are the series inductance (H/m), series resistance (Ω/m), shunt capacitance (F/m), and shunt conductance (S/m) per unit length of the transmission line, respectively. L and C account for the respective magnetic and electric energy of the signal stored in the transmission line. R and G account for the respective losses due to the imperfect conductors and dielectrics used in the transmission line. The loss due to the dielectrics is sometimes called leakage loss. Each transmission line is characterized by its own R, L, G, and C, or its series impedance $Z = R + j\omega L$ and shunt admittance $Y = G + j\omega C$ per unit length.

A (finite-length) transmission line can be considered consisting of many minuscule sections connected in cascade or, electrically, a continuous distribution of R, L, G, and C along the line. As such, a transmission line can be modeled as a cascade of many equivalent circuits, each corresponding to an infinitesimal section as shown in Figure 4.2(b). This distributed-circuit model for a transmission line is valid only when we assume no mutual impedance exists between adjacent sections making up the transmission line. This implies that, there is no mutual coupling between these adjacent elements. Consequently, we can assume that there are no longitudinal fields E_z and H_z for the wave traversing the line, assuming z is the direction of propagation. The electric and magnetic fields of the wave thus have only x and y components; that is, they lie in the planes orthogonally transverse to the propagation direction. Such a wave is called transverse electromagnetic (TEM) wave or mode. In practice, however, there exists very small longitudinal fields along transmission lines and the corresponding wave is referred to as quasi-TEM wave. All equations derived in this section are thus valid only for the TEM or quasi-TEM mode. Other waves having either E_z or H_z, or both, can also exist on transmission lines and are called higher-order waves or modes. These modes are significant only near the feed points to the transmission line, around a discontinuity on the line, or at very high frequencies. For most uniform transmission lines used in practice, TEM of quasi-TEM mode is the principal mode and is referred to as the transmission line's dominant mode.

Applying Kirchhoff's voltage law to the circuit in Figure 4.2(b), we obtain

$$V(z+dz,t) - V(z,t) + Ldz\frac{\partial I(z,t)}{\partial t} + RdzI(z,t) = 0 \tag{4.2}$$

Dividing Eq. (4.2) by dz and taking the limit as dz approaching zero (for infinitesimally short sections) leads to

$$\frac{\partial V(z,t)}{\partial z} + L\frac{\partial I(z,t)}{\partial t} + RI(z,t) = 0 \tag{4.3}$$

Similarly, we can apply Kirchhoff's current law to the circuit in Figure 4.2(b), divide the resultant equation by dz, and take the limit as dz approaches zero to obtain

$$\frac{\partial I(z,t)}{\partial z} + C\frac{\partial V(z,t)}{\partial t} + GV(z,t) = 0 \tag{4.4}$$

There exists a shunt current flowing across the medium between the two conductors due to the finite conductivity of the imperfect medium, which represents the leakage current. This leakage current is substantial for silicon substrate and has significant effects in RFICs, not only because of the resulting dielectric loss in silicon but also because of the coupling that current causes among different circuit elements. Taking the derivatives of Eqs. (4.3) and (4.4) with respect to z and t, respectively, gives

$$\frac{\partial^2 V(z,t)}{\partial z^2} + L\frac{\partial^2 I(z,t)}{\partial t \partial z} + R\frac{\partial I(z,t)}{\partial z} = 0 \tag{4.5}$$

$$\frac{\partial^2 I(z,t)}{\partial z \partial t} + C\frac{\partial V^2(z,t)}{\partial t^2} + G\frac{\partial V(z,t)}{\partial t} = 0 \tag{4.6}$$

Substituting Eqs. (4.4) and (4.6) into (4.5), we get

$$\frac{\partial^2 V(z,t)}{\partial z^2} - LC\frac{\partial^2 V(z,t)}{\partial t^2} - (RC + LG)\frac{\partial V(z,t)}{\partial t} - RGV(z,t) = 0 \tag{4.7}$$

Similarly, taking the derivatives of Eqs. (4.3) and (4.4) with respect to t and z, respectively, and combining them leads to

$$\frac{\partial^2 I(z,t)}{\partial z^2} - LC\frac{\partial^2 I(z,t)}{\partial t^2} - (RC + LG)\frac{\partial I(z,t)}{\partial t} - RGI(z,t) = 0 \tag{4.8}$$

For lossless transmission lines,[1] Eqs. (4.3), (4.4), (4.7), and (4.8) become, after letting $R = G = 0$:

$$\frac{\partial V(z,t)}{\partial z} + L\frac{\partial I(z,t)}{\partial t} = 0 \tag{4.9}$$

$$\frac{\partial I(z,t)}{\partial z} + C\frac{\partial V(z,t)}{\partial t} = 0 \tag{4.10}$$

$$\frac{\partial^2 V(z,t)}{\partial z^2} - LC\frac{\partial^2 V(z,t)}{\partial t^2} = 0 \tag{4.11}$$

$$\frac{\partial^2 I(z,t)}{\partial z^2} - LC\frac{\partial^2 I(z,t)}{\partial t^2} = 0 \tag{4.12}$$

Eqs. (4.3), (4.4) and (4.9), (4.10) are classified as the telegraphist's equations, and (4.7), (4.8) and (4.11), (4.12) are called the transmission-line equations. Equations (4.11) and (4.12) can also be viewed as the wave equations for lossless transmission lines – similar to the wave equations involving electric and magnetic fields derived from Maxwell's equation discussed in Section 2.6. All these equations are general and hold for

[1] A lossless transmission line is defined as one having perfect conductors and dielectrics. This results in no resistance and conductance per unit length ($R = G = 0$). Although all practical transmission lines are lossy, lossless transmission lines are sometimes used for easy and convenient analysis. This assumption is also used to illustrate some concepts in transmission lines without lost of generality.

voltages and currents with arbitrary time dependence and for any transmission line. They become simpler when the applied voltage or current is the sinusoidal time-varying function.

Another form of the transmission-line equations can be obtained by introducing new functions $V_e(z,t)$ and $I_e(z,t)$ defined by

$$V(z,t) = e^{-at}V_e(z,t) \tag{4.13}$$

and

$$I(z,t) = e^{-at}I_e(z,t) \tag{4.14}$$

where a is a constant. Substituting $V(z,t)$ from (4.13) and its derivatives into (4.7) yields

$$\frac{\partial^2 V_e(z,t)}{\partial z^2} - LC\frac{\partial^2 V_e}{\partial t^2} - (RC + LG - 2aLC)\frac{\partial V_e(z,t)}{\partial t} - [RG + a^2LC - a(RC + LG)V_e(z,t)] = 0 \tag{4.15}$$

Letting

$$a = \frac{RC + LG}{2LC} \tag{4.16}$$

and substituting it into (4.15), we get

$$\frac{\partial^2 V_e(z,t)}{\partial z^2} - LC\frac{\partial^2 V_e(z,t)}{\partial t^2} - \frac{1}{2}(RC - LG)V_e(z,t) = 0 \tag{4.17}$$

Similarly, using (4.14), (4.8), and (4.16), we obtain

$$\frac{\partial^2 I_e(z,t)}{\partial z^2} - LC\frac{\partial^2 I_e(z,t)}{\partial t^2} - \frac{1}{2}(RC - LG)I_e(z,t) = 0 \tag{4.18}$$

The solutions of the transmission-line equations (4.17) and (4.18) are related to the voltage and current by (4.13), (4.14), and (4.16). All the foregoing transmission-line and telegraphist's equations can also be derived from the field analysis using Maxwell's equations as described in APPENDIX A4.

To facilitate the demonstration of wave propagation on a transmission line, without lost of generality, we consider the lossless transmission-line equation (4.11). Its solution has the following form:

$$V(z,t) = V^+\left(t - \frac{z}{v}\right) + V^-\left(t + \frac{z}{v}\right) \tag{4.19}$$

where $v = 1/\sqrt{LC}$ and V^+ and V^- are arbitrary functions. Taking the derivatives of $V(z,t)$ with respect to t and z twice gives

$$\frac{\partial V(z,t)}{\partial t} = V^{+\prime} + V^{-\prime} \tag{4.20a}$$

$$\frac{\partial^2 V(z,t)}{\partial t^2} = V^{+\prime\prime} + V^{-\prime\prime} \tag{4.20b}$$

$$\frac{\partial V(z,t)}{\partial z} = -\frac{1}{v}V^{+\prime} + \frac{1}{v}V^{-\prime} \tag{4.20c}$$

$$\frac{\partial^2 V(z,t)}{\partial z^2} = \frac{1}{v^2}V^{+\prime\prime}\frac{1}{v^2}V^{-\prime\prime} \tag{4.20d}$$

where

$$V^{a\prime} \equiv \frac{dV^a}{d\left(t \mp \frac{z}{v}\right)} \tag{4.20e}$$

$$V^{a\prime\prime} \equiv \frac{d^2V^a}{d^2\left(t \mp \frac{z}{v}\right)} \tag{4.20f}$$

with the superscript a standing for $+$ or $-$. Substituting (4.20b) and (4.20d) into (4.11) verifies that $V(z,t)$ given by (4.19) is indeed a solution of the transmission-line equation (4.11). The velocity of wave propagation is obtained by setting the argument $t - z/v$ in (4.19) equal to a constant and differentiating it with respect to t, resulting in

$$\frac{dz}{dt} = v \tag{4.21}$$

which shows that the velocity is given by

$$v = \frac{1}{\sqrt{LC}} \tag{4.22}$$

We will see in Section 4.3 that this velocity is also given as

$$v = \frac{1}{\sqrt{\varepsilon\mu}} \tag{4.23}$$

for a uniform transmission line immersed in a homogeneous medium whose permittivity and permeability are ε and μ, respectively, which is the same as the velocity of a uniform plane wave propagating the same medium. This is expected as both the (dominant) transmission-line propagating wave and uniform plane wave are TEM wave.

The functions $V^+(t - z/v)$ and $V^-(t + z/v)$ constituting the voltage $V(z,t)$ can now be interpreted as the voltage waves propagating with velocity v in the $+z$ and $-z$ directions, respectively. They are commonly referred to as the respective forward and backward traveling waves. Similarly, we can show that

$$I(z,t) = I^+\left(t - \frac{z}{v}\right) + I^-\left(t + \frac{z}{v}\right) \tag{4.24}$$

is a solution of Eq. (4.12) with $I^+(t - z/v)$ and $I^-(t + z/v)$ representing the current waves traveling with velocity v in the $+z$ and $-z$ directions, respectively. Both forward and backward traveling waves can exist simultaneously on a transmission line at any instant and at any location on the line. Therefore, we can conclude that, at any location on a transmission line and at any time, the voltage and current (both magnitude and phase) are contributed by two separate voltage and current waves propagating in two opposite directions along the line, respectively.

Taking the partial derivative of the voltage given by Eq. (4.19) with respect to z and substituting it into Eq. (4.9), we get

$$\frac{\partial I(z,t)}{\partial t} = \frac{1}{Lv}\left[\frac{\partial V^+\left(t - \frac{z}{v}\right)}{\partial z} - \frac{\partial V^-\left(t + \frac{z}{v}\right)}{\partial z}\right] \tag{4.25}$$

Taking the integral of (4.25) with respect to t gives

$$I(z,t) = \frac{1}{Lv}\left[V^+\left(t - \frac{z}{v}\right) - V^-\left(t + \frac{z}{v}\right)\right] + I_o(z) \tag{4.26}$$

Now we take the partial derivatives of $V(z,t)$ given in Eq. (4.19) and $I(z,t)$ given in Eq. (4.26) with respect to t and z, respectively, and substitute the results into Eq. (4.10) to obtain

$$\frac{1}{Lv}\left[\frac{\partial V^+\left(t - \frac{z}{v}\right)}{\partial z} - \frac{\partial V^-\left(t + \frac{z}{v}\right)}{\partial z}\right] + \frac{\partial I_o(z)}{\partial z} = -C\left[\frac{\partial V^+\left(t - \frac{z}{v}\right)}{\partial t} + \frac{\partial V^-\left(t + \frac{z}{v}\right)}{\partial t}\right] \tag{4.27}$$

which is satisfied only if $I_o(z)$ is a constant. This constant is a possible DC portion of the solution of the current wave. It, however, does not contribute to the propagation of the current wave and is normally neglected. The current is therefore given from (4.26) as

$$I(z,t) = \frac{1}{Z_o}\left[V^+\left(t - \frac{z}{v}\right) - V^-\left(t + \frac{z}{v}\right)\right] \tag{4.28}$$

where $Z_o = Lv = \sqrt{L/C}$ is the characteristic impedance of the transmission line to be discussed in Section 4.3.

Figure 4.3 shows the forward traveling-voltage component $V^+(t - z/v)$ versus time at three distinct locations $z_1 < z_2 < z_3$ on a lossless transmission line, assuming a Gaussian pulse. Note that the voltage remains the same shape and amplitude as expected for a lossless line. It, however, is delayed according with the locations, thus depicting a wave movement along the line in the $+z$ direction. Similarly, wave propagation as a function of time in the $-z$ direction, representing backward-traveling wave, can also be demonstrated graphically by plotting $V^-(t + z/v)$.

Practical transmission lines are lossy and the voltage and current waves or signals will be attenuated as they propagate along the line.

4.2.2 Sinusoidal Steady-State Transmission-Line Equations

We now derive another form of transmission-line equations by assuming that the applied voltage varies sinusoidally with time and the transmission line is operated under a steady-state condition (i.e., no transients). Using the phasor representations, we can write the instantaneous voltage and current, with reference to $\cos \omega t$, as

$$V(z,t) = \mathrm{Re}[V(z)e^{j\omega t}] \tag{4.29}$$

$$I(z,t) = \mathrm{Re}[I(z)e^{j\omega t}] \tag{4.30}$$

where $V(z)$ and $I(z)$ represent the phasor voltage and current, respectively, which are functions of location (z) only, and Re (.) denotes the real part. Substituting Eqs. (4.29), (4.30) into Eqs. (4.3), (4.4) and (4.7), (4.8),

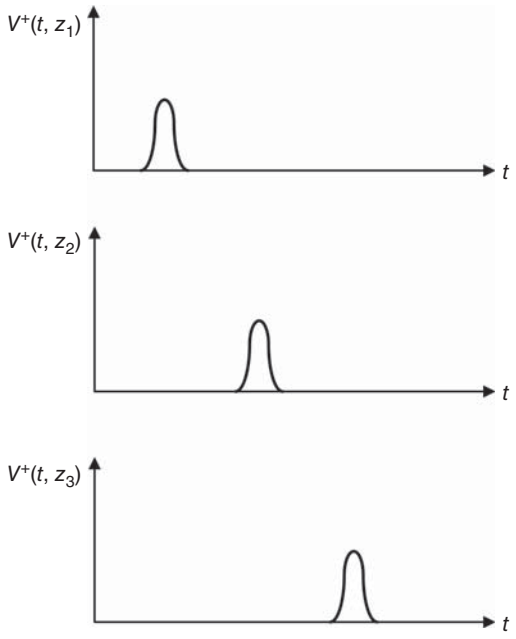


Figure 4.3. Illustration of a forward-travelling (Gaussian-pulse) voltage wave.

we obtain

$$\frac{dV(z)}{dz} + (R + j\omega L)I(z) = 0 \tag{4.31}$$

$$\frac{dI(z)}{dz} + (G + j\omega C)V(z) = 0 \tag{4.32}$$

$$\frac{d^2V(z)}{dz^2} - \gamma^2 V(z) = 0 \tag{4.33}$$

$$\frac{d^2I(z)}{dz^2} - \gamma^2 I(z) = 0 \tag{4.34}$$

where

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{4.35}$$

As will be seen later, γ dictates the propagation of voltage and current waves along the line and hence is called the propagation constant of the transmission line. Similarly, α (Neper/m or Np/m) and β (rad/m) determines the attenuation and phase per unit length of the transmission line and are called the attenuation and phase constant, respectively. α in Np/m can be converted into α in decibel per m by α (dB/m) = 8.686α (Np/m).

Equations (4.31)–(4.34) can also be obtained by replacing the partial derivative with respect to t, $\partial/\partial t$, with $j\omega$ in Eqs. (4.3), (4.4) and (4.7), (4.8), as discussed in Section 2.6 for the time-harmonic wave equations. Equations (4.31) and (4.32) are called the sinusoidal steady-state or time-harmonic telegraphist's equations, and (4.33) and (4.34) are known as the sinusoidal steady-state or time-harmonic transmission-line equations or the transmission-line equations in the frequency domain. The transmission-line equations (4.33) and (4.34) are especially used to analyze transmission lines operating under sinusoidal steady state, which is encountered in continuous-wave (CW) operations of electrical circuits. It should also be noted that all these telegraphist's and transmission-line equations are applicable to any transmission lines. For lossless transmission lines, these equations become, after setting $R = G = 0$ in Eqs. (4.31)–(4.34):

$$\frac{dI(z)}{dz} + j\omega C V(z) = 0 \tag{4.36}$$

$$\frac{dV(z)}{dz} + j\omega L I(z) = 0 \tag{4.37}$$

$$\frac{d^2V(z)}{dz^2} + \beta^2 V(z) = 0 \tag{4.38}$$

$$\frac{d^2I(z)}{dz^2} + \beta^2 I(z) = 0 \tag{4.39}$$

Equations (4.33) and (4.34) are of the same form of the one-dimensional wave equations seen in Section 3.6 and have the following well-known solutions:

$$\begin{aligned} V(z) &= V_o^+ e^{-\gamma z} + V_o^- e^{\gamma z} = V_o^+ e^{-\alpha z} e^{-j\beta z} + V_o^- e^{\alpha z} e^{j\beta z} \\ &= V^+(z) + V^-(z) \end{aligned} \tag{4.40}$$

$$\begin{aligned} I(z) &= I_o^+ e^{-\gamma z} + I_o^- e^{\gamma z} = I_o^+ e^{-\alpha z} e^{-j\beta z} + I_o^- e^{\alpha z} e^{j\beta z} \\ &= I^+(z) + I^-(z) \end{aligned} \tag{4.41}$$

respectively. $V^+(z)$ and $I^+(z)$ are commonly referred to as the forward-traveling voltage and current waves (in the $+z$ direction), and $V^-(z)$ and $I^-(z)$ are called the backward-traveling waves (in the $-z$ direction). The

total instantaneous voltage and current at any location on the line and at any time can be written, using Eqs. (4.29), (4.30), (4.35), (4.40), and (4.41), as

$$V(z,t) = V_o^+ e^{-\alpha z} \cos\left[\omega\left(t - \frac{\beta}{\omega}z\right)\right] + V_o^- e^{\alpha z} \cos\left[\omega\left(t + \frac{\beta}{\omega}z\right)\right] \tag{4.42}$$

$$I(z,t) = I_o^+ e^{-\alpha z} \cos\left[\omega\left(t - \frac{\beta}{\omega}z\right)\right] + I_o^- e^{\alpha z} \cos\left[\omega\left(t + \frac{\beta}{\omega}z\right)\right] \tag{4.43}$$

Note that the first and second terms in the right-hand side of Eqs. (4.42) and (4.43) correspond to the forward traveling waves $V^+(t - z/v)$, $I^+(t - z/v)$ and backward traveling wave $V^-(t + z/v)$, $I^-(t + z/v)$ given in Eqs. (4.19) and (4.24), respectively. These waves travel at a velocity of

$$v = \frac{\omega}{\beta} \tag{4.44}$$

as obtained by comparing (4.19) to (4.42) or (4.24) to (4.43). Equations (4.42) and (4.43) indicate that as the forward or backward voltage and current wave travel along the line at any instant and frequency, their amplitudes at a particular location z are reduced exponentially in accordance with $-\alpha|z|$ and their phases are undergone a phase change of βz. α and β determines the attenuation and phase per unit length of the transmission line and are therefore called the attenuation and phase constant, respectively. $\gamma = \alpha + j\beta$ dictates the propagation of voltage and current waves along the line and hence is called the propagation constant of the transmission line.

Similar to Figure 4.3, we can sketch the instantaneous (sinusoidal) voltage $V(z,t)$ versus time at different locations along the line to demonstrate (again) that the voltage travels along the line as time progresses. Note that in contrary to the circuit theory, in which voltage and current are assumed stationary and constant along the conductor of a transmission line, they are indeed (propagating) waves and are a function of the transmission line's characteristics.

4.3 TRANSMISSION-LINE PARAMETERS

As we recall, components in electronic circuits are physical elements, and we use electrical parameters to describe them so that they can be used in circuit analysis and design. For example, an inductor is described by an electrically equivalent inductance; a resistor is represented by a resistance, etc. Similarly, a transmission line is a physical structure and it is also described by certain electrical parameters. The most useful parameters of a transmission line, at least for circuit design and analysis, are characteristic impedance, propagation constant or attenuation and phase constants, and velocity. The phase constant and velocity are directly related and, many times, we also use the parameter "effective relative dielectric constant" in lieu of these. Transmission-line parameters can be determined from the per-unit-length inductance, resistance, capacitance and conductance of the transmission line described in Section 4.4. They are independent of the time variation of the applied voltage or current and, in general, the magnitude of the applied signal. As such, we can derive expressions for the transmission-line parameters considering a transmission line operating under a sinusoidal steady state. This would facilitate the solution process. To aid the use of these expressions, we divide into three cases: lossy (or general), lossless, and low loss transmission lines.

4.3.1 General Transmission Lines

For any transmission line operated under a sinusoidal steady state, we can write the phasor current using Eq. (4.31) as

$$I(z) = -\frac{1}{R + j\omega L}\frac{dV(z)}{dz} \tag{4.45}$$

Taking the derivative of the voltage in Eq. (4.40) and substituting it into Eq. (4.45), we get

$$I(z) = \frac{\gamma}{R + j\omega L}(V_o^+ e^{-\gamma z} - V_o^- e^{\gamma z}) \tag{4.46}$$

Comparing Eqs. (4.41) and (4.46) and making use of (4.35) gives

$$Z_o \triangleq \frac{V_o^+}{I_o^+} = -\frac{V_o^-}{I_o^-} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{4.47}$$

which is defined as the characteristic impedance of the transmission line (ohms). Its reciprocal

$$Y_o \triangleq \frac{I_o^+}{V_o^+} = -\frac{I_o^-}{V_o^-} = \sqrt{\frac{G + j\omega C}{R + j\omega L}} \tag{4.48}$$

is called the characteristic admittance of the line (mhos). Note that the characteristic impedance and admittance, as expected, are generally complex as can be seen in (4.47) and (4.48). It is important to note that these parameters are simply mathematically defined quantities and, as expected through their expressions in Eqs. (4.47) and (4.48), they are not and cannot be used as the normal impedance and admittance encountered in the electrical circuit theory.[2] The propagation constant $\gamma = \alpha + j\beta$ of the transmission line is given in Eq. (4.35), where α and β, as described in Section 4.2.2, are the attenuation and phase constants of the line, respectively. They are obtained by taking the respective real and imaginary parts of the propagation constant. It should be noted that the attenuation constant is contributed by the conductor attenuation constant (due to imperfect conductors) and the dielectric attenuation constant (due to imperfect dielectrics), which will be examined in Section 4.5 . As the per-unit-length resistance R, inductance L, conductance G, and capacitance C of a transmission line are different for different transmission lines, the characteristic impedance Z_o and propagation constant γ are also different as well. In practice, Z_o and γ (or α and β) are normally used to characterize a transmission line instead of R, L, G, and C.

A closed-form expression for the attenuation constant can also be derived based on the time-average transmitted power and power loss per unit length on the line. To that end, we consider an infinitely long transmission line and write the phasor voltage and current at any location, from Eqs. (4.40), (4.41), as

$$V(z) = V_o^+ e^{-\alpha z} e^{-j\beta z} \tag{4.49}$$

$$I(z) = I_o^+ e^{-\alpha z} e^{-j\beta z} \tag{4.50}$$

The time-average power or, simply, average power transmitted at any point on the line is given by

$$P_T(z) = \frac{1}{2}\mathrm{Re}[V(z)I^*(z)] \tag{4.51}$$

which becomes, after using Eqs. (4.49) and (4.50) and replacing $I_o^+ = V_o^+/Z_o$ with $Z_o = R_o + jX_o$,

$$\begin{aligned} P_T(z) &= \frac{1}{2}\mathrm{Re}\left[\frac{\left(V_o^+\right)^2}{R_o - jX_o} e^{-2\alpha z}\right] \\ &= \frac{(V_o^+)^2 R_o}{2|Z_o|^2} e^{-2\alpha z} \end{aligned} \tag{4.52}$$

[2]For example, multiple transmission lines connected in cascade are not equivalent to a single transmission line, whose characteristic impedance is the sum of the individual lines' characteristic impedances.

Let P_{T0} being the power flow at $z = 0$ and $P_{T0} - \Delta P_T$ being the transmitting power at $z = \Delta z$. The average power loss per unit length is obtained as the difference in these powers over the distance, which is essentially the rate of decrease of the average transmitted power as the wave travels along the line and can, therefore, be derived from Eq. (4.52) as

$$P_L(z) = \lim_{\Delta z \to 0} \frac{(P_{T0} - \Delta P_T) - P_{T0}}{\Delta z} = -\frac{dP_T(z)}{dz} = 2\alpha P_T(z) \tag{4.53}$$

from which

$$\alpha = \frac{P_L(z)}{2P_T(z)} \tag{4.54}$$

Note that z is arbitrary and so, for uniform transmission lines whose geometry along the direction of propagation is constant, the power loss per unit length itself is independent of z. This power loss is due to the currents flowing through the resistance R and the conductance G and, hence, can be determined from Figure 4.2(b) as

$$P_L(z) = \frac{1}{2}[R|I(z)|^2 + G|V(z)|^2] \tag{4.55}$$

which becomes, after utilizing Eqs. (4.49) and (4.50) and $I_o^+ = V_o^+/Z_o$,

$$P_L(z) = \frac{(V_o^+)^2}{2|Z_o|^2}[R + G|Z_o|^2]e^{-2\alpha z} \tag{4.56}$$

The attenuation constant can now be obtained from Eqs. (4.52), (4.54), and (4.56) as

$$\alpha = \frac{1}{2R_o}[R + G|Z_o|^2] \tag{4.57}$$

For transmission lines fabricated using good dielectrics and conductors, which are normally used in practice, the rate of change of the attenuation constant with frequency is, in general, relatively small. At low frequencies, the attenuation constant is very small and may be neglected. However, the attenuation of a transmission line cannot be ignored when the line is either long, used at high frequencies, or carries a large amount of power creating heat due to $\frac{1}{2}(I^2R + V^2G)$ that must be dissipated. It is noted that the attenuation calculated from (4.57) neglects the attenuation due to radiation. We can define an additional resistance representing the loss due to radiation and include it in the per-unit-length resistance R, and hence Z_o, to account for the loss due to radiation. However, radiation loss is usually very small in practical transmission lines and is, therefore, normally neglected. For some RFICs that implement transmissions directly on silicon substrate or without proper shield from silicon substrate, the attenuation constant, however, cannot be neglected due to substantial loss caused by the silicon substrate. As this attenuation constant is large, it not only reduces the signal amplitude significantly, but also causes large (substrate) coupling among circuit elements, degrading the circuit performance and, if not suitably taken care of, may cause the circuit not functioning properly under certain conditions.

The velocity can be derived, by letting the phase $\omega t - \beta z$ of the forward-traveling wave in Eq. (4.42) equal to a constant and differentiating it with respect to time, as

$$v = \frac{dz}{dt} = \frac{\omega}{\beta} \tag{4.58}$$

This velocity corresponds to a constant phase of the wave and so is normally called the phase velocity or, simply velocity, of the wave propagating on the line. It indicates how fast points of equal phase on the wave move along the transmission line. As all points move at the same velocity, the phase velocity describes the rate at which the wave propagates along the transmission line and, hence, can be considered the velocity of

propagation. The distance, over which the phase is changed by 2π radians at any instant, is defined as the wavelength and is thus derived as

$$\lambda = \frac{2\pi}{\beta} = \frac{v}{f} \tag{4.59}$$

where f denotes the frequency. Besides phase velocity, *group velocity* is also used as another kind of velocity. The group velocity is discussed in Section 4.7.

4.3.2 Lossless Transmission Lines

Although all practical transmission lines have loss, lossless transmission lines are sometimes used for easy and fast circuit analysis and design. This assumption is also often used to illustrate some concepts in transmission lines without lost of generality. A lossless transmission line is defined as one having perfect conductors and dielectrics. This results in no resistance and conductance per unit length; that is, $R = G = 0$. Setting $R = G = 0$ in Eqs. (4.47) and (4.35), we obtain

$$Z_o = \sqrt{\frac{L}{C}} \tag{4.60}$$

$$\gamma = j\beta = j\omega\sqrt{LC} \tag{4.61}$$

The velocity is obtained from (4.22) and (4.61) as

$$v = \frac{1}{\sqrt{LC}} = \frac{\omega}{\beta} \tag{4.62}$$

4.3.3 Low Loss Transmission Lines

For most practical transmission lines, the per-unit-length resistance R and conductance G are very small as compared to the per-unit-length reactance ωL and susceptance ωC, respectively. Under these conditions, a transmission line is considered low loss transmission line and approximate closed-form equations can be derived for the characteristic impedance and attenuation and phase constants. Equation (4.47) can be rewritten as

$$Z_o = \sqrt{\frac{L}{C}}\left(1 + \frac{R}{j\omega L}\right)^{1/2}\left(1 + \frac{G}{j\omega C}\right)^{-1/2} \tag{4.63}$$

Applying the binomial series[3]

$$(1+x)^a = 1 + ax + \frac{a(a-1)}{2!}x^2 + \cdots + \frac{a(a-1)\cdots(a-n+1)}{n!}x^n + \cdots \tag{4.64}$$

where x and a are real numbers, and retaining up to second-order terms, we obtain

$$Z_o \cong \sqrt{\frac{L}{C}}\left(1 + \frac{R}{j2\omega L}\right)\left(1 - \frac{G}{j2\omega C}\right) + \frac{1}{8}\left(\frac{R}{\omega L}\right)^2 - \frac{3}{8}\left(\frac{G}{\omega C}\right)^2 \tag{4.65}$$

Expanding Eq. (4.65) and making use of the fact that $R \ll \omega L$ and $G \ll \omega C$ yields

$$Z_o \cong \sqrt{\frac{L}{C}}\left[\left(1 + \frac{1}{8}\left(\frac{R}{\omega L}\right)^2 - \frac{3}{8}\left(\frac{G}{\omega C}\right)^2 + \frac{RG}{4\omega^2 LC}\right) + j\left(\frac{G}{2\omega C} - \frac{R}{2\omega L}\right)\right] \tag{4.66}$$

[3]This series is convergent if $|x| < 1$ and divergent when $|x| > 1$.

Similarly, we can rewrite Eq. (4.35) and apply the binomial expression (4.64) and the conditions $R \ll \omega L$ and $G \ll \omega C$ to get

$$\gamma \cong \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) + j\omega\sqrt{LC}\left[1 + \frac{1}{8}\left(\frac{R}{\omega L}\right)^2 + \frac{1}{8}\left(\frac{G}{\omega C}\right)^2 - \frac{RG}{4\omega^2 LC}\right] \tag{4.67}$$

from which, we can obtain

$$\alpha \cong \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) \tag{4.68}$$

$$\beta \cong \omega\sqrt{LC}\left[1 + \frac{1}{8\omega^2}\left(\frac{R}{L} - \frac{G}{C}\right)^2\right] \tag{4.69}$$

For most engineering design purposes, it is sufficient to retain only the first-order terms in the binomial expansions. Under this approximation, Eqs. (4.66), (4.68), and (4.69) reduce to

$$Z_o \cong \sqrt{\frac{L}{C}}\left[1 + j\left(\frac{G}{2\omega C} - \frac{R}{2\omega L}\right)\right] \tag{4.70}$$

$$\alpha \cong \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) \tag{4.71}$$

$$\beta \cong \omega\sqrt{LC} \tag{4.72}$$

The phase velocity is then obtained from (4.58) as

$$v \cong \frac{1}{\sqrt{LC}} \tag{4.73}$$

It is apparent that the phase constant and velocity of a low loss transmission line are approximately equal to those of a lossless line.

It should be mentioned here that, for transmission lines implemented directly on (or near) silicon substrate or without proper shield from silicon substrate, the transmission lines cannot be considered low loss and the above-derived equations for low loss transmission lines may not be applicable.

4.4 PER-UNIT-LENGTH PARAMETERS *R*, *L*, *C*, AND *G*

In the previous sections, we have learned that a transmission line and all of its parameters are characterized by its series inductance L, series resistance R, shunt capacitance C, and shunt conductance G per unit length. We now derive expressions for these per-unit-length parameters.

4.4.1 General Formulation

We consider an infinitely long transmission line consisting of two conductors C_1 and C_2 as shown in Figure 4.4. The inductance per unit length is contributed by two parts, one from the conductors as if they were perfect and another from the skin effect resulted from the imperfect conductors. The earlier inductance portion is called the external inductance and the latter is referred to as the internal inductance. The external inductance

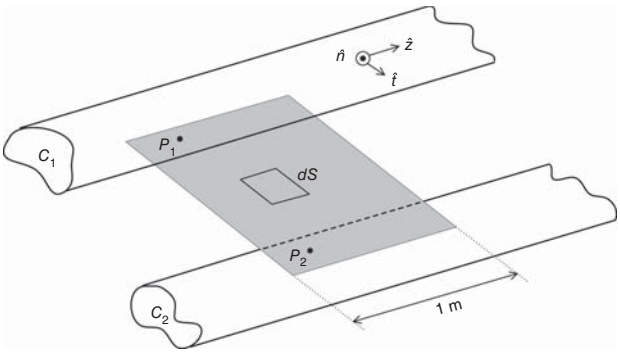


Figure 4.4. A transmission line consisting of two conductors C_1 and C_2. $\hat{n}$ is the unit vector normal to the conducting surface; $\hat{t}$ is the unit vector tangential to the surface; and $\hat{z}$ is the unit vector along the longitudinal direction.

L_e, capacitance C, and conductance G per unit length are given, under static conditions,[4] as

$$L_e = \frac{\psi}{I_o} \tag{4.74}$$

$$C = \frac{Q}{V_o} \tag{4.75}$$

$$G = \frac{I_G}{V_o} \tag{4.76}$$

where ψ is the total magnetic flux per unit length linking the current, I_o is the total current on the conductor C_1, V_o is the voltage difference between the two conductors, Q is the total charge per unit length on the conductor C_1, and I_G represents the current flowing across the conductance G, which is the total shunt conduction current per unit length. Note that L_e, C, and G are not a function of frequency, assuming frequency-independent permittivity, permeability and conductivity for the dielectric material.

The total charge per unit length on the conductor C_1 can be determined, assuming perfect conductor, as

$$Q = \oint_{C_1} \rho_s d\ell = \varepsilon \oint_{C_1} \vec{E} \cdot \hat{n} d\ell \tag{4.77}$$

where $d\ell$ is the differential length and the charge density per unit length $\rho_s = D_n = \varepsilon \vec{E} \cdot \hat{n}$ from the boundary condition, with D_n representing the normal component of the electric flux density, $\vec{E}$ being the electric field, $\hat{n}$ standing for the unit vector perpendicular to the conductor surface, and ε being the permittivity of the medium surrounding the conductors. Note that the integral is carried along a contour enclosing the conductor C_1. The voltage between the two conductors is obtained by taking the integral of the electric field along a path between any points on the surfaces of the conductors C_1 and C_2 as

$$V_o = \int_{C_1}^{C_2} \vec{E} \cdot d\vec{\ell} \tag{4.78}$$

where $d\ell$ is the differential-length vector. The capacitance per unit length can be obtained from Eqs. (4.75), (4.77), and (4.78) as

$$C = \frac{Q}{V_o} = \frac{\varepsilon \oint_{C_1} \vec{E} \cdot \hat{n} d\ell}{\int_{C_1}^{C_2} \vec{E} \cdot d\vec{\ell}} \tag{4.79}$$

[4]Static capacitance and inductance can be used for transmission lines even though the operating frequency is different from DC. This is due to the same boundary conditions existing at the perfect conductors for both static and time-harmonic excitations.

Applying Ampere's law, the total current flowing on the conductor C_1 is obtained by taking the line integral of the tangential magnetic field H_t around a contour enclosing the conductor as

$$I_o = \oint_{C_1} H_t d\ell \tag{4.80}$$

At the perfectly conducting surface C_1, the electric and magnetic fields have only normal and tangential components, respectively, and, for TEM propagating wave, they are related by

$$H_t = \frac{E_n}{\eta} = \frac{\vec{E} \cdot \hat{n}}{\eta} \tag{4.81}$$

where $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the dielectric medium of the transmission line. Substituting H_t from (4.81) into (4.80) and using (4.77), we obtain

$$I_o = \frac{1}{\eta} \oint_{C_1} \vec{E} \cdot \hat{n} d\ell = \frac{Q}{\sqrt{\mu\varepsilon}} \tag{4.82}$$

The characteristic impedance of the transmission line can be derived, making use of (4.79) and (4.82), as

$$Z_o = \frac{V_o}{I_o} = \frac{\sqrt{\mu\varepsilon}}{C} \tag{4.83}$$

from which,

$$C = \frac{\sqrt{\varepsilon\mu}}{Z_o} \tag{4.84}$$

The total magnetic flux linkage per unit length is obtained as an integration of the magnetic flux density over a surface area extended from the conductor C_1 to conductor C_2 that has a longitudinal length of 1 m between the two conductors, as shown in Figure 4.4,

$$\begin{aligned} \psi &= \int_S \vec{B} \cdot d\vec{S} = \int_{P_1}^{P_2} \mu H d\ell \\ &= \frac{\mu}{\eta} \int_{P_1}^{P_2} -\vec{E} \cdot d\vec{l} = \frac{\mu}{\eta} V_o \end{aligned} \tag{4.85}$$

where P_1 and P_2 are arbitrary points on the surfaces of the conductors C_1 and C_2, respectively, and P_1P_2 is perpendicular to the flux lines. The external inductance per unit length is, from (4.74) and (4.85),

$$L_e = \frac{\psi}{I_o} = \frac{\mu}{\eta} \frac{V_o}{I_o} = \sqrt{\varepsilon\mu} \frac{V_o}{I_o} \tag{4.86}$$

or, upon using (4.83),

$$L_e = Z_o \sqrt{\varepsilon\mu} \tag{4.87}$$

From Eqs. (4.84) and (4.87) we can also obtain

$$Z_o = \sqrt{\frac{L_e}{C}} \tag{4.88}$$

which is the same as that obtained in (4.60), with $L_e = L$ for lossless transmission lines. Equations (4.84) and (4.87) also lead to

$$v = \frac{1}{\sqrt{L_e C}} = \frac{1}{\sqrt{\mu\varepsilon}} \tag{4.89}$$

as given in Eqs. (4.22) and (4.23). It is noted again that the velocity for a lossless transmission line immersed in a medium characterized by ε and μ is exactly the same as that for a uniform plane wave propagating in the same medium due to their same TEM mode.

The conductance per unit length given by (4.76) can be rewritten as

$$G = \frac{I_G}{I_D}\frac{I_D}{V_o} \tag{4.90}$$

where I_D is the (total) displacement current in the transmission line and is given as

$$\begin{aligned} I_D &= \int_S \vec{D} \cdot d\vec{S} \\ &= j\omega\varepsilon' \int_S \vec{E} \cdot d\vec{S} \end{aligned} \tag{4.91}$$

where S represents a surface enclosing one conductor (e.g., C_1) and ε' is the real part of the dielectric's complex permittivity $\hat{\varepsilon} = \varepsilon' - j\varepsilon''$. Note that $\varepsilon' = \varepsilon$ where ε is the permittivity of the medium used earlier. This current flows through the shunt capacitance per unit length C and is, thus, also given as

$$I_D = j\omega C V_o \tag{4.92}$$

The total shunt conduction current per unit length can be obtained as

$$\begin{aligned} I_G &= \sigma \int_S \vec{E} \cdot d\vec{S} \\ &= \omega\varepsilon'' \int_S \vec{E} \cdot d\vec{S} \end{aligned} \tag{4.93}$$

upon replacing the dielectric's conductivity σ with $\omega\varepsilon''$, where ε'' is the imaginary part of the dielectric's complex permittivity, given in Eq. (2.37) of Chapter 2. The conductance per unit length is then obtained from (4.90)–(4.93) as

$$G = \frac{\omega\varepsilon''}{\varepsilon'} C \tag{4.94}$$

Equations (4.94) and (2.38) show that

$$\frac{G}{C} = \frac{\omega\varepsilon''}{\varepsilon'} = \frac{\sigma}{\varepsilon'} = \omega \tan\delta \tag{4.95}$$

upon using the loss tangent of the dielectric $\tan\delta = \sigma/\omega\varepsilon'$, which is a useful relationship between the shunt conductance G and capacitance C per unit length of a transmission line, allowing one parameter to be determined from the other for a given transmission-line medium.

The resistance per unit length R, as we recall, is used to represent the power loss due to imperfect conductors and can be determined from the following known relation:

$$P_L = \frac{1}{2} R I_o^2 \tag{4.96}$$

where P_L is the total power loss per unit length along the conductors. P_L is the sum of the individual power losses on the two conductors and is given as

$$\begin{aligned} P_L &= \frac{R_s}{2}\oint_{C_1+C_2} \vec{J}_S \cdot \vec{J}_S^{\,*} dl \\ &= \frac{R_s}{2}\oint_{C_1+C_2} (\hat{n}\times\vec{H})\cdot(\hat{n}\times\vec{H}^{\,*}) dl \\ &= \frac{R_s}{2}\oint_{C_1+C_2} |\vec{H}|^2 dl \end{aligned} \tag{4.97}$$

upon using

$$\begin{aligned} (\hat{n}\times\vec{H})\cdot(\hat{n}\times\vec{H}^{\,*}) &= \hat{n}\cdot\vec{H}\times(\hat{n}\times\vec{H}^{\,*}) \\ &= \hat{n}\cdot[(\vec{H}\cdot\vec{H}^{\,*})\hat{n} - (\vec{H}\cdot\hat{n})\vec{H}^{\,*}] \\ &= |\vec{H}|^2 \end{aligned}$$

which makes use of the fact that $\hat{n}\cdot\vec{H} = 0$ at a perfect conductor. $\vec{J}_S$ is the surface current density along the z axis; $R_s = 1/\sigma_c\delta_s$ is the (frequency-dependent) surface resistance of the conductor with the skin depth $\delta_s = \sqrt{2/\omega\mu_c\sigma_c}$ discussed in Section 2.8 of Chapter 2, σ_c being the conductor's skin depth and conductivity, respectively, and μ_c being the permeability of the conductor (typically equal to μ_o). The magnetic field intensity $\vec{H}$ is taken as that for perfect conductors. Note that the integration is carried around both conductors to obtain the total loss. The resistance per unit length is obtained from (4.80) and (4.96), (4.97) as

$$R = R_s \frac{\oint_{C_1+C_2} |\vec{H}|^2 dl}{\left(\oint_{C_1} \left|\vec{H}\right| dl\right)^2} \tag{4.98}$$

after replacing H_t with $|\vec{H}|$, which is valid since the magnetic field has only tangential component at the surface of a perfect conductor.

Equation (4.98) implies that R depends on the frequency as expected due to the frequency-dependent skin effect of the conductors. When the conductors are imperfect, there will be a penetration of the magnetic field into the conductors, causing currents to flow internally within the conductors. This results in an additional inductance for the transmission line and hence increasing the inductance per unit length of the line. This extra inductance (L_i) is called "internal inductance" or "skin-effect inductance." The magnetic energy stored in this inductance is given by

$$W_{L_i} = \frac{1}{4} L_i I_o^2 \tag{4.99}$$

As the conductors have a finite conductivity, the transmission line exhibits a surface (or skin-effect) impedance on the conductors as given in Eq. (2.55):

$$Z_S = R_s + j\omega L_s = \frac{1}{\sigma_c\delta_s}(1+j) \tag{4.100}$$

where R_s and L_s are the surface resistance (or resistivity) and inductance of the conductor, respectively. The magnetic energy stored in the surface inductance L_s is

$$W_{L_s} = \frac{1}{4} L_s \oint_{C_1+C_2} |\vec{J}_S|^2 dl$$
$$= \frac{1}{4} L_s \oint_{C_1+C_2} |\vec{H}|^2 dl \tag{4.101}$$

making use of $\vec{J}_S = \hat{n} \times \vec{H}$ on a perfect conductor. Replacing the integral with that from Eq. (4.97), and using (4.96) and the fact that $R_s = \omega L_s$ from (4.100), we get

$$W_{L_s} = \frac{RI_o^2}{4\omega} \tag{4.102}$$

which must be equal to the magnetic energy stored in the internal inductance. This leads to

$$\omega L_i = R \tag{4.103}$$

The total series inductance per unit length can now be obtained as

$$L = L_e + \frac{R}{\omega} \tag{4.104}$$

and is frequency-dependent. For low loss transmission lines, $R \ll \omega L_e$, and hence this additional inductance due to the conductor's finite conductivity is very small and normally neglected. Note that, from Eqs. (4.98) and (4.103), R increases as $\sqrt{\omega}$, whereas L_i decreases as $\sqrt{\omega}$. However, the reactance ωL_i is proportional to $\sqrt{\omega}$ and hence cannot be neglected at high frequencies.

As have been seen, the per-unit-length resistance R and internal inductance L_i can be evaluated from Eqs. (4.98) and (4.103), respectively. In practice, transmission lines with good conductors are used. For these transmission lines, R and L_i are typically small. Therefore, the transmission line's parameters like characteristic impedance can be accurately determined even approximations are used for R and L_i. For good conductors, we can derive approximate closed-form equations for R and L_i as follows.

We consider a plane conductor shown in Figure 4.5. Its resistance is given, assuming the current is uniformly distributed over the cross section of the conductor, as

$$R_{\text{dc}} = \frac{\ell}{\sigma_c A} = \frac{\ell}{\sigma_c h W} \tag{4.105}$$

where A is the cross-sectional area of the conductor and σ_c is the conductivity, whose value remains almost the same from DC to infrared frequencies. At DC, the skin depth is infinity and the current would reside completely within the conductor, thus distributing uniformly over the entire cross section. Equation (4.105)

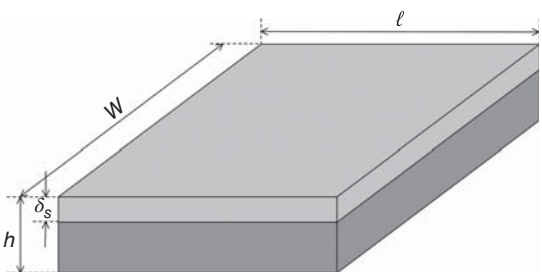


Figure 4.5. A plane conductor.

is therefore strictly valid at DC and approximate at low frequencies where the skin depth is large. At high frequencies, the skin depth is small and the current approximately resides within a few skin depths from the surface, causing nonuniform current distribution over the conductor's entire cross section. However, within one skin depth δ_s of the surface, the current does not vary much and we can approximate the current distribution as uniform. As a result, the resistance[5] per unit length of the conductor at high frequency, corresponding to a cross-sectional area of $\delta_s W$, can be approximated as

$$R_f = \frac{1}{\sigma_c \delta_s W} \tag{4.106}$$

It is noted herein that the current does not vanish at a depth equal to the skin depth; instead, it takes several skin depths to reduce to a negligible value (e.g., $4.6\delta_s$ to reach 1% of its initial value) which should be used in determining R_f (for conductors or very lossy substrates.) These distances (e.g., $4.6\delta_s$), however, are not considered here since the resistance is not uniformly distributed over several skin depths. The use of one skin depth results in possible inaccuracy for R_f.

For good conductors, $\delta_s = 1/\alpha \simeq \sqrt{2/\omega\sigma_c\mu_o}$, where α is the attenuation constant, and so

$$R_f = \frac{1}{W}\sqrt{\frac{\omega\mu_o}{2\sigma_c}} \tag{4.107}$$

for good conductors. Note the relative permeability μ_r for (nonmagnetic) conductors is approximately equal to one. We can then see that, for a given conductor with conductivity σ_c, as the frequency is increased, the skin depth reduces leading to increased resistance and attenuation constant and hence loss. This phenomenon is indeed expected for increasing frequency. R_f can be rewritten using the surface resistance R_s and δ_s as

$$R_f = \frac{R_s}{W} \tag{4.108}$$

which is expected since the conductor width can be considered consisting of W unit widths connected in parallel. The transmission line's per-unit-length resistance can now be estimated as

$$R = \frac{R_{s1}}{W_1} + \frac{R_{s2}}{W_2} \tag{4.109}$$

where W_1, W_2 and R_{s1}, R_{s2} are the widths and surface resistances of the transmission line's two conductors, respectively, assuming plane conductors.

The internal inductance per unit length L_i of the transmission line is obtained from (4.103) and (4.109) as

$$L_i = \frac{1}{\omega}\left(\frac{R_{s1}}{W_1} + \frac{R_{s2}}{W_2}\right) \tag{4.110}$$

As an example, we now consider a coaxial transmission line whose cross section is shown in Figure 4.6(a). Consider the unit-length inner conductor shown in Figure 4.6(b), and assume the current is concentrated within a skin depth of the surface and the frequency is sufficiently high so that the skin depth is relatively small as compared to the conductor's radius, we can neglect the conductor's curvature and approximate the inner conductor per unit length as a plane conductor having width equal to the circumference $2\pi a$ and thickness equal to the skin depth δ_s, as shown in Figure 4.6(c). Similarly, the outer conductor of radius b per unit length can also be approximately represented by a plane conductor of width $2\pi b$ and thickness δ_s The

[5]The resistance per square of a plane conductor (of any size) of thickness δ_s is obtained when $\ell = W$ as $R_f(\Omega/\text{sq}) = 1/\sigma_c\delta_s$.

per-unit-length resistance R and internal inductance $L_{\rm i}$ can be obtained from (4.109) and (4.110), respectively, as

$$R = \frac{R_s}{2\pi a} + \frac{R_s}{2\pi b} = \frac{(a+b)R_s}{2\pi ab} \tag{4.111}$$

$$L_i = \frac{(a+b)R_s}{2\pi\omega ab} \tag{4.112}$$

assuming the same material for the inner and outer conductors. As will be seen later, these results are exactly the same as those in (4.122) and (4.123) obtained from a more rigorous analysis. It should be noted that for some planar transmission lines, the width of the second conductor (normally, ground plane), along which the current concentrates, may be much wider than that of the first conductor; in that case, a width smaller than the actual width should be used in the calculations. Otherwise, the portions of R and L_i corresponding to that conductor would be zero. For instance, for microstrip lines, the width of the ground plane used in calculations can be approximated as about five times of the width of the strip. Figure 4.7 shows sketches of the current distributions within the conductors of coaxial and microstrip lines.

4.4.2 Formulation for Simple Transmission Lines

For transmission lines with simple geometry, such as coaxial line, closed-form expressions for R, L, G, and C can be easily derived. For most practical printed-circuit transmission lines, however, these expressions are very difficult – if not impossible – to derive and numerical methods, for example, spectral-domain method, are normally employed to obtain numerical results for the parameters.

As an illustration of the procedure to determine the transmission line's per-unit-length parameters for simple transmission lines, we consider a (lossy) coaxial transmission line shown in Figure 4.6(a). The relative dielectric constant of the medium, assumed to be nonmagnetic, between the two conductors is characterized by $\widehat{\varepsilon} = \varepsilon' - \varepsilon''$ and the conductors' conductivity is σ_c. We first determine the electric and magnetic fields for the coaxial line. To facilitate this formulation, we assume that the conductors are perfect ($\sigma \to \infty$) so that the transmission line supports TEM mode[6] regardless whether the medium is lossless or lossy. For practical transmission lines including coaxial line, the conductors are typically good and, consequently, the fields can be assumed to be almost the same as that of the TEM mode with perfect conductors. Therefore, the fields for a lossless transmission line are normally employed to evaluate the parameters of these (low loss) practical transmission lines.[7] For TEM mode, the fields lie on a transverse plane and, consequently, the electric field

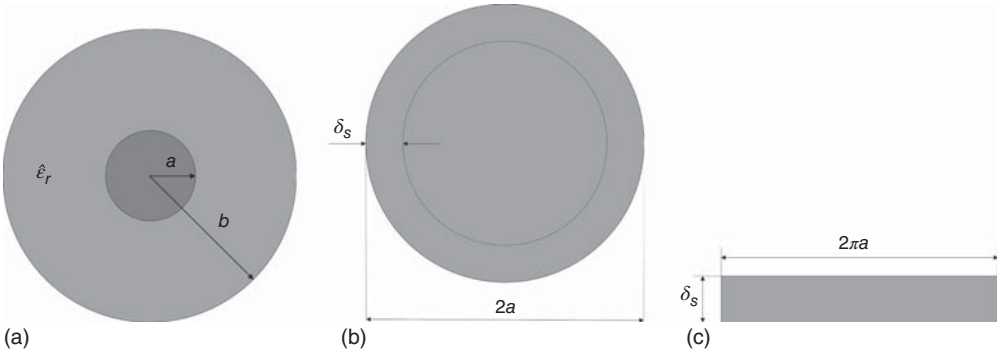


Figure 4.6. Cross section of a coaxial line (a) and the unit-length inner conductor (b) with its approximate plane conductor (c).

[6]TEM mode exists only on transmission lines with perfect conductors immersed in a single dielectric (i.e., the medium surrounding the conductors is homogeneous.)

[7]This technique, in which the fields for lossless lines are used for low-loss lines, is known as *perturbation method*.

for the coaxial line has only radial component. Applying Gauss' law over a cylindrical surface enclosing the inner conductor, we obtain the total charge per unit length on the inner conductor as

$$\begin{aligned} Q &= \oint_S \vec{D} \cdot d\vec{S} = \varepsilon \int_0^{2\pi} E_r \hat{a}_r \cdot \hat{a}_r r d\varphi \\ &= 2\pi\varepsilon r E_r \end{aligned} \tag{4.113}$$

The voltage between the two conductors is

$$\begin{aligned} V_o &= -\int_b^a \vec{E} \cdot d\vec{l} = \int_a^b E_r dr \\ &= \frac{Q}{2\pi\varepsilon} \int_a^b \frac{dr}{r} \\ &= \frac{Q}{2\pi\varepsilon} \ln \frac{b}{a} \end{aligned} \tag{4.114}$$

upon substituting E_r from (4.113) and taking the integral. The electric field is then obtained from (4.113) and (4.114) as

$$E = \frac{V_o}{r \ln(b/a)} a_r \tag{4.115}$$

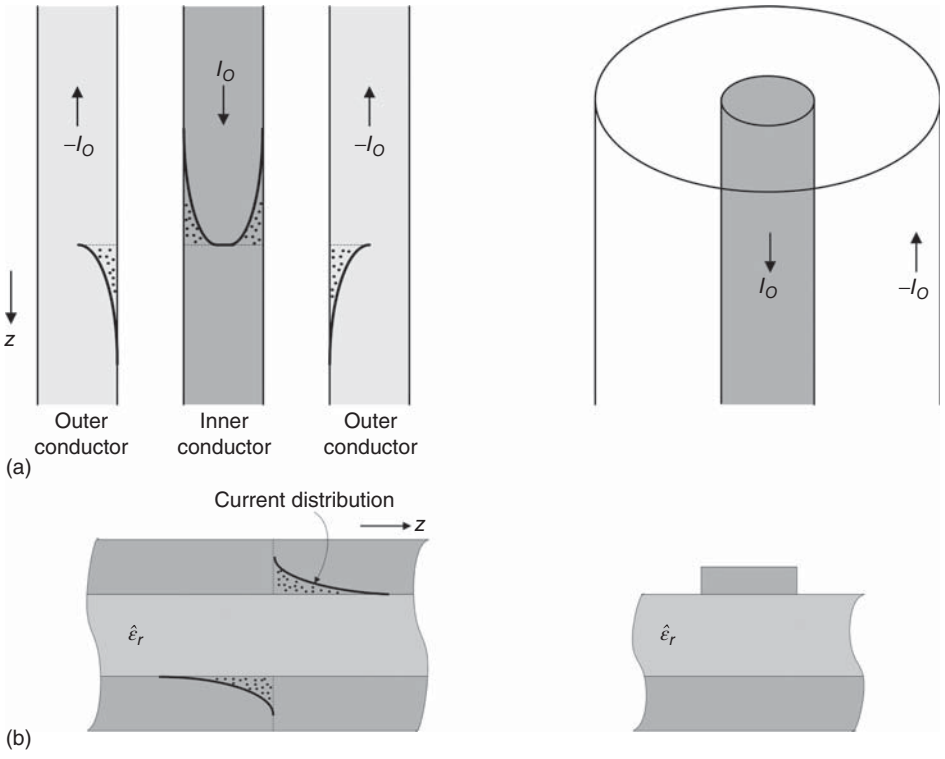


Figure 4.7. Sketch of the current distribution in coaxial line (a) and microstrip line (b).

The magnetic field is given by

$$\begin{aligned}\vec{H} &= \frac{\hat{a}_z \times \vec{E}}{\eta} \\ &= \frac{V_o}{r\eta \ln(b/a)}\hat{a}_\varphi \end{aligned} \tag{4.116}$$

The total current is obtained from (4.80) as

$$\begin{aligned} I_o &= \oint_{C_1} \vec{H} \cdot d\hat{l} = \int_0^{2\pi} \vec{H} \cdot \hat{a}_\varphi a d\varphi \\ &= \frac{2\pi V_o}{\eta \ln(b/a)} \end{aligned} \tag{4.117}$$

Note that the contour C_1 corresponds to $r = a$. The external series inductance per unit length L_e can now be obtained by substituting ψ from Eq. (4.85) and I_o from (4.117) into (4.74) as

$$\begin{aligned} L_e &= \frac{\psi}{I_o} \\ &= \frac{\mu}{2\pi} \ln\left(\frac{b}{a}\right) \end{aligned} \tag{4.118}$$

The shunt capacitance per unit length C is given from (4.75) as

$$\begin{aligned} C &= \frac{Q}{V_o} \\ &= \frac{2\pi\varepsilon}{\ln(b/a)} \end{aligned} \tag{4.119}$$

upon using (4.114).

The shunt conductance per unit length G is obtained by substituting C from (4.119) into (4.94) as

$$G = \frac{2\pi\omega\varepsilon''}{\ln(b/a)} \tag{4.120}$$

The power loss per unit length on the conductors is obtained from (4.97), making use of (4.116), as

$$\begin{aligned} P_L &= \frac{R_s}{2} \oint_{C_1+C_2} |\vec{H}|^2 dl \\ &= \frac{R_s}{2} \left\{ \frac{V_o^2}{a^2\eta^2 \left[\ln(b/a)\right]^2} \int_0^{2\pi} a d\varphi + \frac{V_o^2}{b^2\eta^2[\ln(b/a)]^2} \int_0^{2\pi} b d\varphi \right\} \\ &= \frac{R_s}{2} \frac{2\pi V_o^2}{\eta^2[\ln(b/a)]^2} \left(\frac{1}{a} + \frac{1}{b}\right) \\ &= \frac{a+b}{ab} \frac{\pi R_S V_o^2}{[\eta \ln(b/a)]^2} \end{aligned} \tag{4.121}$$

Substituting P_L in (4.121) into (4.96) and solving for R, we get, upon using (4.117),

$$R = \frac{(a+b)R_s}{2\pi ab} \tag{4.122}$$

The internal inductance per unit length is obtained from (4.103) as

$$L_i = \frac{(a+b)R_s}{2\pi\omega ab} \tag{4.123}$$

Note that R and L_i in Eqs. (4.122) and (4.123), respectively, are the same as those derived in (4.111) and (4.112). The total series inductance per unit length can be obtained from (4.118) and (4.123) as

$$L = \frac{\mu}{2\pi}\ln\frac{b}{a} + \frac{(a+b)R_s}{2\pi\omega ab} \tag{4.124}$$

The foregoing analysis can be employed to determine the parameters per unit length for other transmission lines with simple geometry.

4.5 DIELECTRIC AND CONDUCTOR LOSSES IN TRANSMISSION LINES

We have seen in Sections 4.2 and 4.3 that the loss in transmission lines can be determined exactly from Eq. (4.35) or (4.57) for (lossy) transmission lines, and approximately from (4.68) for low loss lines. This loss is attributed by three different kinds: dielectric, conduction, and radiation loss. It is a common practice to compute these three constituent losses, and the results enable us to get insight to the loss phenomena and helps identify the magnitude of each loss contribution. This information is useful for circuit design – for instance, it can help choose a proper transmission line or possibly optimize a transmission-line structure including its substrates and dimensions to reduce a certain loss.

Radiation loss of uniform transmission lines used in practice is normally very small and usually neglected, especially at low RF regions. As a result, the attenuation constant of a transmission line is commonly described as

$$\alpha = \alpha_d + \alpha_c \tag{4.125}$$

where α_d and α_c represent the dielectric and conductor attenuation constant, respectively, and depend on frequency. Most transmission lines have $\alpha_c \gg \alpha_d$, except for highly lossy dielectrics, such as silicon, in which the dielectric loss is the dominant contribution. For homogeneous transmission lines (e.g., strip line), α_d is independent of the geometry of the line, whereas for inhomogeneous transmission lines (e.g., microstrip line), α_d is in general a function of the line's geometry. On the other hand, α_c is always a function of the geometry. We can write, from Eq. (4.54),

$$\alpha = \alpha_d + \alpha_c = \frac{P_{Ld} + P_{Lc}}{2P_T} \tag{4.126}$$

where

$$\alpha_d = \frac{P_{Ld}}{2P_T} \tag{4.127}$$

$$\alpha_c = \frac{P_{Lc}}{2P_T} \tag{4.128}$$

with P_{Ld} and P_{Lc} representing the power losses per unit length due to imperfect dielectric and conductors, respectively. Practical transmission lines normally have small losses and so their electric and magnetic fields are only slightly changed or perturbed from the lossless case. Therefore, we can employ a perturbation method, in which the fields for the lossless transmission line are used to approximate those of the corresponding lossy transmission line, to determine α_d and α_c. The perturbation method is convenient for analyzing low loss transmission lines since the actual fields of these lines are not needed.

4.5.1 Dielectric Attenuation Constant

From the Poynting theorem, we can write the power loss per unit length of a transmission line due to imperfect dielectric as

$$P_{Ld}(z) = \frac{\sigma}{2}\int_S \vec{E}\cdot\vec{E}^{\,*}\,dS = \frac{\sigma}{2}\int_S |\vec{E}|^2 dS = \frac{\omega\varepsilon''}{2}\int_S |\vec{E}|^2 dS \tag{4.129}$$

where $\sigma = \omega\varepsilon''$ is the conductivity of the dielectric and $\vec{E}$ is the electric field intensity. Note that, because of the dielectric's finite conductivity, there is current $\vec{J} = \sigma\vec{E}$ exist between the two conductors. The average power flow along the line is given as

$$P_T = \frac{1}{2}\mathrm{Re}\int_S \vec{E}\times\vec{H}^{\,*}\cdot d\vec{S} \tag{4.130}$$

Replacing $\vec{H}$ with

$$\vec{H} = \frac{\hat{a}_z\times\vec{E}}{\eta} \tag{4.131}$$

where $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the dielectric medium, and applying the vector identity for three arbitrary vectors $\vec{A}$, $\vec{B}$, and $\vec{C}$

$$\vec{A}\times\vec{B}\times\vec{C} = (\vec{A}\cdot\vec{C})\vec{B} - \vec{C}(\vec{A}\cdot\vec{B}) \tag{4.132}$$

we obtain

$$P_T(z) = \frac{1}{2\eta}\int_S \vec{E}\cdot\vec{E}^{\,*}\,dS = \frac{1}{2\eta}\int_S |\vec{E}|^2 dS \tag{4.133}$$

We can also derive

$$P_T(z) = \frac{\eta}{2}\int_S \vec{H}\cdot\vec{H}^{\,*}\,dS = \frac{\eta}{2}\int_S |\vec{H}|^2 dS \tag{4.134}$$

Here, we assume η is real, which is a good approximation for low loss materials. The dielectric attenuation constant can be derived from Eqs. (4.127), (4.129), and (4.133) as

$$\alpha_d = \frac{P_{Ld}}{2P_T} = \frac{\sigma\eta}{2} = \frac{\omega\varepsilon''\eta}{2} \tag{4.135}$$

For a low loss dielectric with (complex) relative dielectric constant $\hat{\varepsilon}_r = \varepsilon_r' - j\varepsilon''$, $\varepsilon_r'' \ll \varepsilon_r'$ and its intrinsic impedance

$$\eta = \sqrt{\frac{\mu}{\hat{\varepsilon}}} = \frac{\eta_o}{\sqrt{\varepsilon_r' - j\varepsilon_r''}} \simeq \frac{\eta_o}{\sqrt{\varepsilon_r'}} \tag{4.136}$$

where $\eta_o = 120\pi\ \Omega$ is the intrinsic impedance of air. Note that $\varepsilon_r' = \varepsilon_r$ of the dielectric. The dielectric attenuation constant for a low loss TEM transmission line can now be approximated as

$$\alpha_d \cong \frac{k_o}{2}\frac{\varepsilon_r''}{\sqrt{\varepsilon_r}} \tag{4.137}$$

where $k_o = \omega\sqrt{\varepsilon_o\mu_o}$ is the wave number in air and ε_r is used in place of ε_r'.

4.5.2 Conductor Attenuation Constant

It is well known that as a signal propagates along a perfect conductor; its energy cannot penetrate into the conductor. However, if the conductor is nonperfect, some of the energy would enter the conductor, yet attenuated in accordance to the attenuation function $e^{-2\alpha_c r}$, where r is the distance from the conductor surface in the direction perpendicular to it. The density of power flow into a conductor is equal to the power dissipated within the conductor itself. Therefore, the power loss per unit length due to an imperfect conductor is given as

$$\begin{aligned} P_{Lc}(z) &= \frac{1}{2}\mathrm{Re}\int_{C_1+C_2}(\vec{E}\times\vec{H}^*)\cdot\hat{a}_r dS \\ &= \frac{1}{2}\mathrm{Re}\int_{C_1+C_2}(\hat{a}_r\times\vec{E})\cdot\vec{H}^* \\ &= \frac{1}{2}\mathrm{Re}\int_{C_1+C_2}Z_S|\vec{H}|^2 dS \\ &= \frac{1}{2}R_s\int_{C_1+C_2}|\vec{H}|^2 dS \\ &= \frac{1}{2}R_s\int_{C_1+C_2}|\vec{H}|^2 d\ell \end{aligned} \tag{4.138}$$

where C_1 and C_2 represent the two conductors of the transmission line; η_c is the intrinsic impedance of the conductor; Z_S is the surface impedance and R_s is its real part or the surface resistance (or resistivity) of the conductor as given in (4.100). $\vec{H}$ is the magnetic field assuming no loss; that is its magnitude is equal to the magnitude of the magnetic field at the conductor surface.[8] The conductor attenuation constant can be derived from Eqs. (4.128), (4.134), and (4.138) as

$$\begin{aligned} \alpha_c &= \frac{P_{Lc}}{2P_T} \\ &= \frac{R_s\displaystyle\int_{C_1+C_2}\vec{H}\cdot\vec{H}^* dl}{2\eta_c\displaystyle\int_S \vec{H}\cdot\vec{H}^* dS} \end{aligned} \tag{4.139}$$

It can also be derived as

$$\alpha_c = \frac{R_s\displaystyle\int_{C_1+C_2}\vec{E}\cdot\vec{E}^* dl}{2\eta\displaystyle\int_S \vec{E}\cdot\vec{E}^* dS} \tag{4.140}$$

It is particularly noted that, as a signal propagates along transmission lines, the electric and magnetic fields and the current of the signal penetrate into the surfaces of the transmission lines' conductors and reside within a few skin depths of the conductors. The thickness of the conductors in transmission lines should therefore be at least larger than the skin depth at the operating frequencies to minimize the loss due to the conductors. If the conductor thickness is comparable to the skin depth, a significant conduction loss would occur.

[8] $\vec{E}_{\text{tangential}} = Z_S\vec{J}_S$, where $\vec{J}_S$ is the surface current density. This implies that the electric field along the conductor surface $E_z \neq 0$. Therefore, the propagating wave is not exactly TEM. It is commonly referred to as quasi-TEM. Pure TEM exists only in homogeneous transmission lines with perfect conductors for either lossless or lossy dielectric. This finite longitudinal electric field results in a component of the Poynting vector normal to the conductor surface, causing energy propagating into the conductors orthogonally and results in loss in the conductors.

As an example for determining the attenuation constant of transmission lines, we consider a coaxial transmission line whose inner and outer radii are a and b, respectively, as shown in Figure 4.6(a). The dielectric is assumed to be lossy and characterized by $\hat{\varepsilon}_r = \varepsilon_r' - j\varepsilon_r''$ where $\varepsilon_r' = \varepsilon_r$. The conductors are imperfect with a finite conductivity σ_c. Using perturbation method, we assume that the fields of this transmission line are given as those of the lossless coaxial line:

$$\vec{E} = \hat{a}_r \frac{V_o}{\ln(b/a)} \frac{1}{r} e^{-j\beta z} \tag{4.141}$$

$$\vec{H} = \hat{a}_r \frac{V_o}{\eta \ln(b/a)} \frac{1}{r} e^{-j\beta z} \tag{4.142}$$

where V_o is the potential of the inner conductor, r is the radius between a and b, and $\beta = k = k_o\sqrt{\varepsilon_r}$. The potential of the outer conductor is assumed to be zero. The power flow along the coaxial line is obtained as

$$\begin{aligned} P_T &= \frac{1}{2}\mathrm{Re}\int_S \vec{E} \times \vec{H}^* \cdot d\vec{S} \\ &= \frac{1}{2}\int_0^{2\pi}\int_a^b \vec{E} \times \vec{H}^* \cdot \hat{a}_z r dr d\varphi \\ &= \frac{\pi V_o^2}{\eta \ln(b/a)} \end{aligned} \tag{4.143}$$

The power loss per unit length due to dielectric loss is

$$\begin{aligned} P_{Ld} &= \frac{\omega\varepsilon''}{2}\int_S |\vec{E}|^2 dS \\ &= \frac{\omega\varepsilon''}{2}\int_0^{2\pi}\int_a^b |\vec{E}|^2 r dr d\varphi \\ &= \frac{\pi\omega\varepsilon'' V_o^2}{\ln(b/a)} \end{aligned} \tag{4.144}$$

The dielectric attenuation constant can be derived from Eqs. (4.127), (4.143), and (4.144) as

$$\alpha_d = \frac{k_o \varepsilon_r''}{2\sqrt{\varepsilon_r}} \tag{4.145}$$

which is the same as Eq. (4.137) as expected for TEM transmission lines.

The power loss per unit length due to conductor loss is

$$\begin{aligned} P_{Lc} &= \frac{R_s}{2}\int_{C_1+C_2} |\vec{H}|^2 dl \\ &= \frac{R_s}{2}\left[\int_0^{2\pi} \left|\vec{H}(r=a)\right|^2 a d\varphi + \int_0^{2\pi} |\vec{H}(r=b)|^2 b d\varphi\right] \\ &= \frac{R_s \pi V_o^2}{\eta^2 \ln^2(b/a)}\left(\frac{1}{a} + \frac{1}{b}\right) \end{aligned} \tag{4.146}$$

The conductor attenuation constant is obtained from Eqs. (4.128), (4.143), and (4.146) as

$$\alpha_c = \frac{R_s}{2\eta \ln(b/a)}\left(\frac{1}{a} + \frac{1}{b}\right) \tag{4.147}$$

which depends on the transmission line's geometry.

The (total) attenuation constant can now be determined using Eqs. (4.125), (4.145), and (4.147) as

$$\alpha = \frac{k_o \varepsilon_r''}{2\sqrt{\varepsilon_r}} + \frac{R_s}{2\eta \ln(b/a)} \left(\frac{1}{a} + \frac{1}{b} \right) \tag{4.148}$$

4.6 DISPERSION AND DISTORTION IN TRANSMISSION LINES

4.6.1 Dispersion

When the velocity of wave or signal (normally TEM or quasi-TEM) propagating along a transmission line is dependent upon frequency, the transmission line is said to have dispersion. There is no dispersion in an (ideal) lossless transmission line because the velocity, as given by Eq. (4.62), is independent of frequency. Any practical transmission line, however, is lossy and its phase constant, in general, does not vary linearly with frequency. This results in a velocity depending on frequency, thus causing dispersion. As an example, the phase constant for low loss transmission lines, as derived in Eq. (4.69), is in general a nonlinear function of frequency and the resultant velocity, from Eq. (4.58), is hence dependent upon frequency. For good transmission lines, the rate of change of the velocity versus frequency is, in general, relatively small until the frequency reaches an extremely high value. Also, the velocity changes very little at low frequencies. Consequently, narrow-band signals and signals whose spectra consist of very low frequencies suffer relatively small dispersion within the operating bandwidth.

Dielectric substrates used in practical transmission lines usually have relative dielectric constants almost constant versus frequency, particularly in the low RF region. Homogeneous transmission line, such as strip line, has velocity depending exclusively on the relative dielectric constant, as given by $v = c/\sqrt{\varepsilon_r}$, hence possessing very little dispersion. Most transmission lines used in practice, however, are inhomogeneous, having more than one dielectric substrate – for example, microstrip line. An inhomogeneous transmission line has velocity depending on the effective relative dielectric constant, as shown in Eq. (4.246). Assuming the relative dielectric constant is independent of frequency, calculations show that in general the effective relative dielectric constant increases with frequency, and this change can be substantial, especially for high dielectric-constant substrates. Consequently, the dispersion for inhomogeneous transmission lines can be large. Due to the dispersion, signals of different frequencies travel at different velocities. Since the lengths of transmission lines used in RFICs are typically viewed in terms of wavelength which varies as $\lambda = v/f$ (even they are given as physical dimension), which are affected by the transmission line's dispersion, the dispersion in transmission lines needs to be considered in the design and analysis of wideband RFICs. Another effect of the dispersion is the distortion it imposes on a propagating signal whose spectrum contains multiple frequencies occurring simultaneously, which is critical for high frequency and high speed circuits.

Figure 4.8 shows the frequency-dependent effective relative dielectric constant for microstrip lines with SiO_2 and silicon as the substrate that can be used in RFICs. In practice, the relative dielectric constant slightly decreases as the frequency is increased. However, this rate of reduction is relatively small as compared to the rate of increase of the effective relative dielectric constant versus frequency, making it a secondary effect as compared to the effective relative dielectric constant. Some transmission lines have effective relative dielectric constant staying almost constant from DC up to a certain cutoff frequency and then increasing as the operating frequency exceeding that cutoff frequency. These transmission lines can hence have very small dispersion up to the cutoff frequency.

4.6.2 Distortion

There are various kinds of waveforms that may be used in RFICs. However, for the discussion purposes, we consider here three kinds of waveforms: continuous sinusoidal waveform, periodic waveform, and non-periodic waveform. The sinusoid contains only single frequency and hence retains its original wave shape through linear networks; therefore, signal distortion is not much of an issue except a reduction in amplitude.

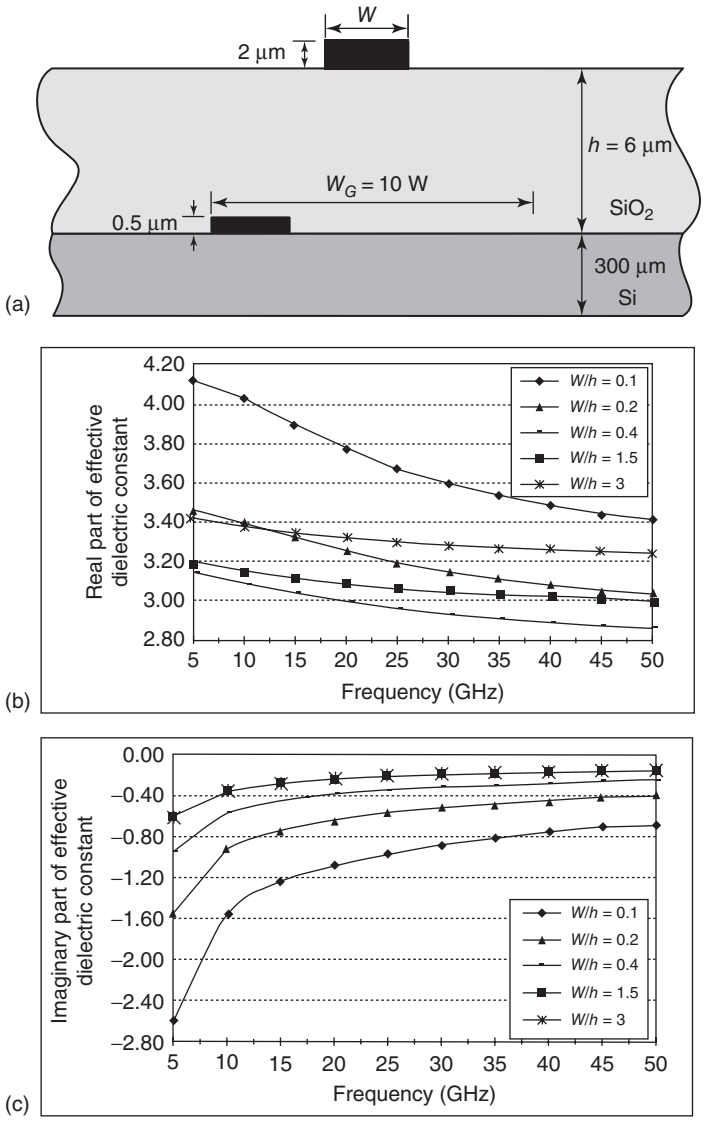


Figure 4.8. Microstrip line (a) and its real (b) and imaginary (c) part of effective relative dielectric constant versus frequency for SiO_2, with loss tangent $\tan\delta = 0.0002$ and $\varepsilon_r = 4.1$, and silicon with resistivity $\rho = 5\ \Omega$-cm and $\varepsilon_r = 12.5$. The conductor is copper having conductivity $\sigma = 5.8 \times 10^7$ mho/m.

It is noted that, even for wide-band circuits, each signal still only has one frequency. The nonperiodic and periodic signals, such as a single sinusoidal pulse or train of pulses, respectively, however, contain multiple constituent signals operating at different frequencies simultaneously. As these composite signals propagate along a transmission line, their amplitude and phase change versus frequency due to the frequency-dependent loss and dispersion of the transmission line, respectively, causing distortion in the signal waveform. This distortion can be significant for some transmission lines and signals with wide bandwidths, especially at very high frequencies, as it can alter the signal substantially as the signal propagates along a transmission line. Distortion, therefore, needs to be considered in transmission-line design as well as in circuit analysis and design. An example is pulse signals propagating in ultra-wideband (UWB)[9] RF circuits or digital circuits, where distortion can be significant within the operating frequency range.

[9]UWB frequencies are from 3.1 to 10.6 GHz and UWB circuits for both communications and radar can be based on pulse signals operating within this frequency range.

Periodic signals such as a train of pulses can be expressed using Fourier series as

$$V(t) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)]$$
$$= \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_0 t} \tag{4.149}$$

where ω_0 is the fundamental frequency, and a_0, a_n, b_n, and c_n are the Fourier coefficients. The signal, as can be seen, consists of multiple sinusoidal signals at discrete frequencies $\omega = n\omega_0$. Nonperiodic signals such as a single pulse may also be represented using the exponential Fourier series given in (4.149), in which the period is extended to infinity, or using Fourier integral as

$$V(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} V(\omega) e^{j\omega t} d\omega \tag{4.150}$$

where $V(\omega)$ is the Fourier transform of $V(t)$, which shows that a nonperiodic signal also contains multiple sinusoidal signals of different frequencies in a continuous spectrum.

We now consider an infinitely long transmission line and assume a signal propagation along the line has a spectrum containing multiple frequencies $f_1, f_2, \ldots, f_N$; that is, the signal contains multiple signals operating at these frequencies concurrently. Since the propagation constant of the transmission line are function of frequency, the amplitude and phase constants of these constituent signal are different. The (phasor) voltage of the composite signal at any location on the line at any given time can be determined as the summation of its constituent signals, using Eq. (4.49), as

$$V(z) = \sum_{n=1}^{N} V_n(z) = \sum_{n=1}^{N} V_n(0) e^{-\alpha_n z} e^{-j\beta_n z} \tag{4.151}$$

where $V_n(z)$ and $V_n(0)$ are the voltages of the constituent signal at frequency f_n at z and $z = 0$, respectively; and α_n and β_n are the respective attenuation and phase constant at frequency f_n. Since the constituent signals arrive at z with different amplitudes and phases, they produce a distorted signal when added. The phase distortion is caused by the dispersion of the transmission line, which typically causes the waveform to spread out and change in shape, while the amplitude distortion is due to the transmission line's loss which results in amplitude reduction. Additional distortion is caused by the reflection and cross-coupling occurred in the transmission line. Figure 4.9 illustrates the distortion of a signal propagating in a (practical) lossy and dispersive transmission line.

4.6.3 Distortion-Less Transmission Lines

The idea of distortion-less was first proposed by Oliver Heaviside in 1887. It is postulated that the per-unit-length parameters R, L, G, and C of a transmission line are related by the following condition:

$$\frac{R}{L} = \frac{G}{C} \tag{4.152}$$

The propagation constant of such a transmission line can be written using (4.35) and the condition in (4.152) as

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)}$$
$$= \sqrt{(R + j\omega L)\left(\frac{RC}{L} + j\omega C\right)} = \sqrt{\frac{L}{C}}(R + j\omega L) \tag{4.153}$$

from which, we obtain

$$\alpha = R\sqrt{\frac{C}{L}} \tag{4.154}$$

and

$$\beta = \omega\sqrt{LC} \tag{4.155}$$

The characteristic impedance and phase velocity are obtained from (4.47) and (4.58), making use of (4.152) and (4.155), respectively, as

$$Z_o = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \sqrt{\frac{R + j\omega L}{\frac{RC}{L} + j\omega C}} = \sqrt{\frac{L}{C}} \tag{4.156}$$

and

$$v = \frac{\omega}{\beta} = \frac{1}{\sqrt{LC}} \tag{4.157}$$

which are exactly the same as those for lossless transmission lines. Equations (4.154) and (4.157) show that the transmission line has constant loss and velocity versus frequency, implying that the waveform of signals traversing the line retains its shape (with reduced amplitude) and hence suffers no (phase) distortion. We can therefore conclude that any transmission line that satisfies the condition stated in (4.152) is distortionless. Since R and G affect the loss of transmission lines, a distortionless transmission should be formed only by adjusting L and/or C to meet the distortionless condition – for instance, placing additional series inductors periodically along the conductors of a transmission line to increase the inductance per unit length, assuming the added inductance is distributed uniformly over the segment where the inductor is added, as shown in Figure 4.10. As R, L, G, and C represent a transmission line whose length is very small when compared to the wavelength, the smaller the interval (period) that the additional inductors and/or capacitors are placed, the better representation for the transmission line and, hence, the better distortionless transmission that can be realized. Using smaller intervals also makes the assumption of uniform distribution for the added inductors and/or capacitors more valid.

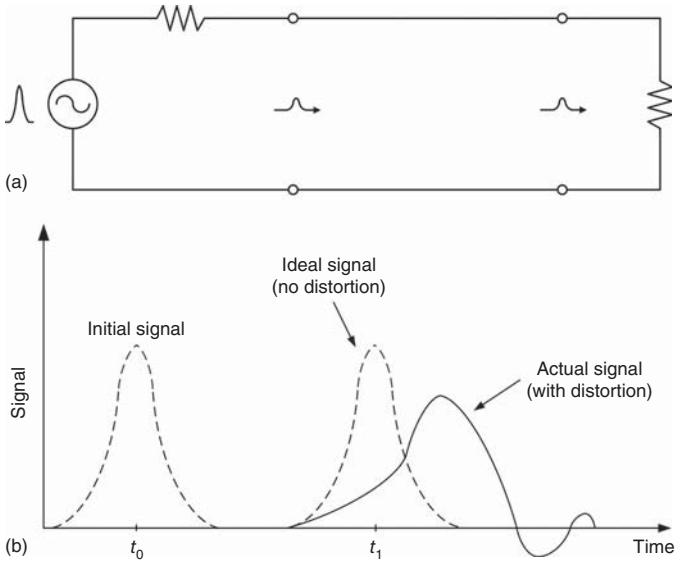


Figure 4.9. (a) Pulse signal propagating in a transmission line and (b) its distortion at t_1 with respect to the initial signal at t_0.

4.7 GROUP VELOCITY

The sinusoidal signals typically encountered in practice travel at their phase velocity at operating frequencies. This velocity is typically concerned in circuit analysis and design. However, there are another class of signals containing multiple signals operating concurrently at different frequencies (i.e., the signal contains different frequency components or spectral components). Examples are carrier-less (or video) pulses, or modulated signals such as a high frequency signal modulated by a video pulse (known as carrier-based pulse), low frequency digital or analog signal. For these signals, we can define another velocity measuring how fast the composite signal or a group of different frequency components moves within a medium or along a transmission line. This velocity is called "group velocity." To illustrate the concept of this velocity, we consider a simple case of signals consisting of two components having equal amplitude V_o at two different frequencies ω_1 and ω_2 as

$$v(t) = V_o(\cos\omega_1 t + \cos\omega_2 t) \tag{4.158}$$

Here, we assume $\omega_1 > \omega_2$. The voltage at any point z along an infinitely long lossless transmission line can be written as

$$v(t,z) = V_o[\cos(\omega_1 t - \beta_1 z) + \cos(\omega_2 t - \beta_2 z)] \tag{4.159}$$

where β_1 and β_2 are the phase constants at ω_1 and ω_2, respectively. Now letting $\omega_1 = \omega_o + \Delta\omega$ and $\omega_2 = \omega_o - \Delta\omega$, and expanding the frequency-dependent phase constant $\beta(\omega)$ in a Taylor series about ω_o as

$$\beta(\omega) = \beta_o + \left.\frac{d\beta}{d\omega}\right|_{\omega_o}(\omega - \omega_o) + \frac{1}{2}\left.\frac{d^2\beta}{d\omega^2}\right|_{\omega_o}(\omega - \omega_o)^2 + \cdots \tag{4.160}$$

where $\beta_o = \beta(\omega_o)$. Substituting β_1 and β_2 obtained from (4.160) into (4.159) and replacing ω_1 and ω_2 with $\omega_o + \Delta\omega$ and $\omega_o - \Delta\omega$, respectively, gives

$$\begin{aligned} v(z,t) = V_o \cos\left[(\omega_o t - \beta_o z) + \left((\Delta\omega)t - (\Delta\omega)\left.\frac{d\beta}{d\omega}\right|_{\omega_o} z\right) - \frac{1}{2}(\Delta\omega)^2\frac{d^2\beta}{d\omega^2} + \cdots\right] \\ V_o \cos\left[(\omega_o t - \beta_o z) - \left((\Delta\omega)t - (\Delta\omega)\left.\frac{d\beta}{d\omega}\right|_{\omega_o} z\right) - \frac{1}{2}(\Delta\omega)^2\frac{d^2\beta}{d\omega^2} + \cdots\right] \end{aligned} \tag{4.161}$$

If $\Delta\omega$ is sufficiently small, that is, narrow frequency band, we can neglect $d^2\beta/d\omega^2$ and other higher-order terms and express (4.161) as

$$v(t,z) = 2V_o \cos\left[\Delta\omega\left(t - \frac{d\beta}{d\omega}z\right)\right]\cos(\omega_o t - \beta_o t) \tag{4.162}$$

This voltage wave now consists of two components: a carrier signal at frequency ω_o moving at velocity $v = \omega_o/\beta_o$ (second component) and an envelope signal travelling at another velocity (first component). The

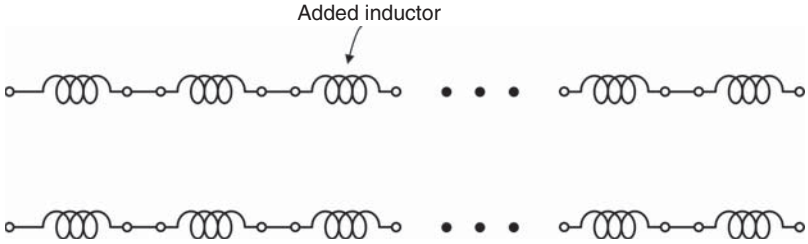


Figure 4.10. Distortionless transmission formed by adding a series of inductors periodically.

envelope has a time delay of

$$t_g = \frac{d\beta}{d\omega} z \tag{4.163}$$

at location z relative to the transmission line's input terminal. The velocity of the envelope can be derived as

$$v_g = \frac{z}{t_g} = \left(\frac{d\beta}{d\omega}\right)^{-1} \tag{4.164}$$

This velocity is the group velocity and is the actual velocity of the composite signal $v(z,t)$. Taking the derivative of $\beta = \omega/v$, where v is a function of ω, with respect to ω and substitute into (4.164), we obtain

$$\begin{aligned} v_g &= \left(\frac{v - \omega dv/d\omega}{v^2}\right)^{-1} \\ &= \frac{v}{1 - \beta\frac{dv}{d\omega}} \end{aligned} \tag{4.165}$$

which has a unique value at a frequency. This group velocity is basically the phase velocity of the envelope and is equal to the phase velocity of the carrier when there is no dispersion with respect to phase velocity (i.e., $dv/d\omega = 0$). When this happens, there is also no dispersion with respect to group velocity.

Typically, as frequency is increased, the effective relative dielectric constant of (inhomogeneous) transmission lines increases and, hence, the phase velocity reduces. Correspondingly, the group velocity, as seen from (4.165), becomes less than the phase velocity. Under this situation, the dispersion and distortion is called normal dispersion and distortion, respectively. In some cases, the phase velocity increases as frequency is increased, leading to the group velocity being greater than the phase velocity. This phenomenon is referred to as anomalous dispersion and distortion.

It is noted that the group velocity as given in (4.165) is valid only for narrow-band signals. If the separation between frequencies ω_1 and ω_2 is large (i.e., the bandwidth is wide), $d^2\beta/d\omega^2$ and other higher-order terms in (4.161) may not be neglected. Under this condition, no unique group velocity can be defined and (4.165) may not be used. Different portions of the (composite) signal or group may travel at different speeds causing dispersion in both time and space and hence signal or group distortion. Specifically, $d^2\beta/d\omega^2$ and other higher-order terms cause the envelope to spread, resulting in an increase of its duration or length. This increase in the envelope length is undesirable. At certain location along the transmission, one envelope may overlap another adjacent envelope, causing additional distortion to the signal, besides the distortion causes by the group dispersion and loss, and ambiguity as well as error in signal and/or data transmission.

The concept of group velocity helps understanding the movement of a composite signal consisting of multiple frequencies. From the circuit design point of view, however, the phase velocity is used more often than the group velocity in transmission line and circuit analysis.

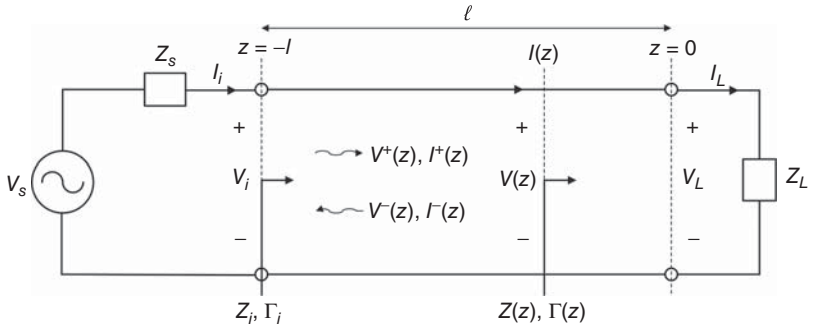


Figure 4.11. A finitely long transmission line.

4.8 IMPEDANCE, REFLECTION COEFFICIENTS, AND STANDING-WAVE RATIOS

4.8.1 Impedance

We consider a transmission line of finite length, terminated with a load $Z_L \neq Z_o$, as shown in Figure 4.11. We also assume that the transmission line is driven by a steady-state sinusoidal voltage source to facilitate the derivation of equations for the impedance, reflection coefficients and standing wave ratios (SWRs). This assumption is valid since, like the transmission-line parameters discussed previously, these parameters are also independent of the driving source's time variation. We assume that the origin of the coordinate ($z = 0$) is located at the load Z_L, and the transmission-line input is at[10] $z = -\ell$. The voltage at any location, obtained as the solution of transmission-line equation, consists of the forward and backward voltage waves as described earlier. The forward voltage wave $V^+(z) = V_o^+ e^{-\gamma z}$ is also called the incident voltage wave, whereas the backward voltage wave $V^-(z) = V_o^- e^{\gamma z}$ can be called the reflected voltage wave. This reflected voltage wave is due to the reflection of the incident voltage wave along the finitely long transmission line.

The voltage V_L and current I_L at the load ($z = 0$) can be derived from Eqs. (4.40) and (4.41) as

$$V_L = V_o^+ + V_o^- \tag{4.166}$$

and

$$I_L = I_o^+ + I_o^- \tag{4.167}$$

Equation (4.167) can be rewritten, making use of Eq. (4.47), as

$$I_L = \frac{V_o^+}{Z_o} - \frac{V_o^-}{Z_o} \tag{4.168}$$

Dividing Eq. (4.166) by Z_o and adding the resulting equation to (4.168), we obtain

$$V_o^+ = \frac{1}{2}(V_L + I_L Z_o) \tag{4.169}$$

or

$$V_o^+ = \frac{1}{2} I_L (Z_L + Z_o) \tag{4.170}$$

after substituting $V_L = Z_L I_L$. Similarly, dividing (4.166) by Z_o, subtracting the resulting equation from (4.168), and using $V_L = Z_L I_L$, we get

$$V_o^- = \frac{1}{2} I_L (Z_L - Z_o) \tag{4.171}$$

Now substituting V_o^+ and V_o^- from (4.170) and (4.171), respectively, into (4.40) gives

$$V(z) = \frac{I_L}{2}[(Z_L + Z_o)e^{-\gamma z} + (Z_L - Z_o)e^{\gamma z}] \tag{4.172}$$

from which, we also obtain

$$V^+(z) = \frac{I_L}{2}(Z_L + Z_o)e^{-\gamma z} \tag{4.173}$$

and

$$V^-(z) = \frac{I_L}{2}(Z_L - Z_o)e^{\gamma z} \tag{4.174}$$

[10]Note that in this convention used for mathematical convenience, the location z on the transmission line with respect to the load location ($z = 0$) is always negative.

Similarly, the current at any point along the transmission line can be derived from (4.41), making use of (4.47), (4.170), and (4.171), as

$$I(z) = \frac{I_L}{2Z_o}[(Z_L + Z_o)e^{-\gamma z} - (Z_L - Z_o)e^{\gamma z}] \tag{4.175}$$

which is formed by

$$I^+(z) = \frac{I_L}{2Z_o}(Z_L + Z_o)e^{-\gamma z} \tag{4.176}$$

and

$$I^-(z) = -\frac{I_L}{2Z_o}(Z_L - Z_o)e^{\gamma z} \tag{4.177}$$

The impedance at any location z looking toward the load is obtained from (4.172) and (4.175) as

$$\begin{aligned} Z(z) = \frac{V(z)}{I(z)} &= Z_o \frac{(Z_L + Z_o)e^{-\gamma z} + (Z_L - Z_o)e^{\gamma z}}{(Z_L + Z_o)e^{-\gamma z} - (Z_L - Z_o)e^{\gamma z}} \\ &= Z_o \frac{Z_L(e^{\gamma z} + e^{-\gamma z}) - Z_o(e^{\gamma z} - e^{-\gamma z})}{Z_L(e^{-\gamma z} - e^{\gamma z}) + Z_o(e^{\gamma z} + e^{-\gamma z})} \end{aligned} \tag{4.178}$$

Using the trigonometry identities

$$\cosh \gamma z = \frac{e^{\gamma z} + e^{-\gamma z}}{2} \tag{4.179}$$

and

$$\sinh \gamma z = \frac{e^{\gamma z} - e^{-\gamma z}}{2} \tag{4.180}$$

in (4.178), we get

$$Z(z) = Z_o \frac{Z_L \cosh \gamma z - Z_o \sinh \gamma z}{-Z_L \sinh \gamma z + Z_o \cosh \gamma z} \tag{4.181}$$

Dividing (4.181) by $\cosh \gamma z$, we obtain

$$Z(z) = Z_o \frac{Z_L - Z_o \tanh \gamma z}{Z_o - Z_L \tanh \gamma z} \tag{4.182}$$

Taking a reciprocal of (4.182), we can obtain the admittance at any location z looking toward the load as

$$Y(z) = Y_o \frac{Y_L - Y_o \tanh \gamma z}{Y_o - Y_L \tanh \gamma z} \tag{4.183}$$

where $Y_o = 1/Z_o$ and $Y_L = 1/Z_L$ are the characteristic admittance of the transmission line and the load admittance, respectively. Note that the impedance and admittance calculations based on (4.182) and (4.183), respectively, only take "negative" values for z. Of particular interest are the impedance and admittance at the input of the transmission line ($z = -\ell$) looking into the load:

$$Z_i = Z_o \frac{Z_L + Z_o \tanh \gamma l}{Z_o + Z_L \tanh \gamma l} \tag{4.184}$$

and

$$Y_i = Y_o \frac{Y_L + Y_o \tanh \gamma l}{Y_o + Y_L \tanh \gamma l} \tag{4.185}$$

4.8.2 Reflection Coefficients

The ratio of the reflected voltage wave to the incident voltage wave is defined as the voltage reflection coefficient, and is obtained from Eqs. (4.173) and (4.174) as

$$\Gamma(z) \equiv \frac{V^-(z)}{V^+(z)} = \frac{Z_L - Z_o}{Z_L + Z_o} e^{2\gamma z} \tag{4.186}$$

Equation (4.178) can be rewritten, upon using (4.186), as

$$Z(z) = Z_o \frac{1 + \Gamma(z)}{1 - \Gamma(z)} \tag{4.187}$$

from which, we can derive

$$\Gamma(z) = \frac{Z(z) - Z_o}{Z(z) + Z_o} \tag{4.188}$$

At the load ($z = 0$), the reflection coefficient, namely the load reflection coefficient, is obtained as

$$\Gamma_L = |\Gamma_L| e^{j\phi_L} = \frac{V_o^-}{V_o^+} = \frac{Z_L - Z_o}{Z_L + Z_o} \tag{4.189}$$

where ϕ_L is its phase. Combining (4.186) and (4.189) gives

$$\Gamma(z) = \Gamma_L e^{2\gamma z} \tag{4.190}$$

Similarly, the current at any point along the transmission line can be derived from (4.41), making use of (4.47), (4.169) and (4.170), as

$$I(z) = \frac{I_L}{2Z_o}[(Z_L + Z_o)e^{-\gamma z} - (Z_L - Z_o)e^{\gamma z}] \tag{4.191}$$

from which, we can obtain the current reflection coefficient $\Gamma_I(z)$ as the ratio between the reflected and incident current waves:

$$\Gamma_I(z) \equiv \frac{I^-(z)}{I^+(z)} = -\frac{Z_L - Z_o}{Z_L + Z_o} e^{2\gamma z} \tag{4.192}$$

Comparing (4.186) and (4.192), we get

$$\Gamma_I(z) = -\Gamma(z) \tag{4.193}$$

We can see, from (4.186) and (4.192), that the magnitudes of the reflection coefficients in a lossless transmission line are always constant; only their phases vary as a function of location. In practice, the voltage reflection coefficient is commonly used to indicate the reflection coefficient. We will adapt this convention in this book. Note that the reflection coefficient is dimensionless and its magnitude varies between 0 and 1. The magnitude of reflection coefficient is also often expressed in decibel (dB) as $10 \log |\Gamma(z)|^2$, which is known as return loss.

4.8.3 Standing-Wave Ratio

The voltage at any location along a transmission line is obtained from Eq. (4.40), upon using (4.189), as

$$\begin{aligned} V(z) &= V_o^+ e^{-\alpha z} e^{-j\beta z} + V_o^- e^{\alpha z} e^{j\beta z} \\ &= V_o^+ e^{-\alpha z} e^{-j\beta z} (1 + |\Gamma_L| e^{2\alpha z} e^{j(2\beta z + \phi_L)}) \end{aligned} \tag{4.194}$$

Equation (4.194) indicates that the amplitude and phase of the incident voltage wave reduces as the load is approached, as illustrated in Figure 4.12, whereas those of the reflected voltage wave increases. So it is expected that there are locations along the transmission line, at which the amplitudes of the incident and reflected waves are equal, and there are also locations at which the phases of these opposite waves are equal or 180° out of phase. At the equal-phase locations, the two waves will add constructively while they will subtract at the out-of-phase locations.

The magnitude of the voltage at any point on a transmission line can be obtained from (4.194) as

$$|V(z)| = |V_o^+ e^{-\alpha z}||1 + |\Gamma_L| e^{2\alpha z} e^{j(2\beta z + \phi_L)}| \tag{4.195}$$

Replacing $e^{j(2\beta z + \phi_L)}$ with $\cos(2\beta z + \phi_L) + j\sin(2\beta z + \phi_L)$ and taking the magnitude of the resulting equation gives

$$\begin{aligned} |V(z)| &= |V_o^+ e^{-\alpha z}| \{[1 + |\Gamma_L| e^{2\alpha z} \cos(2\beta z + \phi_L)]^2 + |\Gamma_L^2| e^{4\alpha z} \sin^2(2\beta z + \phi_L)\}^{1/2} \\ &= |V_o^+ e^{-\alpha z}| \{(1 + |\Gamma_L| e^{2\alpha z})^2 - 2|\Gamma_L| e^{2\alpha z} [1 - \cos(2\beta z + \phi_L)]\}^{1/2} \end{aligned} \tag{4.196}$$

which becomes, after using the trigonometry $\cos 2x = 1 - 2\sin^2 x$,

$$|V(z)| = |V_o^+ e^{-\alpha z}| \left\{ \left(1 + |\Gamma_L|\, e^{2\alpha z}\right)^2 - 4|\Gamma_L| e^{2\alpha z} \sin^2 \left(\beta z + \frac{\phi_L}{2}\right) \right\}^{1/2} \tag{4.197}$$

Examining (4.197) reveals that it has maximum values of

$$|V(z)|_{\max} = |V_o^+ e^{-\alpha z}|(1 + |\Gamma_L e^{2\alpha z}|) \tag{4.198}$$

corresponding to

$$z = \frac{n\pi - \phi_L/2}{\beta}, \quad n = 0, 1, 2, \ldots \tag{4.199}$$

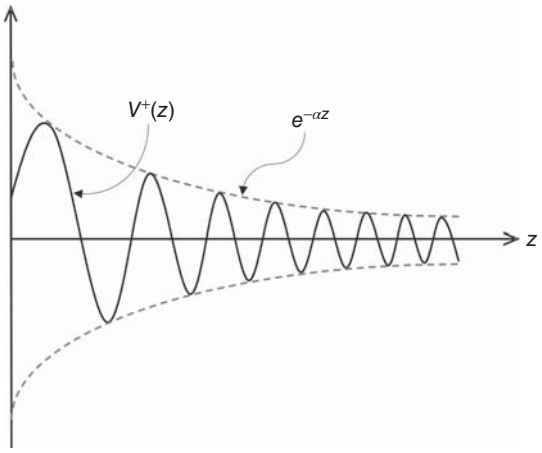


Figure 4.12. Illustration of incident wave $V^+(z)$ along a transmission line.

and minimum values of

$$|V(z)|_{\min} = |V_o^+ e^{-\alpha z}|(1 - |\Gamma_L e^{2\alpha z}|) \tag{4.200}$$

corresponding to

$$z = \frac{(n + 1/2)\pi - \phi_L/2}{\beta}, \quad n = 0, 1, 2, \ldots \tag{4.201}$$

These results show that the magnitude of the voltage along a transmission line fluctuates as a function of location between maximum and minimum values at periodic positions. The maximum and minimum voltages are due to the addition between the incident and reflected voltage waves at locations where they have the same phase and the subtraction at locations where they are 180° out of phase, respectively. The magnitude also reduces as the signal propagates along the transmission line due to loss in the line, which also results in different values for maximum (minimum) at different locations. This is in contrast with the circuit theory, which indicates that the voltage along a conductor (which is part of a two-conductor transmission line) is always constant. The spacing between two successive maxima or minima is a half-wavelength as can be determined from (4.199) and (4.201), respectively. (4.199) and (4.201) also indicate that the distance between a maximum and its nearest minimum is a quarter-wavelength.

The resultant voltage waveform formed by the incident and reflected signals, as given in (4.194) and (4.197), is referred to as the voltage standing-wave pattern. It is noted that, at the load, the incident-voltage amplitude is smallest with respect to those away from the load, whereas the amplitude of the reflected voltage is largest as compared to others at different locations. These incident and reflected voltage amplitudes at the load are thus not equal (unless $Z_L = 0$ or ∞); thereby do not create a pure standing wave at the load except when $Z_L = 0$ or ∞. However, at certain location z_o along the line, the amplitudes of the incident and reflected waves are equal as discussed earlier and, hence, standing waves would be formed in the vicinity of this location, as illustrated in Figure 4.13. Far away from the location z_o in the $+z$ and $-z$ directions, the incident and reflected waves die out, respectively, resulting in the standing waves gradually vanishing as the waves move away from the location. The standing wave is therefore a function of location along the line; its magnitude is large near the location z_o and reduces at locations away from z_o. It should be noted that, in general, a pure standing wave does not exist along a transmission line except when $|V_o^+ e^{-\alpha z}| = |V_o^- e^{\alpha z}|$, for example, in a short-circuited lossless transmission line. Indeed, the wave can be considered having a standing wave and a traveling wave simultaneously at any location and at any time along a transmission line. In practical transmission lines, the loss is normally small, causing an almost constant-magnitude standing-wave pattern along the line as seen in Figure 4.13(b), which is similar to the standing-wave pattern for a lossless transmission line. True standing wave is possible only in transmission lines terminated with a short, open or reactance as seen later.

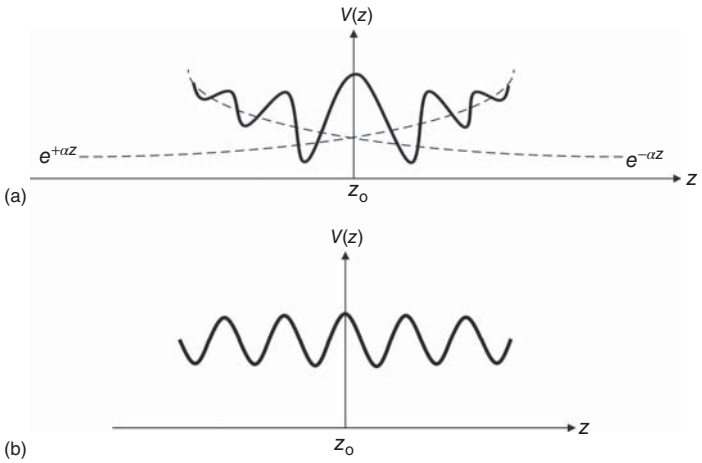


Figure 4.13. Standing wave pattern in high loss (a) and low loss (b) transmission line terminated with load $Z_L \neq Z_o$. The standing wave pattern for lossless line is similar to that in (b).

The ratio between $|V(z)|_{\max}$ and $|V(z)|_{\min}$ is defined as the voltage standing–wave ratio (VSWR):

$$\text{VSWR} = \frac{1 + |\Gamma_L e^{2\alpha z}|}{1 - |\Gamma_L e^{2\alpha z}|} \tag{4.202}$$

which is, like reflection coefficient, dimensionless and varies between 1 and infinity.

Following the same procedure, we can also show that the current along a transmission line oscillates between its maximum and minimum values as the current propagates, forming a current wave consisting of both standing and traveling waves. The maximum and minimum values for the current, however, occur at locations corresponding to the minimum and maximum voltages, respectively. These maximum or minimum locations are also successively spaced a half-wavelength apart. The current SWR is equal to VSWR. In practice, VSWR is commonly used instead of the current SWR.

4.8.4 Perfect Match and Total Reflection

4.8.4.1 Perfect Match. When the load impedance is equal to the transmission-line characteristic impedance, we obtain respectively from (4.189) and (4.202):

$$\Gamma_L = 0 \tag{4.203}$$

and

$$\text{VSWR} = 1 \tag{4.204}$$

which imply that the transmission line has no reflection and is perfectly matched at the load. Under this condition, all incident voltage and current waves, or incident power,[11] will be absorbed by the load. Similarly, when the source impedance is equal to the transmission-line characteristic impedance, the transmission line is perfectly matched to the source and no reflection occurs at the source location.

As can be seen in (4.190), the magnitude of the reflection coefficient reduces exponentially as the wave moves away from the load. In the limit when the distance is large, this magnitude reduces to zero. At that location, VSWR is also equal to 1. This phenomenon is also encountered in a short transmission line when the loss is high. So when a transmission line has a large loss or long length, it has no reflection at locations far away from the load.

4.8.4.2 Total Reflection. Total reflection in a transmission line occurs when the reflection coefficient or VSWR equals to 1 or infinity, respectively, as can be seen in the following cases.

Open-Circuited Transmission Line

An open-circuited transmission line is formed when it is terminated by an open ($Z_L \simeq \infty$). The reflection coefficient at the load and VSWR can be obtained from Eqs. (4.189) and (4.202) as

$$\Gamma_L = 1 \tag{4.205}$$

$$\text{VSWR} \simeq \infty \tag{4.206}$$

Equation (4.205) shows that the reflected and incident voltages are equal, signifying that the incident voltage and power are completely reflected back with equal phase. The transmission line has total reflection. It is

[11] Power is indeed a wave, not a stationary quantity, as can be inferred from its constituent voltage and current waves.

noted that VSWR is only equal to infinity when the transmission line has no loss ($\alpha = 0$) or at the load ($z = 0$), as can be determined from (4.202). The voltage on the line can be expressed using (4.194) and (4.205) as

$$\begin{aligned} V(z) &= V_o^+ e^{-\gamma z}(1 + \Gamma_L e^{2\gamma z}) \\ &= V_o^+ (e^{-\gamma z} + e^{\gamma z}) \\ &= 2V_o^+ \cosh(\gamma z) \end{aligned} \tag{4.207}$$

which clearly represents a pure standing wave. Traveling wave no longer exists along an open-circuited transmission line.

Short-Circuited Transmission Line

A short-circuited transmission line is formed when it is short-circuited at the load ($Z_L = 0$). The reflection coefficient at the load and VSWR are obtained from Eqs. **(4.189)** and (4.202) as

$$\Gamma_L = -1 \tag{4.208}$$

$$\text{VSWR} \simeq \infty \tag{4.209}$$

The reflected and incident voltages have an equal magnitude but opposite phase, indicating that all incident voltage and power are reflected back with the phase reversed. There is total reflection along the line. It is noted that VSWR is only equal to infinity when the transmission line has no loss ($\alpha = 0$) or at the load ($z = 0$), as can be determined from (4.202). The total voltage along the line are found from (4.194) and (4.208) to be

$$V(z) = -2V_o^+ \sinh(\gamma z) \tag{4.210}$$

representing a pure standing wave. Traveling wave no longer exists along a short-circuited transmission line.

Transmission Line Terminated with Reactance

When a transmission line is terminated with a capacitor or inductor, the magnitude of the reflection coefficient at the load is

$$|\Gamma| = 1 \tag{4.211}$$

The reflected and incident voltages have equal magnitude and all incident voltage and power are reflected back.

4.8.5 Lossless Transmission Lines

All equations derived in Sections 4.8.1–4.8.4 are applicable for lossy as well as lossless transmission lines. For lossless transmission lines, the attenuation constant $\alpha = 0$ and the propagation constant $\gamma = j\beta$, where β is the phase constant. Making use of these and recognizing that $\tan(j\beta l) = j\tan(\beta l)$, we can derive from Eqs. (4.182)–(4.185):

$$Z(z) = Z_o \frac{Z_L - jZ_o \tan \beta z}{Z_o - jZ_L \tan \beta z} \tag{4.212}$$

$$Y(z) = Y_o \frac{Y_L - jY_o \tan \beta z}{Y_o - jY_L \tan \beta z} \tag{4.213}$$

$$Z_i = Z_o \frac{Z_L + jZ_o \tan \beta l}{Z_o + jZ_L \tan \beta l} \tag{4.214}$$

$$Y_i = Y_o \frac{Y_L + jY_o \tan \beta l}{Y_o + jY_L \tan \beta l} \tag{4.215}$$

$\theta = \beta l$ is called the electrical length of the line and depends on frequency. Electrical length actually represents the phase of the transmission line and is frequently used in high frequency circuit design. Note that the input impedance Z_i and input admittance Y_i are periodic functions with respect to βl and therefore repeat themselves every half-wavelength along the transmission line. From (4.186), (4.190), and (4.192), we obtain

$$\Gamma(z) = \frac{Z_L - Z_o}{Z_L + Z_o} e^{2j\beta z} \tag{4.216}$$

$$\Gamma(z) = \Gamma_L e^{2j\beta z} \tag{4.217}$$

$$\Gamma_I(z) = -\frac{Z_L - Z_o}{Z_L + Z_o} e^{2j\beta z} \tag{4.218}$$

from which, we get

$$|\Gamma(z)| = |\Gamma_I(z)| = |\Gamma_L| = \left|\frac{Z_L - Z_o}{Z_L + Z_o}\right| \tag{4.219}$$

From (4.202), we obtain

$$\text{VSWR} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} \tag{4.220}$$

and hence,

$$|\Gamma(z)| = |\Gamma_L| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \tag{4.221}$$

These (lossless) equations are commonly used in practice. It is noted that the VSWR and magnitude of the reflection coefficient in a lossless transmission line are always constant. This is expected as the magnitudes of the incident and reflected voltage waves remain constant as they propagate along a lossless transmission line. The voltage standing-wave pattern on a lossless transmission line is similar to that for a low loss transmission line as shown in Figure 4.13(b).

The input impedance of an open-circuited lossless transmission line can be derived from Eq. (4.214), by letting $Z_L = \infty$, as

$$Z_{\text{io}} = Z_o \frac{Z_L + jZ_o \tan \beta l}{Z_o + jZ_L \tan \beta l} = -jZ_o \operatorname{cotan} \beta l \tag{4.222}$$

The input admittance is

$$Y_{\text{io}} = jY_o \tan \beta l \tag{4.223}$$

Similarly, we can derive the input impedance and admittance of a short-circuited lossless transmission line from (4.214) by letting $Z_L = 0$, as

$$Z_{\text{is}} = jZ_o \tan \beta l \tag{4.224}$$

$$Y_{\text{is}} = -jY_o \cot \beta l \tag{4.225}$$

Examining (4.222)–(4.225) reveals that open- and short-circuited lossless lines behave as an inductor or a capacitor depending on the operating frequency and the transmission line's physical length. At a given frequency, any inductor or capacitor can be theoretically realized by choosing a proper length for these transmission lines. Figures 4.14 and 4.15 illustrate the inductive and capacitive behaviors of open- and short-circuited lossless transmission lines, respectively, versus physical and electrical lengths.

When the transmission line is very short compared to wavelength, $\beta l = 2\pi l/\lambda \ll 1$. We can then approximate $\tan \beta l$ as βl and obtain from (4.222)

$$Z_{\text{io}} \cong -j\frac{Z_o}{\beta l} \tag{4.226}$$

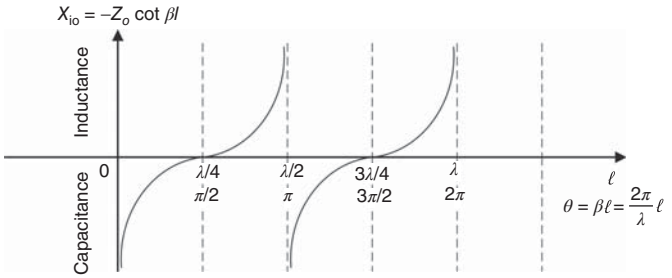


Figure 4.14. Impedance behavior of an open-circuited lossless transmission lines versus lengths.

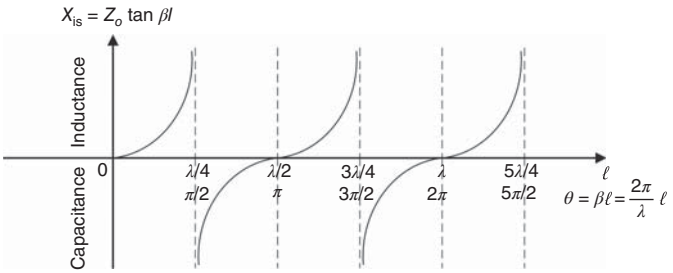


Figure 4.15. Impedance behavior of a short-circuited lossless transmission lines versus lengths.

Replacing $Z_o = \sqrt{L/C}$ from (4.60) and $\beta = \omega\sqrt{LC}$ from (4.61) leads to

$$Z_{\text{io}} \cong \frac{1}{j\omega(Cl)} \tag{4.227}$$

which shows that a very short open-circuited lossless transmission line behaves as a capacitor having a capacitance value of Cl. Note that C is the capacitance per unit length of the line and a transmission line (whether open-circuited or not), whose length is not very short compared to wavelength, does not behave as a capacitor of Cl. Similarly, we can drive the following equation for a short-circuited lossless transmission line when its length is very short compared to wavelength:

$$Z_{\text{is}} \cong j\omega(Ll) \tag{4.228}$$

which indicates that a very short short-circuited line behaves as an inductor with an inductance of Ll. Note that L is the inductance per unit length of the line and a transmission line (whether short-circuited or not), whose length is not very short compared to wavelength, does not behave as an inductor of Ll.

Two special cases worth mentioned are the open- and short-circuited transmission lines whose lengths are multiples of quarter-wavelength. Substituting $l = n\lambda/4$, where n is a positive integer, into Eqs. (4.222) and (4.224), we obtain

$$Z_{\text{io}} = 0 \quad \text{for open-circuited lossless line} \tag{4.229}$$

and

$$Z_{\text{is}} = \infty \quad \text{for short-circuited lossless line} \tag{4.230}$$

These results indicate that open- and short-circuited transmission lines of multiple quarter wavelengths behave as short and open circuits at their inputs, respectively. Quarter-wavelength open- and short-circuited transmission lines ($l = \lambda/4$) are often used in bias networks for high frequency narrow-band circuits.

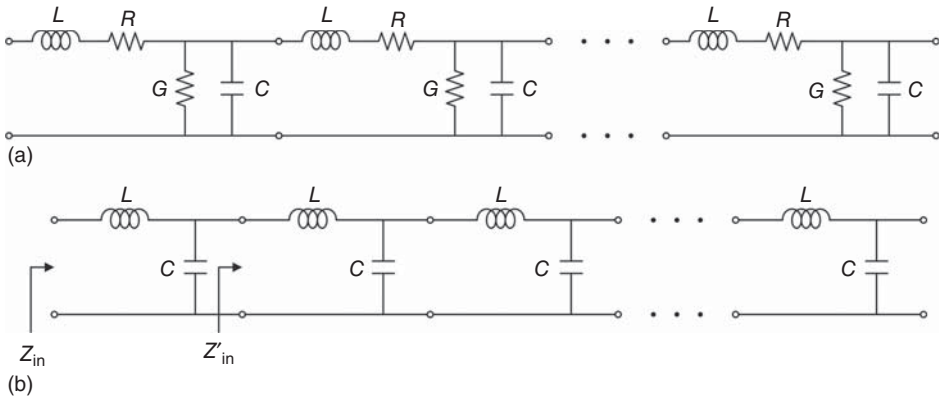


Figure 4.16. Approximate equivalent circuit of lossy (a) and lossless (b) transmission lines.

4.9 SYNTHETIC TRANSMISSION LINES

As discussed in Section 4.2, a transmission line can be approximately represented as an equivalent circuit consisting of many identical subcircuits, each subcircuit corresponding to an infinitesimal section of the transmission line, as shown in Figure 4.16. For lossless transmission lines, $R = G = 0$ as seen in Figure 4.16(b), and a signal would traverse without loss, implying that uniform lossless transmission lines can be operated over extremely wide bandwidths up to the limit of the TEM or quasi-TEM cut-off frequency (theoretically infinite bandwidth).

Examination of Figure 4.16(b) reveals that synthetic (or artificial) transmission lines may be formed by cascading multiple identical LC sections or cells. It is noted that R and G seen in Figure 4.16(a) represent losses due to the conductors and substrate, respectively, which are undesirable. Therefore, they are not used in creating synthetic transmission lines. Nevertheless, R and G are implicitly included in synthetic transmission lines due to the losses associated with the constituent inductors and capacitors, respectively. These losses essentially limit the operating bandwidth of synthetic transmission lines. It is expected that the larger number of sections (N) for a synthetic transmission, the closer resemblance of an actual transmission line. Synthetic transmission lines can exploit the properties of transmission lines to achieve certain functions without using actual transmission lines or eliminate the need of configuring actual transmission lines in particular fashions (e.g., microstrip line) which might be difficult and/or inconvenient to realize under certain cases.

We now consider a lossless synthetic transmission line with infinite number of sections ($N \to \infty$) as represented in Figure 4.16(b), in which we assume that the inductors and capacitors are ideal. Its characteristic impedance and phase velocity can be treated the same as those for lossless transmission lines. That is, they can be obtained from (4.60) and (4.62) as

$$Z_o = \sqrt{\frac{L}{C}} \tag{4.231}$$

and

$$v_p = \frac{1}{\sqrt{LC}} \tag{4.232}$$

respectively. Note that herein L and C are the inductance and capacitance of each section making up the synthetic transmission line. The time delay per section of a synthetic transmission line is considered equivalent to the time delay per unit length of a corresponding actual transmission line and, hence, can be obtained as

$$t_d = \sqrt{LC} \tag{4.233}$$

We can see that a synthetic transmission line provides flexibility in realizing a particular characteristic impedance and phase velocity by properly choosing L and C. It can also achieve certain time delay (small or large) with suitable L, C, and N. Using actual transmission lines may require a long physical length in order to obtain a large time delay.

When the constituent inductors and capacitors have small loss, the corresponding synthetic transmission line can be considered having low loss and, similar to an actual low loss transmission line, its characteristic impedance, attenuation constant, phase constant and velocity can be obtained from (4.70)–(4.73) as

$$Z_o \cong \sqrt{\frac{L}{C}}\left[1 + j\left(\frac{G}{2\omega C} - \frac{R}{2\omega L}\right)\right] \tag{4.234}$$

$$\alpha \cong \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) \tag{4.235}$$

$$\beta \cong \omega\sqrt{LC} \tag{4.236}$$

$$v \cong \frac{1}{\sqrt{LC}} \tag{4.237}$$

where R and G are the resistance and conductance representing loss in the inductor and capacitor of each cell, respectively.

The lossless synthetic transmission line as shown in Figure 4.16(b) resembles a low pass filter and, therefore, it is associated with a cutoff frequency ω_c, which defines pass-band and stop-band below and above ω_c, respectively. As N is of infinity, it is justified to assume that $(N-1)$ also approaches infinity. As such, the input impedances Z_{in} and Z'_{in} would be equal. As indicated in Figure 4.16(b), we can write Z_{in} as a combination of series L and parallel (C, Z'_{in}):

$$Z_{\text{in}} = j\omega L + \frac{1}{\frac{1}{Z_{\text{in}}} + j\omega C} \tag{4.238}$$

where we have replaced Z'_{in} in the right-high side with Z_{in}. Expanding (4.238) gives

$$Z_{\text{in}}^2 - j\omega L Z_{\text{in}} - \frac{L}{C} = 0 \tag{4.239}$$

which can be solved for Z_{in}:

$$Z_{\text{in}} = \frac{1}{2}j\omega L \pm \frac{1}{2}j\sqrt{\frac{L}{C}}\sqrt{\omega^2 LC - 4} \tag{4.240}$$

Equation (4.240) shows that when $\omega^2 LC - 4 < 0$, corresponding to $\omega < 2/\sqrt{LC}$, Z_{in} is complex, and signals with $\omega < 2/\sqrt{LC}$ can propagate along the transmission line, hence signifying a pass-band. In the limit of ω approaching 0, Z_{in} approaches the characteristic impedance $Z_o = \sqrt{L/C}$. On the other hand, when $\omega^2 LC - 4 > 0$, $\omega > 2/\sqrt{LC}$ and correspondingly Z_{in} becomes imaginary. Signals whose radian frequencies are above $2/\sqrt{LC}$ thus cannot propagate on the transmission line, thereby indicating a stop-band corresponding to a cutoff frequency of

$$\omega_c = \frac{2}{\sqrt{LC}} \tag{4.241}$$

For ideal (lossless) synthetic transmission line, the cutoff frequency becomes, making use of (4.231),

$$\omega_c = \frac{2}{CZ_o} \tag{4.242}$$

The cutoff frequency effectively sets the operating bandwidth of synthetic transmission lines from DC to ω_c.

A synthetic transmission line should therefore be designed so that its highest operating frequency is below $\omega_c = 2/\sqrt{LC}$. In practice, the number of sections N making up a synthetic transmission line is always finite, so the actual cutoff frequency is smaller (in fact, may be much smaller) than the theoretical value calculated from (4.241), which corresponds to infinite N. Therefore, the highest operating frequency may need to be much smaller than ω_c for proper operation. Typically, the cutoff frequency is chosen to be at least three times of the highest operating frequency. Actual simulation of a designed synthetic transmission line needs to be performed to make sure its cutoff frequency is well above the intended operating frequency range. It is particularly noted that a synthetic transmission line is not an exact transmission line; it only approximates a transmission line over a certain bandwidth, which may be wide or narrow depending on the number of sections used and the values of L and C.

In some applications, such as those requiring multiple frequencies to be transmitted concurrently, frequency-independent velocity or linear phase response versus frequency in the pass-band is needed to avoid signal distortion. The lossless and low loss synthetic transmission lines have exact and approximate frequency-independent velocities as seen in (4.232) and (4.237), respectively, and correspondingly producing no and minimum signal distortion theoretically.

A synthetic transmission line can be designed by choosing L and C to achieve desired characteristic impedance per (4.231) and desired cutoff frequency (or bandwidth) per (4.241). For a wide bandwidth, LC is kept sufficiently small to produce a high cutoff frequency. Effectively, there is a trade-off between the operating bandwidth and characteristic impedance (and hence matching for circuits implementing the synthetic transmission line). In order to achieve a certain time delay, the number of required sections can be determined from

$$N = \frac{T_d}{t_d} = \frac{T_d}{\sqrt{LC}} \tag{4.243}$$

where T_d is the desired time delay.

It is noted that synthetic transmission lines, as suggested in their equivalent circuit, can also be realized using solid-state devices, such as MOSFETs, operated in the linear (small signal) regime to implement both inductors and capacitors, or a combination between lumped elements and solid-state devices. Using solid-state devices would make smaller synthetic transmission lines, which are attractive for low cost RFIC applications; yet at possible degradation of circuit performance such as increased noise and reduced linearity, which are not desirable, especially in the receiver front-end. Particularly, for large-signal circuits, such as power amplifiers, solid-state devices should not be used due to their nonlinear behavior under large signals, unless they are intended to be used as nonlinear elements in the designed synthetic transmission line.

The main function of a synthetic transmission line is to maintain as close as possible its desired characteristic impedance within a specified tolerance over a bandwidth as wide as possible. Additionally, under some operating conditions, it is also desired to maintain linear phase response as much as possible to avoid signal distortion due to phase nonlinearity. In practice, inductors, capacitors, and semiconductor devices used in synthetic transmission lines have their own parasitics, generally consisting of resistors, inductors, and capacitors, which may severely degrade the transmission-line performance, especially in high RF regions. For instance, the parasitics cause increase in loss and more distortion in the transmission phase, leading to reduced operating bandwidth. Therefore, care must be exercised when designing a synthetic transmission line; all parasitics of its constituent elements must be considered in simulations to make sure that the introduced attenuation and distortion will not ruin the integrity of the signal passing through the transmission line within the operating frequency range. In order to lessen these effects, the number of sections needs to be limited considering trade-off with the operating bandwidth.

4.10 TEM AND QUASI-TEM TRANSMISSION-LINE PARAMETERS

Transmission lines with perfect conductors embedded in a homogeneous medium, such as strip line, support pure TEM mode or wave. However, transmission lines with more than two dielectrics (or embedded

in an inhomogeneous medium), such as microstrip line, can only support quasi-TEM mode which resembles TEM mode. The TEM or quasi-TEM mode is the dominant mode in transmission lines and has no cutoff frequency, implying that transmission lines can support signals from DC up to very high frequencies at which higher-order modes begin to propagate. The most important parameters of a transmission line for circuit design purposes, assuming the line is not very lossy,[12] are perhaps the characteristic impedance and effective dielectric constant for the quasi-TEM mode.

As the operating frequency increases, an infinite number of modes including quasi-TEM and higher-order modes[13] may exist in transmission lines, particularly near discontinuities. Higher-order modes have cutoff frequencies which limit the operating frequency ranges and are undesirable. These modes have either longitudinal electric or magnetic field or both. The modes having longitudinal electric field is called transverse magnetic (TM) modes and those having longitudinal magnetic field is referred to as transverse electric (TE) modes. The modes having both longitudinal electric and magnetic fields are known as hybrid modes. When a higher-order mode coexists with the quasi-TEM mode above a certain frequency, it will disrupt the circuit performance since it will take away some of the energy of the quasi-TEM mode and, particularly, will generate spurious responses in active circuits. The share of energy that sustains the higher-order mode is inevitable due to the conservation of power and the fact that the total energy propagating in a transmission line is finite.

There are two approaches in analyzing a transmission line: static or quasi-static and dynamic or full-wave approaches. The static or quasi-static approach only produces transmission-line parameters for the TEM or quasi-TEM mode that are theoretically valid only at DC. On the other hand, the dynamic approach can not only produce the frequency-dependent transmission-line parameters for the TEM or quasi-TEM mode, but also produce those for the higher-order modes, whose parameters are functions of frequency.

As stated earlier, transmission line parameters obtained by the static or quasi-static approach are strictly valid at DC. In practice, however, these results can also be used at higher frequencies. The question most engineers would ask is how high of frequency they can use static results. There is no clear-cut answer for this. Some tends to believe that as long as the frequency is different from zero, static results cannot be used. On the other hand, most practical engineers perhaps use static results up to high frequencies such as around 18 GHz. In fact, a number of millimeter-wave circuits operating up to *W*-band (75–110 GHz) developed in the 1980s has been successfully designed using only static results. Nevertheless, at high frequencies, especially those in the millimeter-wave region, a dynamic approach should be employed for more accurate determination of the parameters of transmission lines used in RFICs. The main trade-off between using static versus dynamic results is that the former is simpler to calculate but less accurate.

4.10.1 Static or Quasi-Static Analysis

Static or quasi-static analysis produces transmission-line parameters that are frequency independent. We now derive simple equations for the static or quasi-static characteristic impedance and effective relative dielectric constant of transmission lines. Without lost of generality, we consider a lossless microstrip line, as shown in Figure 4.17(a), as a representative transmission line.

The characteristic impedance of this transmission line is given as

$$Z_o = \sqrt{\frac{L}{C}} \tag{4.244}$$

where L and C are the inductance and capacitance per unit length of the transmission line, respectively. Let L_a and C_a represent the inductance and capacitance per unit length of the transmission line when the dielectric is replaced by air (i.e., $\varepsilon_r = 1$) as shown in Figure 4.17(b). We can then rewrite (4.244), using the

[12] For transmission lines with high loss, the attenuation constant is also an important parameter.

[13] All these modes can exist in transmission lines since they all satisfy Maxwell's equations. In fact, higher-order modes exist near discontinuities in transmission lines at any frequencies.

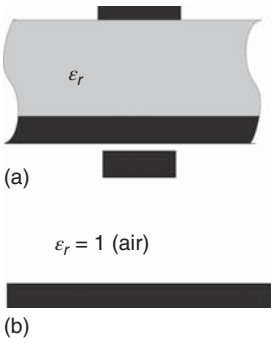


Figure 4.17. Microstrip line with (a) and without (b) substrate.

fact that the inductance per unit length does not depend upon the surrounding dielectrics, as

$$Z_o = \sqrt{\frac{L_a}{C_a}\frac{C_a}{C}} = \frac{1}{c\sqrt{CC_a}} \tag{4.245}$$

where $c = 3 \times 10^8$ m/s is the speed of light in air. The phase velocity of the quasi-TEM wave propagating along the transmission line is given as

$$v_p = \frac{c}{\sqrt{\varepsilon_{\text{reff}}}} \tag{4.246}$$

where $\varepsilon_{\text{reff}}$ is defined as the effective relative dielectric constant. Squaring both sides of (4.246) and making use of $L = L_a$, we obtain

$$\varepsilon_{\text{reff}} = \frac{C}{C_a} \tag{4.247}$$

The corresponding wavelength is given as

$$\lambda = \frac{\lambda_o}{\sqrt{\varepsilon_{\text{reff}}}} \tag{4.248}$$

where $\lambda_o = c/f$ is the free-space wavelength. It is now apparent that the static or quasi-static characteristic impedance and effective relative dielectric constant of any transmission line can be determined solely in terms of the transmission line's capacitances per unit length with and without dielectrics.

4.10.2 Dynamic Analysis

Dynamic analysis produces frequency-dependent characteristic impedance and propagation constant and, hence, attenuation and effective dielectric constant of transmission lines for TEM or quasi-TEM as well as hybrid modes. This approach begins by solving the wave equations discussed in Chapter 2 for the electric and magnetic fields subject to appropriate boundary conditions. The process also yields an eigenvalue equation whose eigenvalue is the propagation constant $\gamma = \alpha + j\beta$.

The phase constant β can be used to calculate the effective dielectric constant as

$$\varepsilon_{\text{eff}} = \left(\frac{\beta}{k_o}\right)^2 \tag{4.249}$$

where $k_o = \omega\sqrt{\varepsilon_o\mu_o}$ is the free-space wave number. It should be noted that the phase constant must follow the condition

$$\omega\sqrt{\varepsilon_o\mu_o} \le \beta \le \omega\sqrt{\varepsilon_o\mu_o\varepsilon_r} \tag{4.250}$$

where ε_r has the highest value of relative dielectric constants used for the transmission line.

Using the electric and magnetic fields, we can calculate the (dynamic) characteristic impedance as a function of frequency. It is well known that the characteristic impedance for the TEM or quasi-TEM mode is unique since the mode's voltage and current are uniquely defined. For a hybrid mode, however, the characteristic impedance is not unique. There exist various definitions for the characteristic impedance, and the common ones are based on the voltage and current, power and current, and power and voltage as given respectively below.

$$Z_o^{VI} = \frac{V_o}{I_o} \tag{4.251}$$

$$Z_o^{PI} = \frac{2P_{\text{avg}}}{|I_o|^2} \tag{4.252}$$

$$Z_o^{PV} = \frac{|V_o|^2}{2P_{\text{avg}}} \tag{4.253}$$

where V_o is the voltage across the conductors, I_o is the longitudinal current, and P_{avg} represents the time-average power transmitted across the transmission line's cross section. These three characteristic impedances are related by

$$Z_o^{VI} = \sqrt{Z_o^{PI} Z_o^{PV}} \tag{4.254}$$

It should be noted that different definitions of the characteristic impedance give different numerical results, except at DC. For TEM or quasi-TEM mode, however, all the three definitions give identical results. These definitions are also identical if $P_{\text{avg}} = \frac{1}{2} VI^*$. Generally, this condition is, however, not valid for hybrid modes in a transmission line. The definitions based on the power and current and power and voltage are deduced from a two-conductor transmission line. The choice of a particular definition is not very clear. A possible choice is to use a definition according to a specific use of the transmission line in a circuit. A common definition for transmission lines containing slots, such as slot line and coplanar waveguide (CPW), is based on the power and voltage. For instance, the characteristic impedance of CPW, shown in Figure 4.20, can be obtained from (4.253) with the voltage and power given as

$$V_o = \int_a^b E_x(x, h)dx \tag{4.255}$$

$$P_{\text{avg}} = \frac{1}{2} R_e \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (E_x H_y^* - E_y H_x^*)dxdy \tag{4.256}$$

The characteristic impedance of the slot line, shown in Figure 4.23, can also be defined by (4.253), where the power is again given by (4.256) and the voltage is obtained as

$$V_o = \int_{-W/2}^{W/2} E_x(x, h)dx \tag{4.257}$$

For other transmission lines without slots, the definitions based on the current and power and voltage and current are normally employed. For the case of a microstrip line, shown in Figure 4.19, the current, voltage, and power are given as

$$I_o = \int_{-W/2}^{W/2} J_z(x, h)dx \tag{4.258}$$

$$V_o = -\int_o^h E_y dy \tag{4.259}$$

$$P_{\text{avg}} = \frac{1}{2} R_e \int_o^{\infty} \int_{-\infty}^{\infty} (E_x H_y^* - E_y H_x^*) dx dy \tag{4.260}$$

where J_z is the current density on the strip along the z direction.

4.11 PRINTED-CIRCUIT TRANSMISSION LINES

Printed-circuit transmission lines, typically configured as planar and uniplanar structures, are important in RF integrated circuits (ICs). Printed-circuit transmission lines have evolved from the advances of electronic ICs and, in turn, they have helped pushing the progress of the IC technology, such as making the ICs more compact, more versatile, having better interconnections and improved performance, etc. Not only that printed-circuit transmission lines fulfill their most fundamental objective of delivering signals, they can also be exploited to create various RF components, such as wide-band hybrid junctions, by appropriately combining them. Various printed-circuit transmission lines have also been developed. The most commonly used structures are microstrip line, CPW, coplanar strip (CPS), strip line, and slot line; among them, the microstrip line and strip line, form the first planar version of the well-known coaxial transmission line. Figure 4.18 describes these printed-circuit transmission lines.

In this section, we will present a brief discussion of these transmission lines along with the closed-form formulas for computing their characteristic impedances, effective dielectric constants, and losses. In all these equations, infinitely wide ground planes are assumed. However, they are also valid for narrow ground planes provided that the field distributions are not significantly altered. Although these parameters can be accurately determined using various dynamic methods; the closed-form expressions allow convenient and fast computer-aided design (CAD) and analysis of RFICs. It is especially noted that all the printed-circuit transmission lines are presented with single dielectric and large ground planes (assumed to be infinitely wide) as conventionally employed. However, multiple dielectrics and/or narrow ground planes can be used to optimize the performance as well the size of the transmission lines. This kind of implementation is particularly

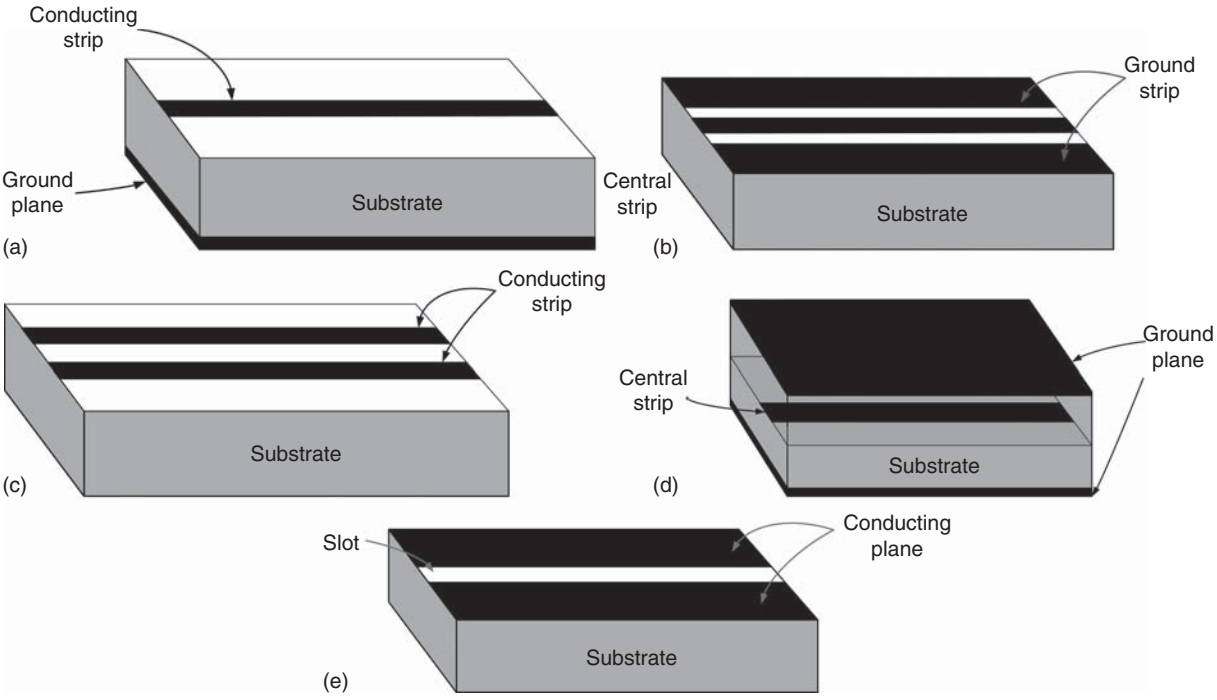


Figure 4.18. Commonly used printed-circuit transmission lines: (a) microstrip line, (b) CPW, (c) CPS, (d) strip line, and (e) slot line.

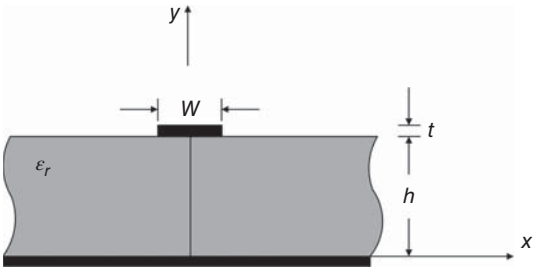


Figure 4.19. Cross-section of microstrip line.

suitable for RFICs due to the inherent existence of multiple dielectric and metal layers and is, in fact, needed since reduced size is one of the main concerns in RFIC design.

4.11.1 Microstrip Line

The microstrip line, as shown in Figure 4.19, was first proposed in 1952 [2], and since then it has become a widely used transmission line due to its planar nature, ease of fabrication using photolithographic processes, easy integration with solid-state devices, good heat sinking, and good mechanical support.

The closed-form expressions for the characteristic impedance and effective dielectric constant of the microstrip line, assuming zero strip thickness ($t = 0$), are given as [3]

$$Z_o = \begin{cases} \dfrac{119.9}{\sqrt{2\,(\varepsilon_r+1)}}\left\{\ln\left[4\dfrac{h}{W}+\sqrt{16\left(\dfrac{h}{W}\right)^2+2}\right]-\dfrac{1}{2}\left(\dfrac{\varepsilon_r-1}{\varepsilon_r-1}\right)\left(\ln\dfrac{\pi}{2}+\dfrac{1}{\varepsilon_r}\ln\dfrac{4}{\pi}\right)\right\}, & \dfrac{W}{h}<3.3 \\ \dfrac{119.9\pi}{2\sqrt{\varepsilon_r}}\left\{\dfrac{W}{2h}+\dfrac{\ln 4}{\pi}+\dfrac{\ln\left(\dfrac{e\pi^2}{16}\right)}{2\pi}\left(\dfrac{\varepsilon_r-1}{\varepsilon_r^2}\right)+\dfrac{\varepsilon_r+1}{2\pi\varepsilon_r}\left[\ln\dfrac{e\pi}{2}+\ln\left(\dfrac{W}{2h}+0.94\right)\right]\right\}^{-1}, & \dfrac{W}{h}>3.3 \end{cases} \tag{4.261}$$

and

$$\varepsilon_{\text{reff}} = \begin{cases} \dfrac{\varepsilon_r+1}{2}+\dfrac{\varepsilon_r-1}{2}\left(1+10\dfrac{h}{W}\right)^{-0.555}, & Z_o<(63-2\varepsilon_r)\,\Omega \\ \dfrac{\varepsilon_r+1}{2}\left[1-\dfrac{1}{2K}\left(\dfrac{\varepsilon_r-1}{\varepsilon_r+1}\right)\left(\ln\dfrac{\pi}{2}+\dfrac{1}{\varepsilon_r}\ln\dfrac{4}{\pi}\right)\right]^{-2}, & Z_o>(63-2\varepsilon_r)\,\Omega \end{cases} \tag{4.262}$$

where $e = 2.7182818$ is the Euler constant and

$$K = \frac{Z_o\sqrt{2(\varepsilon_r+1)}}{119.9}+\frac{1}{2}\left(\frac{\varepsilon_r-1}{\varepsilon_r+1}\right)\left(\ln\frac{\pi}{2}+\frac{1}{\varepsilon_r}\ln\frac{4}{\pi}\right) \tag{4.263}$$

The accuracy of the characteristic impedance and effective dielectric constant calculated from (4.261) and (4.262) is within ±1% and ±0.25%, respectively.

The normalized strip width, W/h, can also be determined from the characteristic impedance and relative dielectric constant as

$$\frac{W}{h} = \begin{cases} \dfrac{2}{\pi}\left[D - 1 - \ln(2D-1)\right] + \dfrac{\varepsilon_r - 1}{\pi\varepsilon_r}\left[\ln(D-1) + 0.293 - \dfrac{0.517}{\varepsilon_r}\right], & Z_o < (44 - 2\varepsilon_r)\,\Omega \\ \left(\dfrac{e^K}{8} - \dfrac{1}{4e^K}\right)^{-1}, & Z_o > (44 - 2\varepsilon_r)\,\Omega \end{cases} \tag{4.264}$$

where

$$D = \frac{59.95\pi^2}{Z_o\sqrt{\varepsilon_r}} \tag{4.265}$$

Using (4.264), W/h can be determined within $\pm 1\%$ of accuracy.

The practical microstrip line has finite strip thickness t, which effectively increases the strip width. Equations (4.261) and (4.262) can be modified to take into account t for more accurate results as [4]

$$Z_o(t) = \begin{cases} \dfrac{60}{\sqrt{\varepsilon_{\text{reff}}(t)}} \ln\left[\dfrac{8h}{W_e} + 0.25\dfrac{W_e}{h}\right], & \dfrac{W}{h} \le 1 \\ \dfrac{120\pi}{\sqrt{\varepsilon_{\text{reff}}(t)}}\left[\dfrac{W_e}{h} + 1.393 + 0.667\ln\left(\dfrac{W_e}{h} + 1.444\right)\right]^{-1}, & \dfrac{W}{h} \ge 1 \end{cases} \tag{4.266}$$

and

$$\varepsilon_{\text{reff}}(t) = \varepsilon_{\text{reff}} - C \tag{4.267}$$

where

$$\frac{W_e}{h} = \begin{cases} \dfrac{W}{h} + \dfrac{1.25}{\pi}\dfrac{t}{h}\left(1 + \ln\dfrac{4\pi W}{t}\right), & \dfrac{W}{h} \le \dfrac{1}{2\pi} \\ \dfrac{W}{h} + \dfrac{1.25}{\pi}\dfrac{t}{h}\left(1 + \ln\dfrac{2h}{t}\right), & \dfrac{W}{h} \ge \dfrac{1}{2\pi} \end{cases} \tag{4.268}$$

$$C = \frac{\varepsilon_r - 1}{4.6}\frac{t/h}{\sqrt{W/h}} \tag{4.269}$$

and $\varepsilon_{\text{reff}}$ is the effective relative dielectric constant with $t = 0$ obtained from (4.262).

The frequency-dependent effective dielectric constant and characteristic impedance can be found from [5] as

$$\varepsilon_{\text{reff}}(f) = \varepsilon_r - \frac{\varepsilon_r - \varepsilon_{\text{reff}}(0)}{1 + G(f/f_p)^2} \tag{4.270}$$

and

$$Z_o(f) = Z_o(0)\frac{\varepsilon_{\text{reff}}(f) - 1}{\varepsilon_{\text{reff}}(0) - 1}\sqrt{\frac{\varepsilon_{\text{reff}}(0)}{\varepsilon_{\text{reff}}(f)}} \tag{4.271}$$

where

$$f_p = \frac{Zo(0)}{2\mu_o h} \tag{4.272}$$

$$G = \frac{\pi^2}{12}\frac{\varepsilon_r - 1}{\varepsilon_{\text{reff}}(0)}\sqrt{\frac{Z_o(0)}{60}} \tag{4.273}$$

$\varepsilon_{\text{eff}}(0)$ and $Z_o(0)$ are the quasi-static effective dielectric constant and characteristic impedance, which can be determined using (4.262) or (4.267) and (4.261) or (4.266), respectively; and $\mu_o = 4\pi \times 10^{-7}$ H/m is the permeability of free space.

As for any transmission line, the loss in a microstrip line is due to imperfect conductors and dielectrics, and is characterized by the attenuation constant $\alpha = \alpha_c + \alpha_d$, where α_c and α_d represent the conductor and dielectric attenuation constants, respectively. α_c (dB/cm) can be determined from [6, 7]

$$\alpha_c = \begin{cases} \dfrac{8.68R_s}{2\pi Z_o h}\left[1-\left(\dfrac{W_e}{4h}\right)^2\right]\left\{1+\dfrac{h}{W_e}+\dfrac{h}{\pi W_e}\left[\ell n\left(\dfrac{4\pi W}{t}\right)+\dfrac{t}{W}\right]\right\}, & 0 < \dfrac{W}{h} \le \dfrac{1}{2\pi} \\ \dfrac{8.68R_s}{2\pi Z_o h}\left[1-\left(\dfrac{W_e}{4h}\right)^2\right]\left\{1+\dfrac{h}{W_e}+\dfrac{h}{\pi W_e}\left[\ell n\left(\dfrac{2h}{t}\right)-\dfrac{t}{h}\right]\right\}, & \dfrac{1}{2\pi} < \dfrac{W}{h} \le 2 \\ \dfrac{8.68R_s}{Z_o h}\left\{1+\dfrac{h}{W_e}+\dfrac{h}{\pi W_e}\left[\ell n\left(\dfrac{2h}{t}\right)-\dfrac{t}{h}\right]\right\}\left[\dfrac{W_e}{h}+\dfrac{\dfrac{W_e}{\pi h}}{\dfrac{W_e}{2h}+0.94}\right] & \\ \qquad \times\left[\dfrac{W_e}{h}+\dfrac{2}{\pi}\ell n\left(\dfrac{W_e}{2h}+0.94\right)\right]^{-2}, & \dfrac{W}{h} \ge 2 \end{cases} \tag{4.274}$$

where $R_s = \sqrt{\omega\mu_o/2\sigma}$ is the surface resistivity of the conductor, with σ being the conductivity. W_e is the effective strip width, taking into account the strip's finite metallization thickness, given in (4.268). α_d can be found by [8]

$$\alpha_d = \frac{27.3\varepsilon_r(\varepsilon_{\text{reff}}-1)\tan\delta}{\sqrt{\varepsilon_{\text{reff}}}(\varepsilon_r-1)\lambda_0} \quad \text{(dB/m)} \tag{4.275}$$

where $\tan\delta$ is the loss tangent of the dielectric and λ_o is the free-space wavelength. For the microstrip line or any other printed-circuit transmission lines employing low loss dielectrics, such as SiO_2 used in complementary metal oxide silicon (CMOS) structures, the dielectric loss is normally much lower than the conductor loss. However, when these transmission lines are deposited directly on low resistivity semiconductor substrates such as silicon in RFICs, the dielectric loss becomes dominant.

4.11.2 Coplanar Waveguide

Figure 4.20 shows cross sections of the conventional and conductor-backed CPW. The (conventional) CPW was first proposed in 1969 [9]. Since then, it has rapidly gained wide-spread use due to its many attractive features, such as elimination of via-holes in connecting circuit elements to ground, easy integration with solid-state devices, ease in realizing compact balanced circuits, and reduction of cross talk between lines. The CPW consists of a central conducting strip (signal line) and two ground conducting strips (ground lines). The two ground lines must be tied together so that they act as a single conductor in order to form a two-conductor transmission line together with the signal line. Otherwise, it can cause propagation problems at high frequencies due to the existence of the (quasi-TEM) c- and π-mode of a three-conductor transmission line as discussed in Section 4.13 for multiconductor transmission lines. In practice, they are connected via air-bridges. In the simplest equivalence, an air-bridge is modeled as a network consisting of series inductor, resistor and shunt capacitor, resistor, hence resulting in different potentials at the ground lines. For the conductor-backed CPW, via-holes are typically used to connect the back-conductor to the ground lines to keep them at equal potential and prevent or minimize the propagation of the other transmission-line modes and parallel-plate waveguide modes within the operating frequency range. However, since a via-hole would electrically consist of at least a resistor in series with an inductor, it inhibits the back-conductor from being at the same potential with the ground strips. The unequal potentials are more pronounced at very high frequencies, resulting in potential problems for the CPW to maintain as a two-conductor transmission line. The conductor-backed CPW particularly increases the mechanical strength as well as facilitate heat sinking and packaging needed

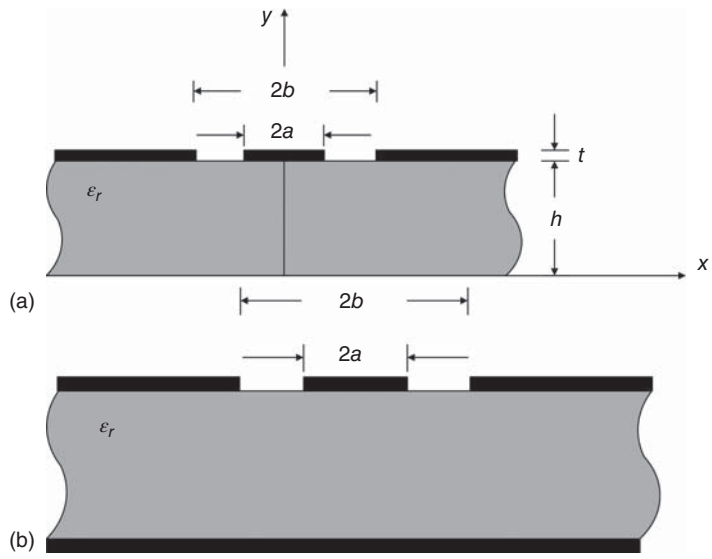


Figure 4.20. Cross-section of conventional (a) and conductor-backed (b) CPW.

for practical applications. It is operated at narrower bandwidth as compared to the conventional CPW due to the possibility of parallel-plate waveguide mode propagation.

Closed-form expressions for the effective relative dielectric constant and characteristic impedance of CPWs with zero strip thickness and infinitely wide ground strips were derived using conformal-mapping methods [10, 11]. For the conventional CPW, they are given according to [10]:

$$\varepsilon_{\text{eff}} = 1 + \frac{\varepsilon_r - 1}{2} \frac{K(k')}{K(k)} \frac{K(k_1)}{K(k_1')} \tag{4.276}$$

and

$$Z_o = \frac{30\pi}{\sqrt{\varepsilon_{\text{eff}}}} \frac{K(k')}{K(k)} \tag{4.277}$$

where

$$k = \frac{a}{b} \tag{4.278}$$

$$k' = \sqrt{1 - k^2} \tag{4.279}$$

$$k_1 = \frac{\sinh(\pi a/2h)}{\sinh(\pi b/2h)} \tag{4.280}$$

$$k_1' = \sqrt{1 - k_1^2} \tag{4.281}$$

K represents the complete integral of the first kind, whose values can be determined from an integral or from tabulated tables. The $K(k)/K(k')$ ratio can also be approximately obtained by [12]

$$\frac{K(k)}{K(k')} = \begin{cases} \dfrac{\pi}{\ell n\left(2\frac{1+\sqrt{k'}}{1-\sqrt{k'}}\right)}, & 0 \le k \le 0.707 \\ \dfrac{1}{\pi}\ell n\left(2\dfrac{1+\sqrt{k}}{1-\sqrt{k}}\right), & 0.707 \le k \le 1 \end{cases} \tag{4.282}$$

For the conductor-backed CPW, the effective dielectric constant and characteristic impedance can be evaluated from [11]

$$\varepsilon_{\text{eff}} = \frac{1 + \varepsilon_r \frac{K(k')}{K(k)} \frac{K(k_1)}{K(k_1')}}{1 + \frac{K(k')}{K(k)} \frac{K(k_1)}{K(k_1')}} \tag{4.283}$$

and

$$Z_o = \frac{60\pi}{\sqrt{\varepsilon_{\text{eff}}}} \frac{1}{\frac{K(k)}{K(k')} + \frac{K(k_1)}{K(k_1')}} \tag{4.284}$$

where

$$k = \frac{a}{b} \tag{4.285}$$

$$k' = \sqrt{1 - k^2} \tag{4.286}$$

$$k_1 = \frac{\tanh(\pi a/2h)}{\tanh(\pi b/2h)} \tag{4.287}$$

$$k_1' = \sqrt{1 - k_1^2} \tag{4.288}$$

When the thickness t of the central strip and ground strips is considered, the widths of the strip and gaps effectively increase and decrease, respectively. The effective dielectric constant and characteristic impedance of the (conventional) CPW can be found, taking into account this effect, using the following formulas [13]:

$$\varepsilon_{\text{reff}}(t) = \varepsilon_{\text{reff}} - \frac{0.7(\varepsilon_{\text{reff}} - 1)\frac{t}{b-a}}{\frac{K(k)}{K(k')} + 0.7\frac{t}{b-a}} \tag{4.289}$$

where $\varepsilon_{\text{reff}}$ is the effective relative dielectric constant when $t = 0$ given in (4.276), and

$$Z_o = \frac{30\pi}{\sqrt{\varepsilon_{\text{reff}}(t)}} \frac{K(k_e')}{K(k_e)} \tag{4.290}$$

where

$$k_e = \frac{S_e}{S_e + 2W_e} \tag{4.291}$$

$$k_e' = \sqrt{1 - k_e^2} \tag{4.292}$$

$$S_e = 2a + \Delta \tag{4.293}$$

$$W_e = b - a - \Delta \tag{4.294}$$

$$\Delta = \frac{1.25t}{\pi}\left[1 + \ell n\left(\frac{8\pi a}{t}\right)\right] \tag{4.295}$$

The conductor attenuation constant for the conventional CPW is given as [14]:

$$\alpha_c = \frac{4.88 \times 10^{-4}}{\pi} R_s \varepsilon_{\text{reff}} Z_0 P \frac{b + a}{(b - a)^2} \left\{ \frac{\frac{1.25t}{\pi} \ell n\left(\frac{8\pi a}{t}\right) + 1 + \frac{1.25t}{2\pi a}}{\left[2 + \frac{2a}{b-a} - \frac{1.25t}{\pi(b-a)}\left(1 + \ell n \frac{8\pi a}{t}\right)\right]^2} \right\} \text{ (dB/m)} \tag{4.296}$$

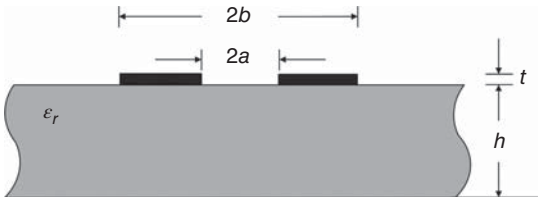


Figure 4.21. Cross-section of coplanar strips.

where

$$P = \begin{cases} \dfrac{k}{(1-k')\,(k')^{3/2}} \left[\dfrac{K(k)}{K(k')}\right]^2, & 0 \le k \le 0.707 \\ \dfrac{1}{(1-k)\sqrt{k}}, & 0.707 \le k \le 1 \end{cases} \tag{4.297}$$

The expression for the dielectric attenuation constant is the same as that for the microstrip line, which is given in (4.275).

4.11.3 Coplanar Strips

CPSs, as shown in Figure 4.21, employ two parallel strips on the same side of a dielectric substrate. The CPS structure is also useful for RFICs, especially balanced circuits due to its inherent balance nature. The CPS structure allows easy connections for series and shunt solid-state devices. The effective dielectric constant and characteristic impedance of the CPS for zero strip thickness can be evaluated using the following closed-form equations derived from a conformal-mapping method [10]:

$$\varepsilon_{\text{eff}} = 1 + \frac{\varepsilon_r - 1}{2} \frac{K(k')}{K(k)} \frac{K(k_1)}{K(k_1')} \tag{4.298}$$

and

$$Z_o = \frac{120\pi}{\sqrt{\varepsilon_{\text{eff}}}} \frac{K(k)}{K(k')} \tag{4.299}$$

where

$$k = \frac{a}{b} \tag{4.300}$$

$$k' = \sqrt{1-k^2} \tag{4.301}$$

$$k_1 = \frac{\sinh(\pi a/2h)}{\sinh(\pi b/2h)} \tag{4.302}$$

$$k_1' = \sqrt{1-k_1^2} \tag{4.303}$$

When the strip thickness is considered, the effective dielectric constant and characteristic impedance can be determined using [13]

$$\varepsilon_{\text{reff}}(t) = \varepsilon_{\text{eff}} - \frac{0.7(\varepsilon_{\text{reff}} - 1)\frac{t}{a}}{\frac{K(k')}{K(k)} + 0.7\frac{t}{a}} \tag{4.304}$$

and

$$Z_o(t) = \frac{120\pi}{\sqrt{\varepsilon_{\text{reff}}(t)}} \frac{\mathrm{K}(k_e)}{\mathrm{K}(k_e')} \tag{4.305}$$

where

$$k_e = \frac{S_e}{S_e + 2W_e} \tag{4.306}$$

$$k_e' = \sqrt{1 - k_e^2} \tag{4.307}$$

$$S_e = S - \Delta \tag{4.308}$$

$$W_e = W + \Delta \tag{4.309}$$

$$\Delta = \frac{1.25t}{\pi}\left[1 + \ell n\left(4\pi\frac{b-a}{t}\right)\right] \tag{4.310}$$

The attenuation due to the conductor loss can be found from [14]

$$\alpha_c = \frac{4.34}{\pi}\frac{R_s}{Z_o}P\frac{a+b}{a^2}\left\{\frac{\frac{1.25}{\pi}\ell n\left(4\pi\frac{b-a}{t}\right) + 1 + \frac{1.25}{\pi}\frac{t}{b-a}}{\left\{1 + \frac{b-a}{a} + \frac{1.25}{2\pi}\frac{t}{a}\left[1 + \ell n\left(4\pi\frac{b-a}{t}\right)\right]\right\}^2}\right\} \tag{4.311}$$

where P is given in (4.297). The dielectric attenuation constant can be determined from (4.275).

4.11.4 Strip Line

Strip line, as shown in Figure 4.22, is essentially a printed-circuit version of the coaxial transmission line. Its dominant mode of propagation is pure TEM, assuming perfect conductors. Because of possible excitement of unwanted parallel-plate modes at high frequencies due to the two parallel ground planes, the strip line is in general more suitable for use in low RF regimes. Its characteristic impedance for zero-thickness strip can be determined using the following conformal-mapping formulas [15]:

$$Z_o = \frac{30\pi}{\sqrt{\varepsilon_r}}\frac{K(k\prime)}{K(k)} \tag{4.312}$$

where

$$k = \tanh\left(\frac{\pi W}{4a}\right) \tag{4.313}$$

$$k' = \sqrt{1 - k^2} \tag{4.314}$$

For finite strip thickness, the characteristic impedance can be found from [16]

$$Z_o = \frac{30}{\sqrt{\varepsilon_r}}\ell n\left\{1 + \frac{4}{\pi}\frac{2a-t}{W_e}\left[\frac{8}{\pi}\frac{2a-t}{W_e} + \sqrt{\left(\frac{8}{\pi}\frac{2a-t}{W_e}\right)^2 + 6.27}\right]\right\} \tag{4.315}$$

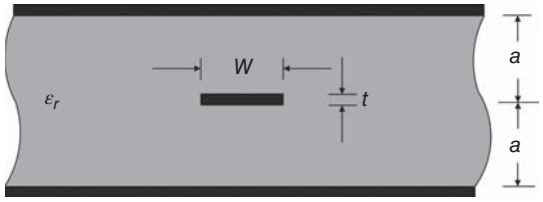


Figure 4.22. Cross-section of strip line.

where

$$\frac{W_e}{2a-t} = \frac{W}{2a-t} + \frac{\Delta W}{2a-t} \tag{4.316}$$

$$\frac{\Delta W}{2a-t} = \frac{x}{\pi(1-x)}\left\{1 - \frac{1}{2}\ell n\left[\left(\frac{x}{2-x}\right)^2 + \left(\frac{0.0796x}{\frac{W}{2a}+1.1x}\right)^m\right]\right\} \tag{4.317}$$

$$m = \frac{2}{1+\frac{2x}{3(1-x)}} \tag{4.318}$$

$$x = \frac{t}{2a} \tag{4.319}$$

The strip width can also be determined from the characteristic impedance and relative dielectric constant. For a zero-thickness strip, the expression for the strip width can be derived from (4.312)–(4.314) as

$$\frac{W}{a} = \frac{4}{\pi}\tanh^{-1}\sqrt{p} \tag{4.320}$$

where

$$p = \begin{cases} \sqrt{1-\left[\dfrac{e^{\pi q-2}}{e^{\pi q+2}}\right]^4}, & q \geq 1 \\ \left[\dfrac{e^{\pi/q}-2}{e^{\pi/q}+2}\right]^2, & 0 \leq q \leq 1 \end{cases} \tag{4.321}$$

$$q = \frac{Z_o\sqrt{\varepsilon_r}}{30\pi} \tag{4.322}$$

Using (4.315)–(4.319), we can drive the following expression for the strip width, taking into account the strip thickness:

$$\frac{W}{2a-t} = \frac{W_e}{2a-t} - \frac{\Delta W}{2a-t} \tag{4.323}$$

where

$$\frac{W_e}{2a-t} = \frac{8}{\pi A} \tag{4.324}$$

$$\frac{\Delta W}{2a-t} = \frac{x}{\pi(1-x)}\left\{1 - \frac{1}{2}\ell n\left[\left(\frac{x}{2-x}\right)^2 + \left(\frac{0.0796x}{\frac{W_e}{2a}+1.1x}\right)^m\right]\right\} \tag{4.325}$$

$$A = \frac{2B}{C} \tag{4.326}$$

$$B = \exp\left(\frac{Z_o\sqrt{\varepsilon_r}}{30}\right) - 1 \tag{4.327}$$

$$C = \sqrt{4B+6.27} \tag{4.328}$$

$$m = \frac{2}{1+\frac{2x}{3(1-x)}} \tag{4.329}$$

$$x = \frac{t}{2a} \tag{4.330}$$

The conductor attenuation constant (dB/m) can be determined from

$$\alpha_c = \begin{cases} \dfrac{23.4 \times 10^{-3} R_s \varepsilon_r Z_o A}{30\pi\,(2a-t)}, & Z_o < \dfrac{120}{\sqrt{\varepsilon_r}} \\ \dfrac{1.4 R_s B}{2 Z_o a}, & Z_o > \dfrac{120}{\sqrt{\varepsilon_r}} \end{cases} \tag{4.331}$$

where

$$A = 1 + \frac{2W}{2a-t} + \frac{1}{\pi}\frac{2a+t}{2a-t}\ell n\left(\frac{4a-t}{t}\right) \tag{4.332}$$

$$B = 1 + \frac{2a}{0.5W + 0.7t}\left(0.5 + \frac{0.414t}{W} + \frac{1}{2\pi}\ell n\frac{4\pi W}{t}\right) \tag{4.333}$$

The dielectric attenuation constant is given by

$$\alpha_d = \frac{27.3\sqrt{\varepsilon_r}\tan\delta}{\lambda_o} \quad \text{(dB/m)} \tag{4.334}$$

4.11.5 Slot Line

The slot line, whose cross section is shown in Figure 4.23, is also useful for RFICs. Its balance nature is especially attractive for circuits requiring balance topology. It is noted that, in contrast with conventional transmission lines, the two conducting strips forming the slot are kept at equal potential (typically grounded) and the slot line does not support a TEM or quasi-TEM propagation mode. Modes propagating on the slot line are quasi-TE modes, which resemble the TE type. The dominant mode is quasi-TE_{10}, similar to the TE_{10} of the rectangular waveguides. The quasi-TE_{10} mode of the slot line, however, has no cutoff frequency. Due to the propagating mode being quasi-TE_{10}, when the slot line is used together with other transmission lines supporting TEM or quasi-TEM mode in circuits, suitable transitions between them need to be used to allow proper transformation of the fields of the slot line into those of the other transmission lines. In order to be used as a good transmission line, the slot line should be fabricated using a substrate with a high dielectric constant to minimize radiation. On the other hand, an antenna using the slot line should be implemented using a low dielectric constant substrate to maximize radiation.

Closed-form expressions for the characteristic impedance, Z_o, based on voltage and power and the wavelength, λ_g, for the slot line on high dielectric constant substrates ($9.7 \le \varepsilon_r \le 20$) were obtained [17] by curve

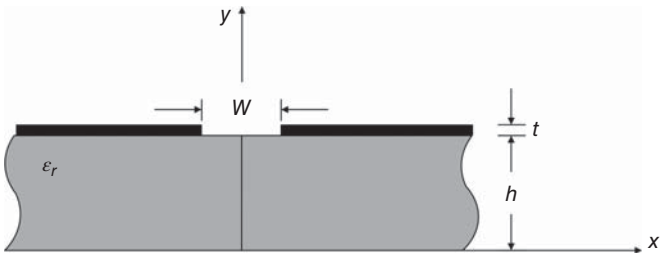


Figure 4.23. Cross-section of slot line.

fitting to the numerical results based on [18]. They are given as:

$$\frac{\lambda_g}{\lambda_o} = \begin{cases} 0.923 - 0.195\ell n\varepsilon_r + 0.2\dfrac{W}{h} - \left(0.126\dfrac{W}{h} + 0.02\right)\ell n\left(\dfrac{h}{\lambda_o}\times 10^2\right), & 0.02 \le \dfrac{W}{h} \le 0.2 \\ 0.987 - 0.21\ell n\varepsilon_r + \dfrac{W}{h}(0.111 - 0.0022\varepsilon_r) & \\ \quad - \left(0.053 + 0.041\dfrac{W}{h} - 0.0014\varepsilon_r\right)\ell n\left(\dfrac{h}{\lambda_o}\times 10^2\right), & 0.2 \le \dfrac{W}{h} \le 1 \end{cases} \tag{4.335}$$

$$Z_o = \begin{cases} 72.62 - 15.283\ell n\varepsilon_r + 50\dfrac{\left(\dfrac{W}{h} - 0.02\right)\left(\dfrac{W}{h} - 0.1\right)}{W/h} + \ln\left(\dfrac{W}{h}\times 10^2\right)[19.23 - 3.693ln\varepsilon_r] & \\ \quad - \left[0.139\ell n\varepsilon_r - 0.11 + \dfrac{W}{h}(0.465\ell n\varepsilon_r + 1.44)\right]\left(11.4 - 2.636\ell n\varepsilon_r - \dfrac{h}{\lambda_o}\times 10^2\right)^2 & 0.02 \le \dfrac{W}{h} \le 0.2 \\ 113.19 - 23.257\ell n\varepsilon_r + 1.25\dfrac{W}{h}(114.59 - 22.531\ell n\varepsilon_r) + 20\left(\dfrac{W}{h} - 0.2\right)\left(1 - \dfrac{W}{h}\right) & \\ \quad - \left[0.15 + 0.1\ell n\varepsilon_r + \dfrac{W}{h}(-0.79 + 0.899\ell n\varepsilon_r)\right] & 0.2 \le \dfrac{W}{h} \le 1 \\ \quad \times\left[10.25 - 2.171\ell n\varepsilon_r + \dfrac{W}{h}(2.1 - 0.617\ell n\varepsilon_r) - \dfrac{h}{\lambda_o}\times 10^2\right]^2, & \end{cases} \tag{4.336}$$

These equations were obtained assuming infinitesimally thin conductors and are accurate to within 2% for the following ranges:

$$9.7 \le \varepsilon_r \le 20 \tag{4.337}$$

$$0.01 \le \frac{h}{\lambda_o} \le \left(\frac{h}{\lambda_o}\right)_c \tag{4.338}$$

where $(h/\lambda_o)_c$ is the cutoff value for the TE_{10} surface-wave mode on the slot line, and is given as

$$\left(\frac{h}{\lambda_o}\right)_c = 0.25\sqrt{\varepsilon_r - 1} \tag{4.339}$$

Closed-form expressions for the characteristic impedance and wavelength for low dielectric constant substrates were also derived by curve fitting to results of the spectral domain method [19].

4.11.6 Field Distributions

The behavior of a transmission line and the signal propagating on it, governed by the power $\vec{P} = 1/2\ Re(\vec{E}\times\vec{H}^*)$, are dictated by the transmission line's electric and magnetic fields, which affect circuit performance at all frequencies, particularly more pronounced at high frequencies. Therefore, in RFIC design, it is important and useful to understand the field distributions of printed-circuit transmission lines used in RFICs. As signals propagate along a transmission line, they carry with them the electric and magnetic fields and the signals may couple from the transmission line to other circuit elements including transmission lines. The coupling between a transmission line and other transmission lines or circuit elements is caused by

the springing of the transmission line's electric and/or magnetic fields onto other lines or circuit elements. There are two kinds of coupling between transmission lines (or between a transmission line and other circuit elements): one is the electric coupling caused by the electric fields between the transmission lines and another is the magnetic coupling caused by the magnetic fields. The field distribution of a transmission line thus helps RFIC designers to visualize and comprehend the effects of coupling from one transmission line to another, effectively enabling them to optimize circuit layout while considering possible effects on circuit performance. Another important utilization of the field distribution is for the matching between the fields of different kinds of transmission lines – for example, between a microstrip line and CPW – when they are connected. From the circuit theory point of view, two interconnected transmission lines only need to have equal characteristic impedances. This, however, is not sufficient considering EM effects; these equal-characteristic-impedance transmission lines should also have their electric and magnetic fields matched to each other to minimize possible reflections due to unmatched fields at the interconnecting junction. In general, a transition between these lines should be used to allow the fields of one line to be transformed into those of the other. Figure 4.24 illustrates the field distributions of the printed-circuit transmission lines shown in Figure 4.18. From these field distributions, we can see that the width of the ground strip, whether or not in the same plane of the central strip, does not have to be large; only sufficient width is needed to contain all or majority of the electric fields emanating from the central strip. This field-confinement is particularly critical in RFICs as the electric fields not terminated by the ground plane may enter the (lossy) silicon substrate, causing significant dielectric loss.

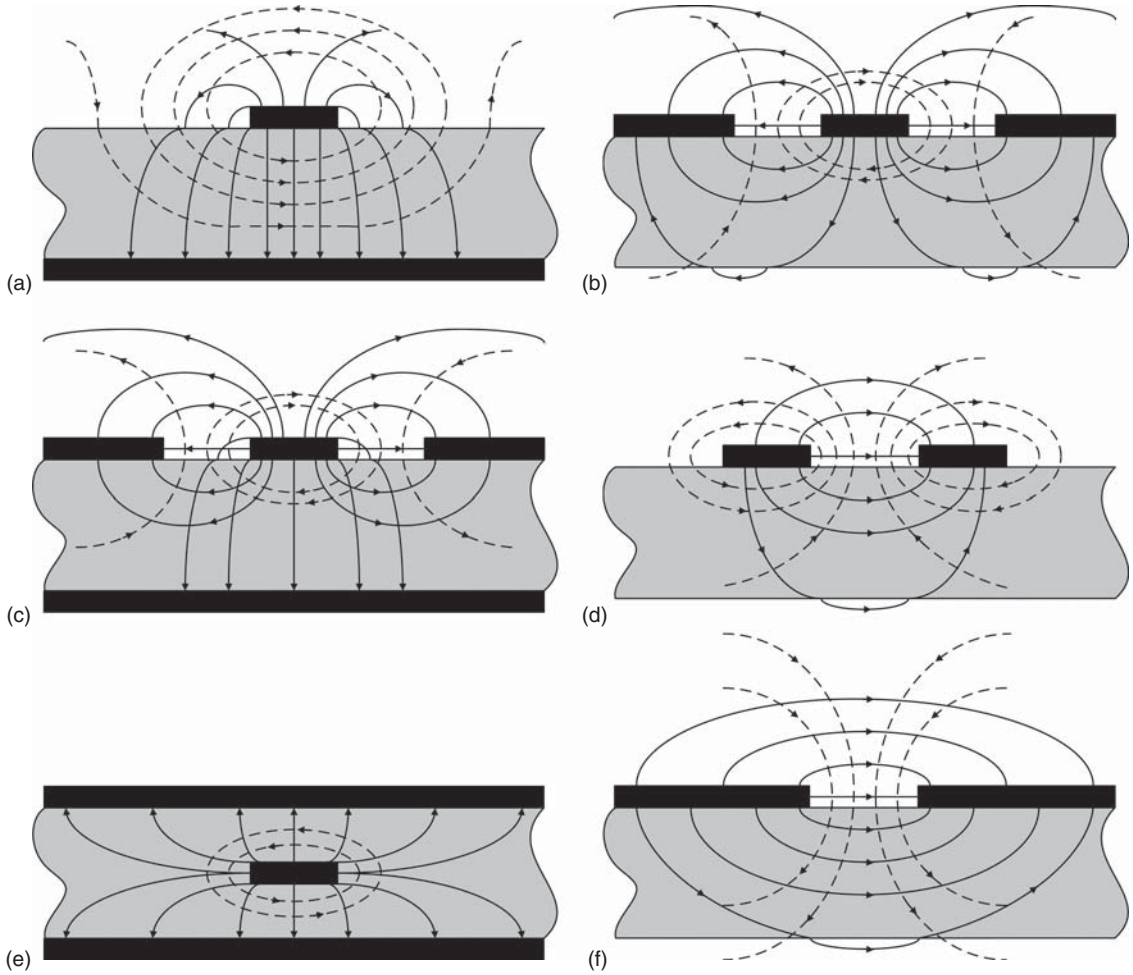


Figure 4.24. Field distributions of microstrip line (a), CPW (b), conductor-backed CPW (c), CPS (d), strip line (e) and slot line (f). Electric field (solid lines) and magnetic field (dashed lines).

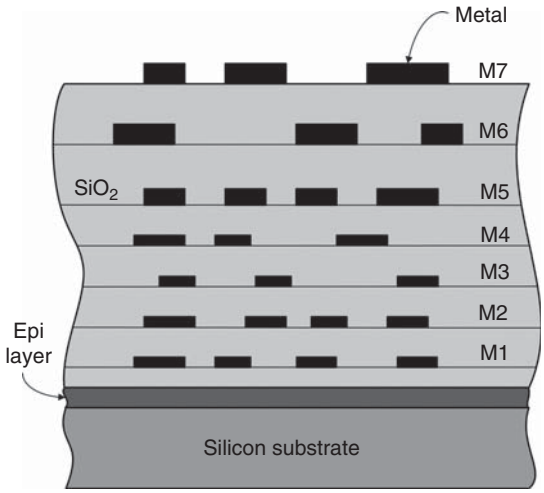


Figure 4.25. Typical multilayer metal-dielectric CMOS structure (seven metal layers are assumed here.)

4.12 TRANSMISSION LINES IN RFICS

Figure 4.25 shows the cross section of a typical CMOS dielectric-metal structure used in RFICs, which consists of multiple metal layers embedded within multilayer dielectric above a silicon substrate. The numbers of metal and dielectric layers vary depending on CMOS processes. The dielectrics are normally SiO_2 with relative dielectric constant of 3.9 and typical layer thickness in the 1-μm range depending on process. The top-most metal layer (e.g., M7 as shown in Figure 4.25) has the thickest metallization (in the range of 3 μm depending on process) and the other metal layers have thinner metallization (about 0.7 μm depending on process.) The thickest top-most metal layer produces the lowest conductor loss among the metal layers and is thus preferred for transmission lines. Due to the availability of multiple conductors and dielectrics as seen in Figure 4.25, together with the use of air-bridges and via-holes, various two-conductor and multiconductor printed-circuit transmission lines can be formed for RFICs. Such transmission lines facilitate the realization of complicated circuits and, along with the ability of achieving both horizontal and vertical integration through multilayer, ultimately allow very compact, high density circuit integration. The inherently thin SiO_2 dielectric layers can further reduce circuit size by achieving narrow line widths. The two fundamentals that should be considered in configuring a transmission line for RFICs are that the characteristic impedance range and the loss should be as wide and low as possible, respectively. It is noted that typical silicon substrates used in RFICs have relative dielectric constant of around 11.7 and particularly low resistivity of about 10^3–10^5 Ω-cm[14] which results in substantial dielectric loss, especially in the RF range. Therefore, transmission lines in RFIC structures should be completely, or at least partially, isolated or shielded from the silicon substrate to reduce loss. Figures 4.26–4.30 show some possible printed-circuit transmission lines with visual of the electric field lines that can be implemented for RFICs. These transmission lines exploit the advantages of multilayer structures as well as the couplings between the constituent strips and slots to improve performance. Using a back conductor with and without slot, they can have wider characteristic-impedance ranges than those of the conventional microstrip line and CPW. In addition, they have several other desirable features such as flexibility and ability to realize complicated, highly dense circuits through appropriate arrangements of the structural elements. The analysis of these transmission lines can be accurately analyzed using full-wave dynamic approaches. Various commercially available EM simulation programs can also be used for accurate calculations of the transmission lines' parameters.

[14]For comparison, typical GaAs substrates used in microwave monolithic integrated circuits (MMIC) have about 10^7–10^9 Ω-cm resistivity and, hence, much lower dielectric loss as compared to silicon.

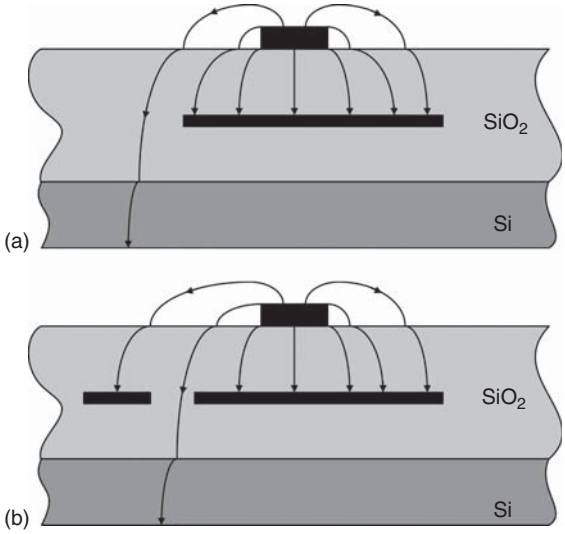


Figure 4.26. (a, b) Representative microstrip lines in RFICs.

4.12.1 Microstrip Line

Figure 4.26 shows two examples of microstrip lines in RFICs. The slot in the bottom conductors in Figure 4.26(b) facilitates adjustment of the characteristic impedance by varying its location and/or size. The bottom conductors are kept at approximately equal potential using air-bridges. The top conductor can be located in any metal layer, typically the top-most metal layer (e.g., M7 in Figure 4.25) for the lowest conductor loss, and the bottom conductor in another metal layer (e.g., M2) depending on desired characteristics such as characteristic impedance, loss, effective relative dielectric constant, dispersion, etc. These microstrip lines allow easy interface with (external) off-chip components and/or instruments for measurement purposes. The bottom conductor in Figure 4.26(a) and the slot in Figure 4.26(b) may be symmetrical or asymmetrical with respect to the top conductor. These bottom conductors should be sufficiently wide and the slot should be properly located with respect to the top conductor to prevent all or majority of the electric fields coming from the top conductor from penetrating into the silicon substrate in order to eliminate or minimize the effects of the silicon substrate on the transmission line, respectively. The complete or partial isolation of the microstrip line from the silicon substrate results in low dielectric loss due to the use of the SiO_2 dielectrics. It also eliminates or reduces the substrate noise propagating into the microstrip line and the coupling between the microstrip line and other nearby transmission lines and circuit elements through the silicon substrate. Since SiO_2 has low loss, multilayer SiO_2 can be used to increase the dielectric thickness to produce wider strips for given characteristic impedances, which result in lower conductor loss. The highest characteristic impedance of the microstrip line shown in Figure 4.26(a) is limited due to the large capacitance per unit length resulting from the bottom conductor. Its characteristic impedance, effective relative dielectric constant and loss can be approximately determined using the closed-form expressions presented in Section 4.11.1 . The microstrip line shown in Figure 4.26(b) is particularly attractive since its characteristic impedance can be varied over a large range by simply changing the location and size of the slot between the bottom conductors. This unique tuning feature can be utilized to realize a larger impedance range than that of the microstrip line shown in Figure 4.26(a). When the slot is far away from the top conductor, the characteristic impedance and effective relative dielectric constant are constant since the slot has negligible influence on the transmission line's characteristics. However, as the slot approaches the top conductor, both the characteristic impedance and effective dielectric constant increase because more of the electric field lines penetrate the silicon substrate whose relative dielectric constant is higher than that of the SiO_2 dielectric. These changes are more drastic for larger slot and also when the slot is close to the top conductor due to increased interaction between the silicon substrate and top conductor via the slot. The maximum characteristic impedance and effective relative dielectric constant occur when the slot is centered underneath the top conductor, where maximum

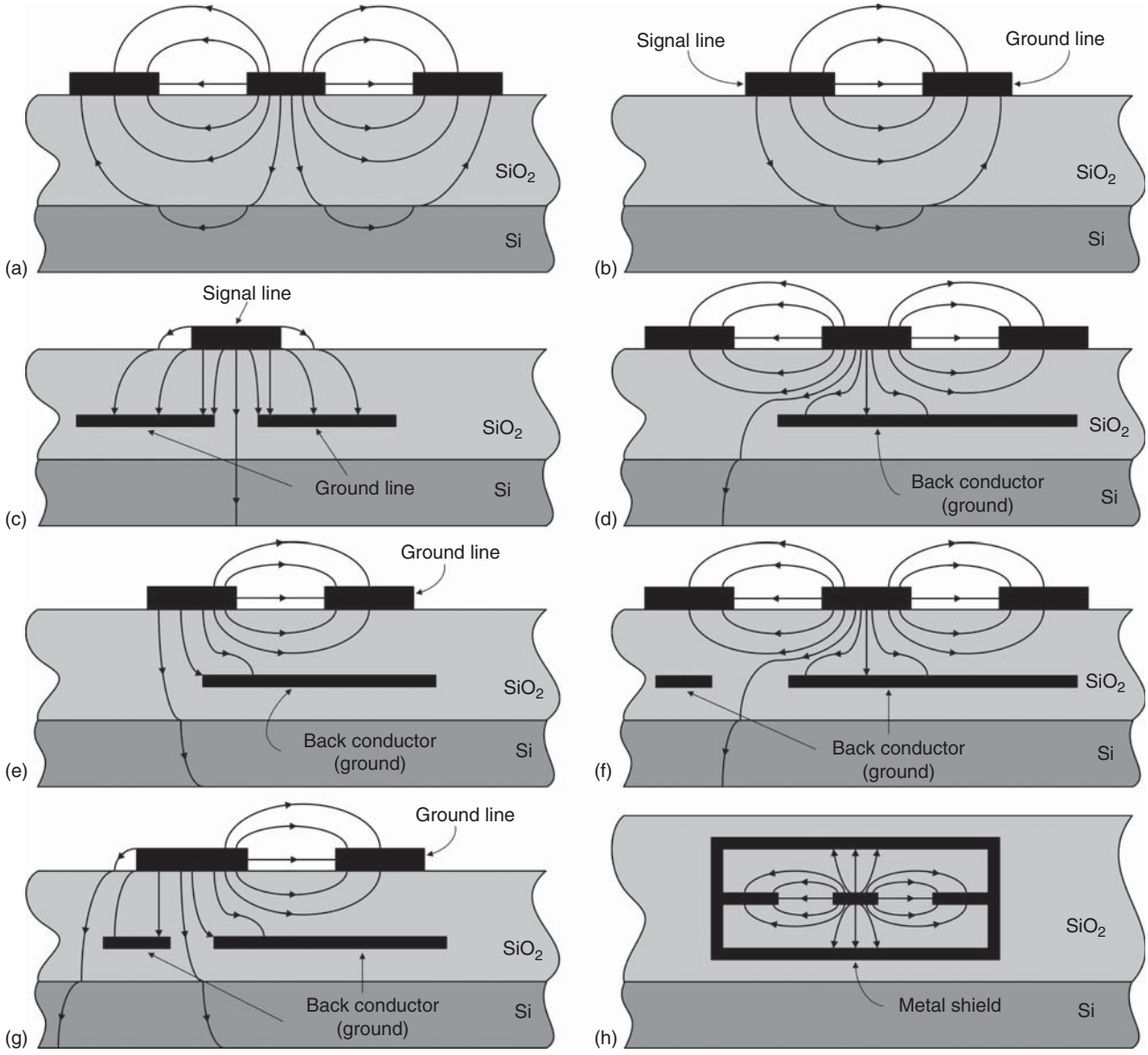


Figure 4.27. Possible CPWs in RFICs: (a) CPW, (b) single-ground CPW, (c) non-coplanar CPW, (d) conductor-backed CPW, (e) single-ground conductor-backed CPW, (f) slot-conductor-backed CPW, (g) single-ground slot-conductor-backed CPW, and (h) shielded CPW. Back conductors and ground lines are kept at equal potential through via-hole interconnects. Back conductors separated by slot are interconnected through air-bridges to maintain the same potential.

field lines enter the bottom silicon substrate. While achieving higher characteristic impedance and hence larger impedance range is desirable, more electric fields penetrating the silicon substrate cause larger dielectric loss and, hence, detrimental circuit effects. Therefore, care needs to be exercised when optimizing the characteristic-impedance range using the slot.

4.12.2 Coplanar Waveguide

Figure 4.27 shows several CPWs that can be implemented in RFICs. Any metal layers can be used; however, like the microstrip line or any other transmission lines, the top-most metal and those close to it should be employed. In general, the CPW has wider characteristic impedance range than the microstrip line due to the possibility of obtaining higher characteristic impedances through using large gaps between the signal and ground lines. The CPW is, in general, has larger size than the microstrip line. However, narrow ground lines

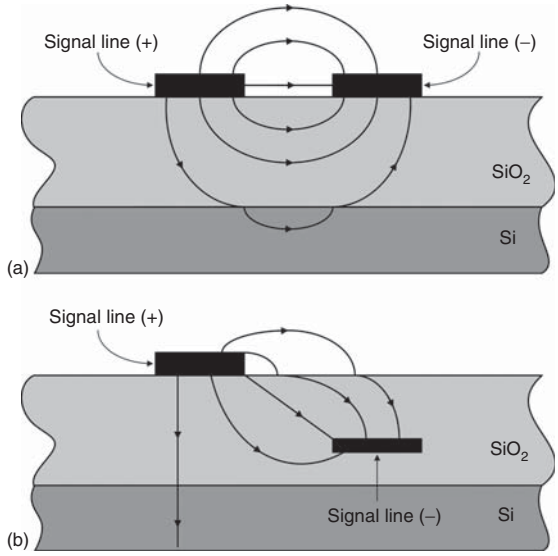


Figure 4.28. Possible CPS in RFICs: (a) CPS and (b) noncoplanar CPS.

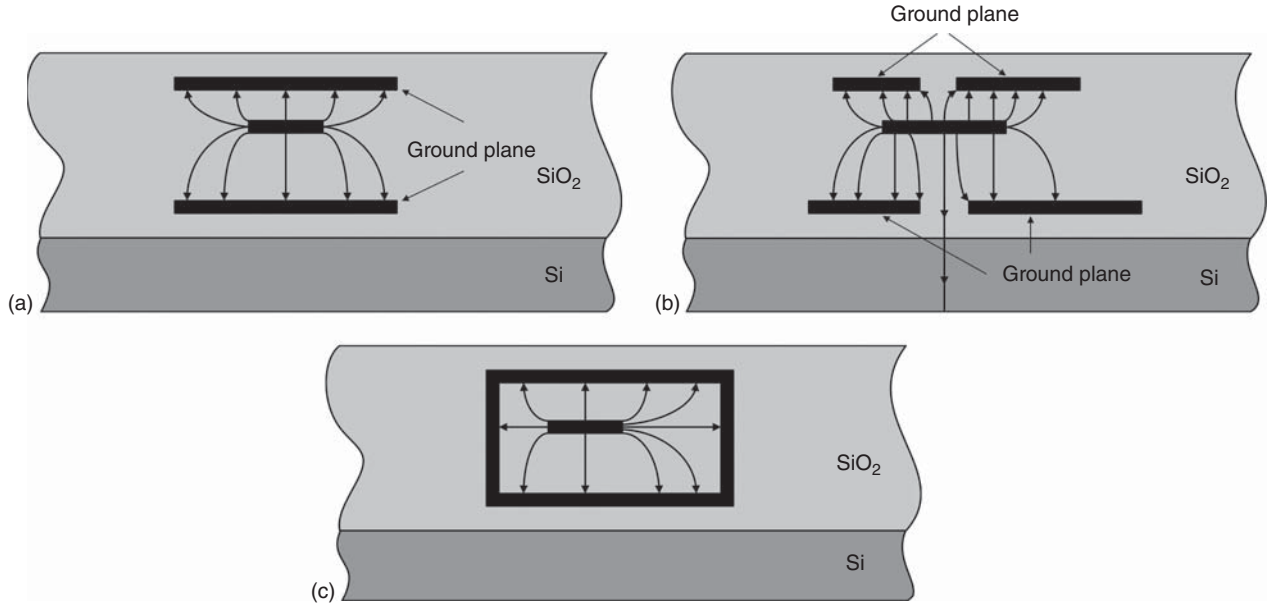


Figure 4.29. Possible strip line in RFICs: (a) strip line, (b) slitting-ground strip line, and (c) shielded strip line.

can be used to reduce the size while not causing significant effects to the line's characteristic impedance and effective relative dielectric constant since the charge and current are distributed primarily along the edges of the strip and the inner edges of all the ground lines. A single ground line can also be used to reduce the size further. Unequal ground lines can be employed to accommodate circuit layout. As for the microstrip line, lower loss can be achieved by using thicker dielectric corresponding to wider signal line. All the ground lines and back conductors need to be tied together through via-holes and air bridges to maintain at approximately equal potential. A back conductor, when connected to the ground lines, also help increase the isolation between the signal line and adjacent transmission lines since the electric fields springing away from the signal line tend to terminate in the back conductor via the ground lines. In RFICs with very small size constraint, which causes significant interactions between elements, CPW with enhanced isolation is indeed desirable. On the other hand, when the back conductor is not interconnected with the ground lines, possibly strong

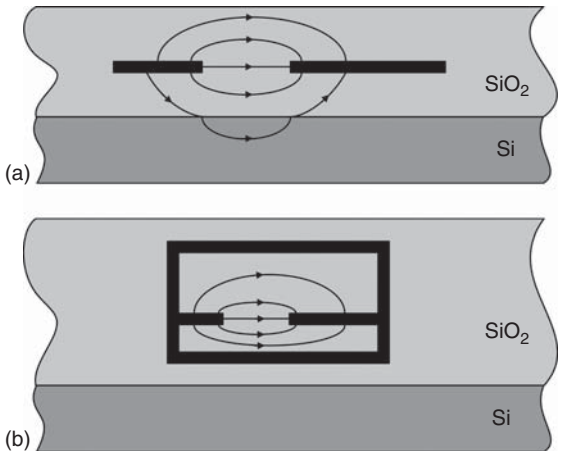


Figure 4.30. (a, b) Slot lines in RFICs.

coupling from the signal line to adjacent lines can result as the ground lines may serve as a bridge between them. As will be discussed later, the use of a back conductor with slot allows the characteristic impedance to be varied over a large range by simply changing the location and size of the slot. This unique tuning feature, in conjunction with the signal line and upper slots, can be utilized to realize much larger impedance ranges than those of the conventional microstrip line and CPW.

Figure 4.27(a) shows the conventional CPW, Figure 4.27(b) shows a modified CPW with single ground line, and Figure 4.27(c) depicts a CPW with the ground lines located in another layer. As can be seen from their electric field distributions, the CPWs in Figure 4.27(a) and (b) are partially shielded from the silicon substrate, which result in low dielectric loss and reduced substrate noise from the silicon and coupling between them and other nearby transmission lines and circuit elements through the silicon substrate. The degree of isolation from the silicon substrate depends on how far the signal line of the CPWs is located from the silicon. Complete isolation may be possible if the CPWs are located far from the silicon substrate and the gaps between the signal and ground lines are sufficiently small to cause the electric fields to focus more within the gap areas. It is noted that the single ground-line CPW, while physically similar to the CPS, is electrically different since its ground line is held at zero potential while, in the CPS, this line is kept at an opposite potential of the other strip. The CPW in Figure 4.27(c) is similar to the microstrip line shown in Figure 4.26(b) and, as expected from its electric field lines, can be fully shielded from the silicon substrate even located near it if the gap between the ground lines is narrow, especially at higher frequencies of the RF range since the gap becomes negligible with respect to the operating wavelengths. As the relative dielectric constant of the SiO_2 ($\varepsilon_r = 3.9$) is much smaller than that of the silicon substrate ($\varepsilon_r = 11.7$), the confinement of the fields within the SiO_2 is weakest when the transmission line is farthest from the silicon substrate, hence resulting in largest characteristic impedance.

Figure 4.27(d) and (e) shows the CPWs with a back conductor. Depending on the width and/or location of the back conductor, complete or partial isolation from the silicon substrate can be obtained. For these CPWs, while the characteristic impedance and effective relative dielectric constant stay constant when the edge of the back conductor closer to the slots is far away from the signal line, they decrease as the edge advances near the signal line. As the back conductor increases in width toward the slots, the beneath dielectrics vary from a combination of SiO_2 and Si to SiO_2, thus the characteristic impedance and the effective relative dielectric constant decrease because more of the field lines remain in SiO_2 as the back conductor gets larger.

The CPWs in Figure 4.27(f) and (g) implement a slitting back-conductor whose size and/or location of the slot and conductor control the degree of shielding from the silicon substrate. It is noted that the signals see the gap in the back conductor in terms of wavelength and so, at very high frequencies, the gap may appear opaque to the signals and hence the transmission line is completely shielded from the silicon substrate. In addition to the upper slot, the lower slot formed by the back conductor can also be used to change the characteristic impedance and effective relative dielectric constant. These parameters increase as the width of the lower slot is increased. The change is significantly pronounced when the width is small, while virtually unnoticed

for large widths. This phenomenon is expected as the transmission line's capacitance per unit length reduces when the lower-slot width is increased due to more electric fields penetrating into the bottom layers. Furthermore, the range of the lower-slot width that affects the behavior is smaller for smaller upper-slot width, as expected, due to strong interaction between the signal line and the coplanar ground strips when the upper slots are small. Similar to the microstrip line with slot in the bottom conductor as shown in Figure 4.26(b), when the slot in the back conductor is far away from the signal line, the characteristic impedance and effective relative dielectric constant are constant. However, as the slot approaches the signal line, both the characteristic impedance and effective dielectric constant increase. These changes are more drastic for larger lower slot and also when the slot is close to the signal line. The maximum characteristic impedance and effective relative dielectric constant occur when the slot is centered underneath the signal line.

Figure 4.27(h) shows a CPW enclosed by conducting walls that completely shield the transmission line from the silicon substrate and surrounding elements. The vertical walls are formed by lattices of periodic metallic via-holes placed very close to each other, similar to the side-walls used in the rectangular resonator discussed in the Section titled An Example of CMOS Rectangular Cavity Resonators of Chapter 5. Narrow metal frames on the metal layers between the top and bottom walls are also used to connect all the via-holes together to further confine the fields within the metal channel and enhance the electrical isolation between the signal line and the exterior of the channel. The metal shield prevents fields from entering and leaving the CPW from its left, right, upper, and lower sides, except through the signal line. The size of the via-holes and distance between them in typical RFIC fabrication processes are extremely small with respect to wavelengths in the high RF range. The via-hole lattices, together with the metal frames, can thus create vertical walls resembling well solid metallic walls electrically and hence can be used to replace them. Similar conducting walls can also be used for other CPWs with back conductors to provide complete shield to the silicon substrate and other surrounding elements.

4.12.3 Coplanar Strips

Figure 4.28(a) and (b) shows the conventional CPSs and another version with strips located on two different metal layers, respectively, both consisting of positive (+) and negative (−) signal lines. The top-most metal layer or those nearest to it are preferred for the strips. The widths of the strips are equal in typical implementations. However, unequal widths can also be used to optimize performance. Similar to the conventional CPW, the CPS is partially shielded from the silicon substrate. The coupling to the silicon substrate, and hence the influence of the silicon, depends on the distance of the strips from the silicon substrate and the gap between the CPSs or the offset between the noncoplanar strips. Under certain conditions, the coupling can be significant, leading to substantial dielectric loss and substrate noise coupling. In general, the dielectric loss depends slightly on the width of the strips, while the conductor loss is a strong function of the strip width. The CPS has larger characteristic impedance than the microstrip line. Compared to the CPSs, the noncoplanar strips generally have a stronger electric field between the strips, which results in larger capacitance per unit length and, hence, lower characteristic impedance. Noncoplanar strips, however, can have less coupling to the silicon substrate, and hence less dielectric loss, due to possible shielding from the lower strip. The coupling is significantly reduced when the lower strip is moved closer to the upper strip. In fact, the transmission line can be completely shielded from the silicon substrate if the lower strip is directly underneath the upper strip and has sufficiently larger size than the upper strip.

4.12.4 Strip Line

Figure 4.29 shows some possible strip-line configurations for RFICs, in which the center conductor is sandwiched between two ground planes. The upper and lower ground planes do not have to be identical and the dielectrics above and below the central strip may have different thickness in order to optimize the strip line's characteristics. The strip line in Figure 4.29(b) particularly has a slit in either or both of the ground planes to allow high characteristic impedances to be obtained for certain applications. Figure 4.29(c) depicts a shielded strip line in which vertical walls are used together with horizontal walls to completely enclose the strip line.

These vertical walls can be formed using arrays of via-holes similar to those used for the shielded CPW shown in Figure 4.27(h). The strip line has a larger size and a less-convenient layout as compared to the microstrip line and CPW and, in general, operates at lower frequencies due to possible propagation of the higher-order parallel-plate waveguide modes. However, it is inherently isolated from the silicon substrate and, with proper use of thin SiO_2 dielectrics and sufficient number of via-holes between the two ground planes, can operate at high frequencies into the millimeter-wave range. Properly configured strip line is thus valuable for RFICs.

4.12.5 Slot Line

Figure 4.30(a) and (b) shows a slot line and shielded slot line, respectively. The shield is realized with horizontal metal strips and vertical metal walls formed by lattices of via-holes. It confines the fields and isolates the slot line from other structures external to the shield including the silicon substrate, other transmission lines, and circuit elements.

4.12.6 Transitions and Junctions Between Transmission Lines

The multilayer of metals and dielectrics in RFICs provide significant advantages in forming various transmission lines, enabling possible optimization of RFIC performance and layout. However, it is important to configure these transmission lines and use them properly, recognizing not only the need of impedance matching from their characteristic impedances but also the behavior of their field distribution. This is particularly important when different kinds of transmission lines are used in the same circuit. We consider two different combinations of transmission lines: one consisting of unbalanced transmission lines such as a microstrip line and CPW and another consisting of both unbalanced (e.g., CPW) and balanced (e.g., CPS) transmission lines.

When two different unbalanced transmission lines, such as a microstrip line and CPW, are used, the interconnection between these transmission lines must have equal characteristic impedance and closely matched field distributions – unmatched field distributions may cause reflection at the junction and degrade performance, particularly in the high RF range. When the fields between the transmission lines are unmatched, a transition at the junction between the lines should be provided to allow the fields of one line to transfer into those of the other line. Although proper transitions between unbalanced transmission lines should always be used for optimum performance, in practice, their use may not precluded in order to minimize the circuit size when such avoidance does not affect the circuit performance significantly. This is facilitated if the field distributions of these transmission lines are not significantly different from each other. Therefore, proper selection of transmission lines with fields as close as possible should be exercised when using them in the same circuit.

While the transitions between different unbalanced transmission lines may be avoidable, those between balanced and unbalanced transmission lines, however, are inevitable. This transition is commonly known as "balun" stemming from the words BAlanced and UNbalanced (balun). The name balun is also used to indicate the transition between single-ended and differential-ended ports. As examples, we show in Figures 4.31

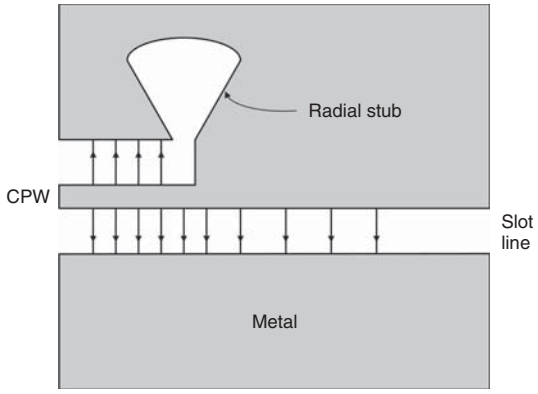


Figure 4.31. Balun between CPW and slot line with illustration of the electric fields.

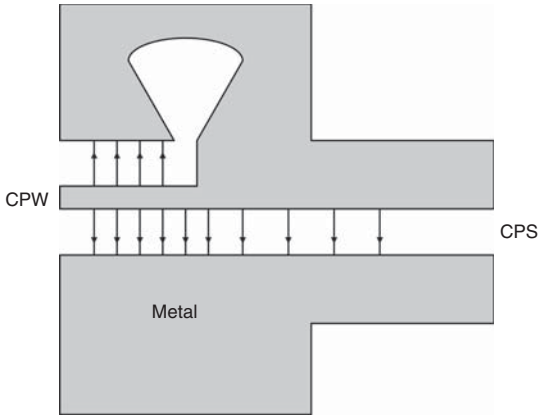


Figure 4.32. Balun between CPW and CPS with illustration of the electric fields.

and 4.32 some baluns between the (unbalanced) CPW and the (balanced) slot line and CPS, respectively [20]. The CPW-slotline and CPW–CPS transitions are configured in such way that the electric fields from the CPW make a direct transition into those of the slot line and CPS, respectively. The slot-line radial stub acts a quarter-wavelength short-circuited transmission line to present a short at the junction over a broad bandwidth. The CPW-slotline balun can operate over more than 2.3 octaves of bandwidth. The CPW–CPS transition is based on the CPW-slotline transition in which the CPW is first transformed into the slot line before making a transition into the CPS, which is facilitated by the similarity between the physical structures of the slot line and CPS. This balun can work well over a frequency range of more than 2 octaves.

Proper configurations between different transmission lines can also provide certain useful functions needed in RFICs over very wide frequency ranges. Figure 4.33 shows an UWB 180° hybrid formed by the CPW and slot line, that can be used for extremely wideband balanced circuits such as balanced mixers. It is noted that, in general, circuits such as 180° hybrids are frequency-dependent due to the dependence of their constituent elements. Extremely broadband components, however, may be realized by using elements less dependent upon frequency such as those based on the transmission lines' field distributions. The broadband hybrid described in Figure 4.33 is one of such components where the operating bandwidth is dictated mainly by the directions of the electric fields of the CPW and slot lines, which are essentially constant across an ultra-wide frequency range over which the corresponding dominant modes (e.g., quasi-TEM) propagate. As can be recognized from the electric field distributions at the junction between the slot line and CPW, a signal coming from the slot line will split equally but opposite in phase into the two slots of the CPW, thus

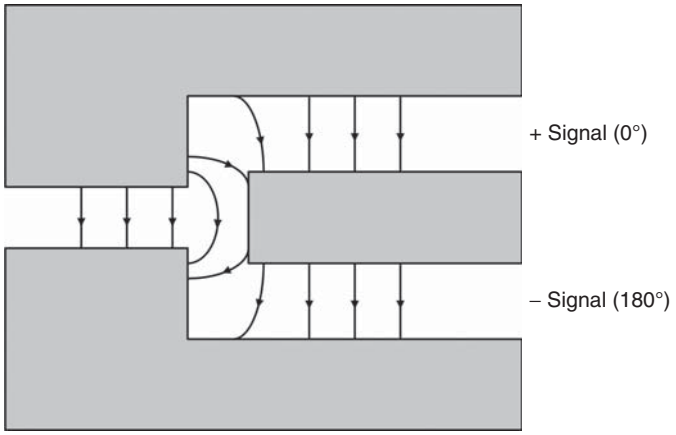


Figure 4.33. Ultra-wideband 180° hybrid created by the junction between the slot line and CPW. The upper and lower slots are at opposite phases (0° and 180°).

forming a 180° hybrid working over an extremely wide bandwidth. The EM simulation for a design of this hybrid shows amplitude and phase imbalances of almost zero and 180°, respectively, from 0.1 MHz up to about 25 GHz.

4.13 MULTI-CONDUCTOR TRANSMISSION LINES

Multiple conductors in parallel are always encountered in RFICs. These conductors are used to realize a certain desired function and/or structure such as parallel-coupled transmission lines. They may also occur undesirably due to layouts, particularly in highly dense RFICs. Regardless whether or not these parallel conductors are used to realize a certain function (e.g., filtering) and/or structure (e.g., parallel-coupled transmission lines), they together form multiconductor transmission lines and hence should be treated as transmission lines in RFIC design, particularly at high frequencies.

For the purpose of analysis, we consider transmission lines formed by n conductors in parallel and a ground plane (represented by another conductor), which are hereafter referred to as n parallel-coupled transmission lines. As stated in Section 4.10, the wave propagating along these transmission lines, when embedded in a homogeneous medium with perfect conductors, is TEM wave. When the media surrounding all the conductors are inhomogeneous, as in microstrip lines, the wave propagation is no longer TEM due to the existence of dielectric interfaces where different intrinsic propagation velocities exist. However, the bulk of the energy is transmitted along the transmission lines with a field distribution quite closely resembling TEM. Thus, the propagation mode along such transmission lines can be considered quasi-TEM mode. The electric and magnetic fields are assumed to be in the transverse plane with respect to the direction of propagation. This section covers the fundamentals of parallel-coupled transmission lines including transmission-line equations, characteristic and admittance impedances, velocities, propagation constants, and impedance and admittance matrices, which are essential for the understanding and design of parallel-coupled transmission lines in RFICs.

4.13.1 Transmission-Line Equations

Figure 4.34 shows n parallel-coupled transmission lines. The n conductors can be on the same or different metal layers and the ground plane can be on a separate metal layer or on the same metal layer as some of the conductors. Figure 4.35 shows an example of n parallel-coupled transmission lines with conductors on different metal layers and Figure 4.36 shows a schematic of the transmission lines consisting of n parallel lines and a ground line. The transmission lines are assumed to be embedded in a homogeneous or inhomogeneous medium, uniform along the length (z-coordinate), and having arbitrary cross-sectional configuration.

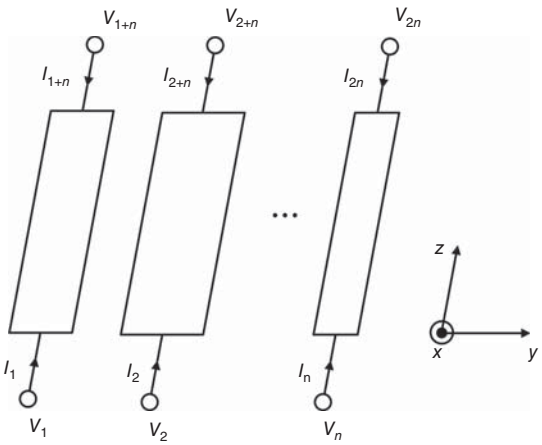


Figure 4.34. n parallel-coupled transmisison lines. Port voltages V_i, $i = 1, 2, \ldots, n$, are with respect to ground.

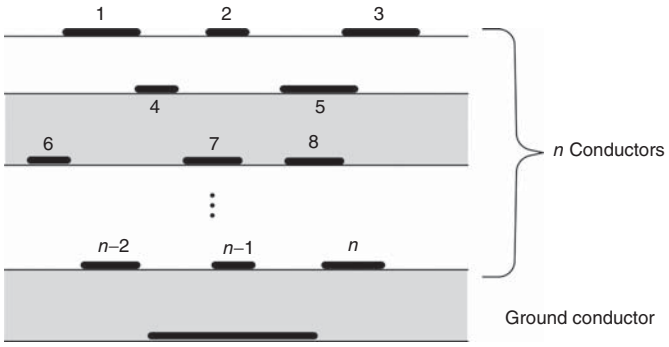


Figure 4.35. Illustration of n parallel-coupled transmission lines with conductors on different metal layers.

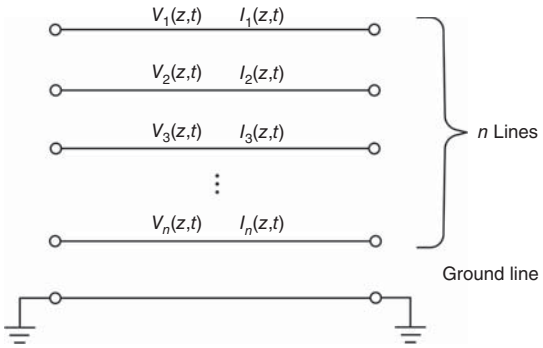


Figure 4.36. Schematic of n parallel-coupled transmission lines.

As proved in the next section, there are n TEM or quasi-TEM modes for waves propagating on the transmission lines. Generalizing the equivalent circuit for an infinitesimal section of two-conductor transmission lines shown in Figure 4.2, we can derive an equivalent circuit for a very short section (compared to wavelength) of n-parallel-coupled transmission lines. For simplicity without loss of generality, we show in Figure 4.37 such an equivalent circuit for three parallel-coupled transmission lines with a common ground plane.

Applying the same formulation based on Kirchhoff's current and voltage laws for the two-conductor transmission line described in Section 4.2 to the n-parallel-coupled transmission lines modeled similar to that shown in Figure 4.37, we can derive the telegraphist's equations:

$$\frac{\partial[V(z,t)]}{\partial z} + [L]\frac{\partial[I(z,t)]}{\partial z} + [R][I(z,t)] = 0 \tag{4.340}$$

$$\frac{\partial[I(z,t)]}{\partial z} + [C]\frac{\partial[V(z,t)]}{\partial z} + [G][V(z,t)] = 0 \tag{4.341}$$

where

$$V(z,t) = \begin{bmatrix} V_1(z,t) & V_2(z,t) & \ldots & V_n(z,t) \end{bmatrix}^T \tag{4.342}$$

$$I(z,t) = \begin{bmatrix} I_1(z,t) & I_2(z,t) & \ldots & I_n(z,t) \end{bmatrix}^T \tag{4.343}$$

are the instantaneous voltage and current matrices containing the voltages and currents on the conductors, respectively, and

$$[R] = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ R_{n1} & R_{n2} & \cdots & R_{nn} \end{bmatrix} \tag{4.344}$$

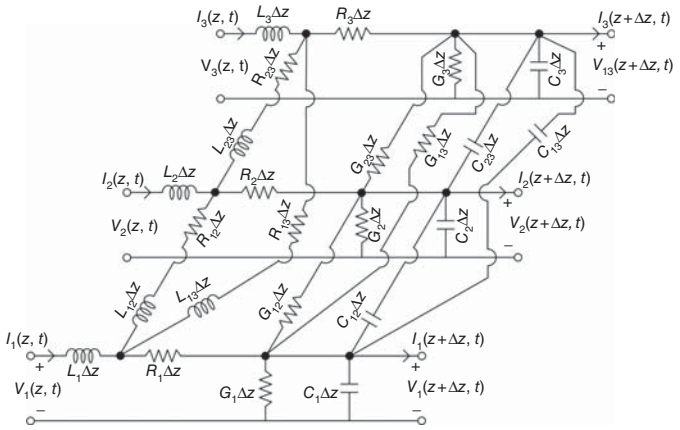


Figure 4.37. Equivalent circuit of a very short section Δz of three parallel-coupled transmission lines ($n = 3$). $V_i(z,t)$, $V_i(z+\Delta z,t)$ and $I_i(z,t)$, $I_i(z+\Delta z,t)$, $i = 1,2,3$, are the instantaneous voltages and currents on conductor i with respect to ground at the input (z) and output ($z+\Delta z$), respectively. R_i, L_i and G_i, C_i represent the (self) series resistance, inductance and (self) shunt conductance, capacitance per unit length for conductor i. $R_{i,i+1}$, $L_{i,i+1}$, $G_{i,i+1}$, and $C_{i,i+1}$, $i = 1,2$, stand for the coupling resistance, mutual inductance, coupling conductance, and mutual capacitance per-unit-length due to the coupling between conductors i and $i+1$. L's and C's account for the respective magnetic and electric energy of the signal in the transmission lines. R's and G's account for the respective conductor and dielectric losses. The couplings between non-adjacent conductors are neglected.

$$[L] = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & L_{22} & \cdots & L_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{bmatrix} \tag{4.345}$$

$$[G] = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix} \tag{4.346}$$

and

$$[C] = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{bmatrix} \tag{4.347}$$

are the resistance, inductance, conductance, and capacitance matrices per unit length of the transmission lines, respectively. We can write from (4.340) and (4.341):

$$\frac{\partial^2[V(z,t)]}{\partial z^2} - [L][C]\frac{\partial^2[V(z,t)]}{\partial t^2} - \{[R][C] + [L][G]\}\frac{\partial[V(z,t)]}{\partial t} - [R][G][V(z,t)] = 0 \tag{4.348}$$

$$\frac{\partial^2[I(z,t)]}{\partial z^2} - [C][L]\frac{\partial^2[I(z,t)]}{\partial t^2} - \{[C][R] + [G][L]\}\frac{\partial[I(z,t)]}{\partial t} - [G][R][I(z,t)] = 0 \tag{4.349}$$

which represent the transmission-line equations for n parallel-coupled transmission lines. These equations are general and hold for arbitrarily time-dependent voltages and currents and for any kinds of transmission lines. It is noted that, in general, the matrix products in (4.348) and (4.349) do not commute, and so the order of multiplication needs to be followed as indicated. It is not straight-forward to determine the solutions for (4.348) and (4.349). A possible way to derive these solutions is to use Fourier transform.

For lossless transmission lines, $[R] = [G] = 0$, and (4.348) and (4.349) become

$$\frac{\partial^2[V(z,t)]}{\partial z^2} - [L][C]\frac{\partial^2[V(z,t)]}{\partial t^2} = 0 \tag{4.350}$$

and

$$\frac{\partial^2[I(z,t)]}{\partial z^2} - [C][L]\frac{\partial^2[I(z,t)]}{\partial t^2} = 0 \tag{4.351}$$

respectively, which are the transmission-line equations for lossless transmission lines. Following the solution for the voltage on the lossless two-conductor transmission line as given in (4.19), we can write the solutions for the voltages and currents on n parallel-coupled lossless transmission lines in terms of the forward-traveling (+) and backward-traveling (−) voltages and currents as

$$[V(z,t)] = \left[V^+\left(t - \frac{z}{[v]}\right)\right] + \left[V^-\left(t + \frac{z}{[v]}\right)\right] \tag{4.352}$$

$$[I(z,t)] = \left[I^+\left(t - \frac{z}{[v]}\right)\right] + \left[I^-\left(t + \frac{z}{[v]}\right)\right] \tag{4.353}$$

where $[v]$ is a diagonal matrix whose diagonal elements are velocities. By introducing subscript k indicating mode the voltages and currents on line i for mode k can be written as

$$V_{ik}(z,t) = V_{ik}^+\left(t - \frac{z}{v_k}\right) + V_{ik}^-\left(t + \frac{z}{v_k}\right) \tag{4.354}$$

$$I_{ik}(z,t) = I_{ik}^+\left(t - \frac{z}{v_k}\right) + I_{ik}^-\left(t + \frac{z}{v_k}\right) \tag{4.355}$$

where v_k is the velocity of the wave in mode k, and V_{ik}^+ and V_{ik}^- represent the voltage waves propagating in the $+z$ and $-z$ direction, respectively. As for the two-conductor transmission line, we can demonstrate waves moving in both $+z$ and $-z$ direction for the voltage and current in (4.354) and (4.355).

Assuming sinusoidal voltages and currents and the transmission lines are operated under a steady-state condition, we can then use the phasor voltages and currents, and accordingly derive the following telegraphist's equations [21]:

$$\frac{d[V(z)]}{dz} + [Z][I(z)] = 0 \tag{4.356}$$

$$\frac{d[I(z)]}{dz} + [Y][V(z)] = 0 \tag{4.357}$$

or, in matrix form:

$$\begin{bmatrix} \dfrac{d[V(z)]}{dz} \\ \dfrac{d[I(z)]}{dz} \end{bmatrix} = \begin{bmatrix} [0] & -[Z] \\ -[Y] & [0] \end{bmatrix} \begin{bmatrix} [V(z)] \\ [I(z)] \end{bmatrix} \tag{4.358}$$

where $[V(z)]$ and $[I(z)]$ are n-dimensional column vectors with elements $V_i(z)$ and $I_i(z)$, $i = 1, 2, \ldots, n$, representing the phasor voltage and current on conductor i, respectively, which are functions of location (z) only; and $[Z] = [R] + j\omega[L]$ and $[Y] = [G] + j\omega[C]$ are the $n \times n$ series impedance and shunt admittance matrices per unit length, respectively, and are independent of z for uniform transmission lines. $[R]$ and $[G]$ are the per-unit-length $n \times n$ series resistance and shunt conductance matrices accounting for the conductor and dielectric losses of the transmission lines due to the conductors' finite conductivity and lossy dielectrics,

respectively. $[L]$ and $[C]$ are the per unit length $n \times n$ series inductance and shunt capacitance matrices representing the stored magnetic and electric energy on the transmission lines, respectively. Both $[L]$ and $[C]$ are real and symmetric. Also, $[L]$ has all positive elements, and $[C]$ has all positive diagonal elements and all negative off-diagonal elements.

Differentiating (4.356) with respect to z and substituting (4.357) into the resulting equation, we obtain the following transmission-line equation:

$$\frac{d^2[V(z)]}{dz^2} - [Z][Y][V(z)] = 0 \tag{4.359}$$

Similarly, differentiating (4.357) and making use of (4.356) gives another transmission-line equation:

$$\frac{d^2[I(z)]}{dz^2} - [Y][Z][I(z)] = 0 \tag{4.360}$$

It is noted that the matrix products $[Z][Y]$ and $[Y][Z]$ are not necessarily the same for general n parallel transmission lines.

A general solution to (4.359) is

$$[V(z)] = [e^{-[\Gamma]z}][V]^+ + [e^{[\Gamma]z}][V]^- \tag{4.361}$$

where $[\Gamma] = [[Z][Y]]^{1/2}$, $[e^{[\Gamma]z}]$ is an $n \times n$ diagonal matrix whose diagonal elements are $\exp(\gamma z)$, with γ being the wave propagation constant for each mode, and $[V]^+$ and $[V]^-$ are n-dimensional column vectors whose elements, V_i^+ and V_i^-, $i = 1, 2, \ldots, n$, are arbitrary amplitude constants for wave propagating on conductor i in the $+z$ and $-z$ directions, respectively. Equation (4.361) can be rewritten as

$$[V(z)] = [V]^+ e^{-\gamma z} + [V]^- e^{\gamma z} \tag{4.362}$$

The currents on the parallel-coupled lines can be found from (4.360) or (4.356) and (4.362) as

$$\begin{aligned}[I(z)] &= \gamma [Z]^{-1}[[V]^+ e^{-\gamma z} - [V]^- e^{\gamma z}] \\ &= [I]^+ e^{-\gamma z} - [I]^- e^{\gamma z}\end{aligned} \tag{4.363}$$

Each set of voltages and currents corresponds to one propagation mode.

4.13.2 Propagation Modes

Differentiating (4.362) and (4.363) with respect to z twice gives

$$\frac{d^2[V(z)]}{dz^2} = \gamma^2 [V(z)] \tag{4.364}$$

and

$$\frac{d^2[I(z)]}{dz^2} = \gamma^2 [I(z)] \tag{4.365}$$

Substituting (4.364) and (4.365) into (4.359) and (4.360), respectively, results in the following equations for voltages and currents:

$$[Z][Y][V(z)] = \gamma^2 [V(z)] \tag{4.366}$$

$$[Y][Z][I(z)] = \gamma^2 [I(z)] \tag{4.367}$$

Equations (4.366) and (4.367) represent two eigenvalue problems, in which the squares of the propagation constants, γ^2, are the eigenvalues of both matrices $[Z][Y]$ and $[Y][Z]$, and the associated eigenvectors of these

matrices are $[V(z)]$ and $[I(z)]$, respectively. The eigenvalues γ_k^2, $k = 1, \ldots, n$, are the roots of the following matrix determinant that represents a nth-degree characteristic equation:

$$|[Z][Y] - \gamma^2[U]| = 0 \tag{4.368}$$

where $[U]$ is a diagonal unity matrix and $||$ denotes the determinant of matrix. There are in general n distinct roots for (4.368). By taking γ_k to be the positive (negative) square roots of the eigenvalues, we can confirm that there are n propagating modes representing waves travelling in the negative (positive) z direction on the n parallel-coupled lines with distinct velocities. Each mode (or wave) is characterized by a certain distribution of voltages between the conductors and currents on the conductors described by (4.362) and (4.363), respectively. In other words, each mode is characterized by its own electric and magnetic field structures.

Because $[Z]$ and $[Y]$ are not a function of distance z for uniform lines, the eigenvectors $[V(z)]$ and $[I(z)]$ are always constant vectors. Since $[Z][Y]$ and $[Y][Z]$ are different in general, their eigenvectors can be assumed different as well. The voltage and current eigenvector matrices can be described by the following general forms:

$$[V] = \begin{bmatrix} [V_1] & [V_2] & \ldots & [V_n] \end{bmatrix} \tag{4.369}$$

and

$$[I] = \begin{bmatrix} [I_1] & [I_2] & \ldots & [I_n] \end{bmatrix} \tag{4.370}$$

respectively, where

$$[V_k] = \begin{bmatrix} V_{1k} & V_{2k} & \ldots & V_{nk} \end{bmatrix}^T \tag{4.371}$$

and

$$[I_k] = \begin{bmatrix} I_{1k} & I_{2k} & \ldots & I_{nk} \end{bmatrix}^T \tag{4.372}$$

represent the eigenvectors of $[Z][Y]$ and $[Y][Z]$ corresponding to the eigenvalues γ_k^2, respectively.

Now we define R_{ik} and R'_{ik}, $i = 2, 3, \ldots, n$ and $k = 1, 2, \ldots, n$, as the ratios of voltages and currents, respectively, on the ith and first line for mode k:

$$R_{ik} = \frac{V_{ik}}{V_{1k}} \tag{4.373}$$

$$R'_{ik} = \frac{I_{ik}}{I_{1k}} \tag{4.374}$$

and, from which, we can rewrite (4.371) and (4.372) as

$$[V_k] = \begin{bmatrix} V_{1k} & R_{2k}V_{1k} & \ldots & R_{nk}V_{1k} \end{bmatrix}^T \tag{4.375}$$

and

$$[I_k] = \begin{bmatrix} I_{1k} & R'_{2k}I_{1k} & \ldots & R'_{nk}I_{1k} \end{bmatrix}^T \tag{4.376}$$

4.13.3 Characteristic Impedance and Admittance Matrix

Differentiating (4.362) and (4.363) with respect to z and substituting into (4.356) and (4.357), respectively, we obtain the following equations:

$$[I(z)] = \gamma[Z]^{-1}[[V]^+ e^{-\gamma z} - [V]^- e^{\gamma z}] \tag{4.377}$$

$$[V(z)] = \gamma[Y]^{-1}[[I]^+ e^{-\gamma z} + [I]^- e^{\gamma z}] \tag{4.378}$$

The characteristic impedance matrix $[Z_o]$ is defined as $[V]^+/[I]^+$ or $-[V]^-/[I]^-$ and can be determined from (4.377) and (4.378) as

$$[Z_o] = \frac{[Z]}{\gamma} = \gamma[Y]^{-1} \tag{4.379}$$

The characteristic admittance matrix $[Y_o]$ is defined simply as the inverse of the impedance matrix:

$$[Y_o] = \frac{[Y]}{\gamma} = \gamma[Z]^{-1} \tag{4.380}$$

Eq. (4.379) can be rewritten as

$$[Z_o] = \gamma[U][Y]^{-1} \tag{4.381}$$

Solving (4.379) for γ and substituting into (4.381) gives

$$[Z_o] = [Z_o]^{-1}[Z][Y]^{-1} \tag{4.382}$$

Multiplying both sides of (3.382) by $[Z_o]$, we obtain:

$$[Z_o]^2 = [Z][Y]^{-1} \tag{4.383}$$

or

$$[Z_o] = [Y_o]^{-1} = [[Z][Y]^{-1}]^{1/2} \tag{4.384}$$

which can be reduced to the familiar result of the characteristic impedance for the two-conductor transmission line:

$$Z_o = Y_o^{-1} = \sqrt{Z/Y} \tag{4.385}$$

It is worthwhile to mention that the square of a matrix $[A]$ is $[A]^2 = [A][A]$, whereas the square-root of a matrix $[A]$ is $[A']$ obtained by decomposing it into two identical matrices as $[A] = [A'][A']$. It is noted that if the conductors of the n parallel-coupled transmission lines are sufficiently far from each other, $[Z_o]$ and $[Y_o]$ would have negligible off-diagonal elements, whereas when the conductors are near each other, these matrices would have significant off-diagonal elements. When the parallel-coupled lines are terminated by a network-load represented by $[Z_o]$ or $[Y_o]$, there will be no reflections on the lines at the load. Otherwise, there will be reflections on the lines, causing additional cross-talk or coupling between them which, in turn, results in more distortion to the signals and currents on the lines. In general, we can classify the cross-talk on parallel-coupled lines as termination cross-talk and reflection cross-talk, and corresponding distortion as termination distortion and reflection distortion. These cross-talks and distortions do not depend on the length of the transmission lines. Another cross-talk is the propagation cross-talk, which is a function of the location (z) on the lines. It is worthwhile to recall that for the two-conductor transmission line, the distortion or signals is caused by either dispersion, frequency-dependent loss, or reflection, while for parallel-coupled transmission lines, the distortion results from either dispersion, loss, reflection, or coupling.

It is apparent that when a voltage is applied to one conductor of the n parallel-coupled transmission lines, there will be current flowing on all the other conductors according to (4.373), and if a current is imposed upon one conductor then there exists voltages on other conductors as deduced from (4.374), which are intuitively expected. These effects are desired (e.g., those in filters or couplers) or undesired (e.g., interconnects in RFIC) depending on a particularly design. The undesired effects are normally referred to as cross-talk or unwanted coupling.

4.13.4 Mode Characteristic Impedances and Admittances

For general n parallel-coupled transmission lines consisting of n parallel conductors and a ground plane, there are n distinct (mode) characteristic impedances (or admittances) for each line corresponding to n modes. When the conductors are arranged symmetrically, degeneracy may occur among these characteristic impedances (or admittances).

Equations (4.356) and (4.357) can be expanded as

$$\frac{d}{dz}\begin{bmatrix} V_1(z) \\ V_2(z) \\ \vdots \\ V_n(z) \end{bmatrix} + \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{21} & Z_{22} & \dots & Z_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{nn} \end{bmatrix} \begin{bmatrix} I_1(z) \\ I_2(z) \\ \vdots \\ I_n(z) \end{bmatrix} = 0 \tag{4.386}$$

and

$$\frac{d}{dz}\begin{bmatrix} I_1(z) \\ I_2(z) \\ \vdots \\ I_n(z) \end{bmatrix} + \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1n} \\ Y_{21} & Y_{22} & \dots & Y_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nn} \end{bmatrix} \begin{bmatrix} V_1(z) \\ V_2(z) \\ \vdots \\ V_n(z) \end{bmatrix} = 0 \tag{4.387}$$

respectively. As done earlier, by introducing subscript k for mode k into (4.362) and (4.363), we can obtain the voltages and currents of the waves on line i of the n parallel-coupled lines for mode k as

$$V_{ik}(z) = V_{ik}^{+} e^{-\gamma k z} + V_{ik}^{-} e^{\gamma_k z} \tag{4.388}$$

and

$$I_{ik}(z) = I_{ik}^{+} e^{-\gamma k z} - I_{ik}^{-} e^{\gamma_k z} \tag{4.389}$$

The characteristic impedance Z_{oik} of line i for mode k is defined as

$$Z_{oik} = \frac{V_{ik}^{+}}{I_{ik}^{+}} = -\frac{V_{ik}^{-}}{I_{ik}^{-}} \tag{4.390}$$

and the corresponding characteristic admittance Y_{oik} is simply defined as the inverse of the characteristic impedance as

$$Y_{oik} = \frac{1}{Z_{oik}} \tag{4.391}$$

Substituting (4.373) and (4.374) into (4.388) and (4.389), respectively, we can obtain the following relations:

$$R_{ik} = \frac{V_{ik}^{+}}{V_{1k}^{+}} = \frac{V_{ik}^{-}}{V_{1k}^{-}} \tag{4.392}$$

$$R'_{ik} = \frac{I_{ik}^{+}}{I_{1k}^{+}} = \frac{I_{ik}^{-}}{I_{1k}^{-}} \tag{4.393}$$

Using (4.388), (4.389), and (4.390) in (4.386), we get

$$\gamma_k \left[\begin{bmatrix} Z_{o1k} I_{1k}^{+} \\ Z_{02k} I_{2k}^{+} \\ \vdots \\ Z_{onk} I_{nk}^{+} \end{bmatrix} e^{-\gamma_k z} - \begin{bmatrix} Z_{o1k} I_{1k}^{-} \\ Z_{o2k} I_{2k}^{-} \\ \vdots \\ Z_{onk} I_{nk}^{-} \end{bmatrix} e^{\gamma_k z} \right] = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{21} & Z_{22} & \dots & Z_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nn} \end{bmatrix} \left[\begin{bmatrix} I_{1k}^{+} \\ I_{2k}^{+} \\ \vdots \\ I_{nk}^{+} \end{bmatrix} e^{-\gamma_k z} - \begin{bmatrix} I_{1k}^{-} \\ I_{2k}^{-} \\ \vdots \\ I_{nk}^{-} \end{bmatrix} e^{\gamma_k z} \right] \tag{4.394}$$

A system of n equations can be deduced from (4.394) as

$$\begin{aligned}
\gamma_k Z_{o1k} I_{1k}^{+} &= Z_{11} I_{1k}^{+} + Z_{12} I_{2k}^{+} + \cdots + Z_{1n} I_{nk}^{+} \\
\gamma_k Z_{o2k} I_{2k}^{+} &= Z_{21} I_{1k}^{+} + Z_{22} I_{2k}^{+} + \cdots + Z_{2n} I_{nk}^{+} \\
&\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \cdots \quad \vdots \\
\gamma_k Z_{onk} I_{nk}^{+} &= Z_{n1} I_{1k}^{+} + Z_{n2} I_{2k}^{+} + \cdots + Z_{nn} I_{nk}^{+}
\end{aligned} \tag{4.395}$$

Substituting (4.393) into (4.395) and eliminating the current amplitude I_{1k}^{+} results in another n-equation system:

$$\begin{aligned}
\gamma_k Z_{o1k} &= Z_{11} + Z_{12} R'_{2k} + \cdots + Z_{1n} R'_{nk} \\
\gamma_k Z_{o2k} R'_{2k} &= Z_{21} + Z_{22} R'_{2k} + \cdots + Z_{2n} R'_{nk} \\
&\vdots \quad \vdots \quad \vdots \qquad \vdots \quad \cdots \quad \vdots \\
\gamma_k Z_{onk} R'_{nk} &= Z_{n1} + z_{n2} R'_{2k} + \cdots + z_{nn} R'_{nk}
\end{aligned} \tag{4.396}$$

By solving this system of equations, the characteristic impedances of the n parallel-coupled transmission lines for the n modes can be found, assuming $R'_{jk} \neq 0$, as

$$\begin{aligned}
Z_{o1k} &= \frac{1}{\gamma_k}(Z_{11} + Z_{12} R'_{2k} + \cdots + Z_{1n} R'_{nk}) \\
Z_{ojk} &= \frac{1}{\gamma_k R'_{jk}}(Z_{j1} + Z_{j2} R'_{2k} + \cdots + Z_{jn} R'_{nk})
\end{aligned} \tag{4.397}$$

$j = 2, 3, \ldots, n$. Using (4.386)–(4.397), other sets of n equations can be derived in a similar fashion as

$$\begin{aligned}
\gamma_k &= Y_{11} Z_{o1k} + Y_{12} Z_{o2k} R'_{2k} + \cdots + Y_{1n} Z_{onk} R'_{nk} \\
\gamma_k R'_{2k} &= Y_{21} Z_{o1k} + Y_{22} Z_{o2k} R'_{2k} + \cdots + Y_{2n} Z_{onk} R'_{nk} \\
&\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \cdots \qquad \vdots \\
\gamma_k R'_{nk} &= Y_{n1} Z_{o1k} + Y_{n2} Z_{o2k} R'_{2k} + \cdots + Y_{nn} Z_{onk} R'_{nk}
\end{aligned} \tag{4.398}$$

$$\begin{aligned}
\gamma_k Y_{o1k} &= Y_{11} + Y_{12} R_{2k} + \cdots + Y_{1n} R_{nk} \\
\gamma_k R_{2k} Y_{o2k} &= Y_{21} + Y_{22} R_{2k} + \cdots + Y_{2n} R_{nk} \\
&\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \cdots \quad \vdots \\
\gamma_k R_{nk} Y_{onk} &= Y_{n1} + Y_{n2} R_{2k} + \cdots + Y_{nn} R_{nk}
\end{aligned} \tag{4.399}$$

and

$$\gamma_k = Z_{11}Y_{o1k} + Z_{12}Y_{o2k}R_{2k} + \cdots + Z_{1n}Y_{onk}R_{nk}$$

$$\gamma_k R_{2k} = Z_{21}Y_{o1k} + Z_{22}Y_{o2k}R_{2k} + \cdots + Z_{2n}Y_{onk}R_{nk}$$

$$\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \cdots \qquad \vdots$$

$$\gamma_k R_{nk} = Z_{n1}Y_{o1k} + Z_{n2}Y_{o2k}R_{2k} + \cdots + Z_{nn}Y_{onk}R_{nk} \tag{4.400}$$

assuming $R_{jk} \neq 0$. Equation (4.399) can be solved for the characteristic admittances of the n parallel-coupled transmission lines for the n modes as

$$Y_{o1k} = \frac{1}{\gamma_k}(Y_{11} + Y_{12}R_{2k} + \cdots + Y_{1n}R_{nk})$$

$$Y_{ojk} = \frac{1}{\gamma_k R_{jk}}(Y_{j1} + Y_{j2}R_{2k} + \cdots + Y_{jn}R_{nk}) \tag{4.401}$$

4.13.5 Impedance and Admittance Matrix

The behavior of n parallel-coupled transmission lines can be conveniently described by an impedance or admittance matrix. In order to derive the impedance matrix $[Z]$ or admittance matrix $[Y]$ for n parallel-coupled transmission lines, we must relate the voltages and currents at its input ($z = 0$) to the voltages and currents at its output ($z = l$), where l is the length of the lines.

In general, there exist n simultaneous pairs of forward and backward traveling waves of propagation constants $(\gamma_1, \ldots, \gamma_n)$ on each line (with respect to ground). The total voltage and current on each line may therefore be written as a superposition of the voltages and currents of these waves, respectively, as

$$V_i(z) = \sum_{k=1}^{n}(V_{ik}^{+}e^{-\gamma_k z} + V_{ik}^{-}e^{\gamma_k z}) \tag{4.402}$$

and

$$I_i(z) = \sum_{k=1}^{n}(I_{ik}^{+}e^{-\gamma_k z} + I_{ik}^{-}e^{\gamma_k z})$$

$$= \sum_{k=1}^{n} Y_{oik}(V_{ik}^{+}e^{-\gamma_k z} - V_{ik}^{-}e^{\gamma_k z}) \tag{4.403}$$

Using (4.392) in (4.402) and (4.403), the voltages and currents on the lines can now be expressed in terms of the voltages on the first line as

$$V_1(z) = V_{11}^{+}e^{-\gamma_1 z} + V_{11}^{-}e^{\gamma_1 z} + V_{12}^{+}e^{-\gamma_2 z} + V_{12}^{-}e^{\gamma_2 z} + \cdots + V_{1n}^{+}e^{-\gamma_n z} + V_{1n}^{-}e^{\gamma_n z}$$

$$V_2(z) = R_{21}V_{11}^{+}e^{-\gamma_1 z} + R_{21}V_{11}^{-}e^{\gamma_1 z} + R_{22}V_{12}^{+}e^{-\gamma_2 z} + R_{22}V_{12}^{-}e^{\gamma_2 z} + \cdots + R_{2n}V_{1n}^{+}e^{-\gamma_n z} + R_{2n}V_{1n}^{-}e^{\gamma_n z}$$

$$\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \cdots \qquad \vdots \qquad \vdots$$

$$V_n(z) = R_{n1}V_{11}^{+}e^{-\gamma_1 z} + R_{n1}V_{11}^{-}e^{\gamma_1 z} + R_{n2}V_{12}^{+}e^{-\gamma_2 z} + R_{n2}V_{12}^{-}e^{\gamma_2 z} + \cdots + R_{nn}V_{1n}^{+}e^{-\gamma_n z} + R_{nn}V_{1n}^{-}e^{\gamma_n z} \tag{4.404}$$

and

$$I_1(z) = Y_{o11}V_{11}^{+}e^{-\gamma_1 z} - Y_{o11}V_{11}^{-}e^{\gamma_1 z} + Y_{o12}V_{12}^{+}e^{-\gamma_2 z} - Y_{o12}V_{12}^{-}e^{\gamma_2 z} + \cdots + Y_{o1n}V_{1n}^{+}e^{-\gamma_n z} - Y_{o1n}V_{1n}^{-}e^{\gamma_n z}$$

$$I_2(z) = R_{21}Y_{o21}V_{11}^{+}e^{-\gamma_1 z} - R_{21}Y_{o21}V_{11}^{-}e^{\gamma_1 z} + R_{22}Y_{o22}V_{12}^{+}e^{-\gamma_2 z} - R_{22}Y_{o22}V_{12}^{-}e^{\gamma_2 z} + \cdots + R_{2n}Y_{o2n}V_{1n}^{+}e^{-\gamma_n z} - R_{2n}Y_{o2n}V_{1n}^{-}e^{\gamma_n z}$$

$$I_n(z) = R_{2n}Y_{on1}V_{11}^{+}e^{-\gamma_1 z} - R_{n1}Y_{on1}V_{11}^{-}e^{\gamma_1 z} + R_{n2}Y_{on2}V_{12}^{+}e^{-\gamma_2 z} - R_{n2}Y_{on2}V_{12}^{-}e^{\gamma_2 z} + \cdots + R_{nn}Y_{onn}V_{1n}^{+}e^{-\gamma_n z} - R_{nn}Y_{onn}V_{1n}^{-}e^{\gamma_n z} \tag{4.405}$$

Applying the terminating conditions:

$$V_i(z=0) = V_i$$

$$V_i(z=l) = V_{i+n}$$

$$I_i(z=0) = I_i$$

$$I_i(z=l) = I_{i+n} \tag{4.406}$$

where V_i, V_{i+n} and I_i, I_{i+n} are the port voltages and currents, respectively, as defined in Figure 4.34, to (4.404) and (4.405), yields the following equations:

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \\ V_{1+n} \\ V_{2+n} \\ \vdots \\ V_{2n} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ R_{21} & R_{21} & R_{22} & R_{22} & \cdots & R_{2n} & R_{2n} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ R_{n1} & R_{n1} & R_{n2} & R_{n2} & \cdots & R_{nn} & R_{nn} \\ e^{-\gamma_1 l} & e^{\gamma_1 l} & e^{-\gamma_2 l} & e^{\gamma_2 l} & \cdots & e^{-\gamma_n l} & e^{\gamma_n l} \\ R_{21}e^{-\gamma_1 l} & R_{21}e^{\gamma_1 l} & R_{22}e^{-\gamma_2 l} & R_{22}e^{\gamma_2 l} & \cdots & R_{2n}e^{-\gamma_n l} & R_{2n}e^{\gamma_n l} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ R_{n1}e^{-\gamma_1 l} & R_{n1}e^{\gamma_1 l} & R_{n2}e^{-\gamma_2 l} & R_{n2}e^{\gamma_2 l} & \cdots & R_{nn}e^{-\gamma_n l} & R_{nn}e^{\gamma_n l} \end{bmatrix} \begin{bmatrix} V_{11}^{+} \\ V_{11}^{-} \\ V_{12}^{+} \\ V_{12}^{-} \\ \vdots \\ V_{1n}^{+} \\ V_{1n}^{-} \end{bmatrix} \tag{4.407}$$

and

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \\ I_{1+n} \\ I_{2+n} \\ \vdots \\ I_{2n} \end{bmatrix} = \begin{bmatrix} Y_{o11} & -Y_{o11} & Y_{o12} & -Y_{o12} & \cdots & Y_{o1n} & -Y_{o1n} \\ R_{21}Y_{o21} & -R_{21}Y_{o21} & R_{22}Y_{o22} & -R_{22}Y_{o22} & \cdots & R_{2n}Y_{o2n} & -R_{2n}Y_{o2n} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ R_{n1}Y_{on1} & -R_{n1}Y_{on1} & R_{n2}Y_{on2} & -R_{n2}Y_{on2} & \cdots & R_{nn}Y_{onn} & -R_{nn}Y_{onn} \\ -Y_{o11}e^{-\gamma_1 l} & Y_{o11}e^{\gamma_1 l} & -Y_{o12}e^{-\gamma_2 l} & Y_{012}e^{\gamma_2 l} & \cdots & -Y_{o1n}e^{-\gamma_n l} & Y_{o1n}e^{\gamma_n l} \\ -R_{21}Y_{021}e^{-\gamma_1 l} & R_{21}Y_{021}e^{\gamma_1 l} & -R_{22}Y_{022}e^{-\gamma_2 l} & R_{22}Y_{022}e^{\gamma_2 l} & \cdots & -R_{2n}Y_{02n}e^{-\gamma_n l} & R_{2n}Y_{02n}e^{\gamma_n l} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ -R_{n1}Y_{on1}e^{-\gamma_1 l} & R_{n1}Y_{on1}e^{\gamma_1 l} & -R_{n2}Y_{on2}e^{-\gamma_2 l} & R_{n2}Y_{on2}e^{\gamma_2 l} & \cdots & -R_{nn}Y_{onn}e^{-\gamma_n l} & R_{nn}Y_{onn}{}^{\gamma_n l} \end{bmatrix} \begin{bmatrix} V_{11}^{+} \\ V_{11}^{-} \\ V_{12}^{+} \\ V_{12}^{-} \\ \vdots \\ V_{1n}^{+} \\ V_{1n}^{-} \end{bmatrix} \tag{4.408}$$

The impedance, admittance, and chain matrices for the $2n$-port network of the n parallel-coupled transmission lines, as shown in Figure 4.34, can now be derived from the relationships between the port currents and voltages in (4.407), (4.408) obtained by eliminating the amplitude coefficients V_{1i}^{+} and V_{1i}^{-}. For example, the impedance matrix $[Z]$ is found by solving for the port voltages in terms of the port currents according to $[Z] = [V]/[I]$ and the admittance matrix $[Y]$ is obtained by determining the port currents as a function of the port voltages according to $[Y] = [I]/[V]$.

4.13.6 Lossless Multiconductor Transmission Lines

The analysis of lossless multiconductor transmission lines can be derived using the results obtained previously. As for the lossless two-conductor transmission line, there are no resistances along the conductors, between the conductors, and between the conductors and ground for the lossless n parallel-coupled transmission lines. The lossless n parallel-coupled transmission line is thus characterized by $[R] = [G] = [0]$, and its impedance and admittance matrices per unit length and propagation constant are

$$[Z] = j\omega[L] \tag{4.409}$$

$$[Y] = j\omega[C] \tag{4.410}$$

and the propagation constant for each mode is

$$\gamma = j\beta = j\frac{\omega}{v} \tag{4.411}$$

where β and v are the phase constant and phase velocity for each mode, respectively.

4.13.6.1 Characteristic Impedance Matrix. Using (4.409)–(4.411) in (4.379) and (4.384), the characteristic impedance matrix can be expressed as

$$[Z_o] = \frac{1}{v[C]} = v[L] = [[L][C]^{-1}]^{1/2} \tag{4.412}$$

If the dielectric is replaced by air, then from (4.412) we obtain

$$[L_o] = \frac{1}{v_o^2}[C_o]^{-1} \tag{4.413}$$

where v_o is the free-space propagation velocity and $[C_o]$ is the per-unit-length (static) capacitance matrix of the n parallel-coupled transmission lines without the dielectric. If the dielectric is nonmagnetic then its permeability is equal to that of air, and $[L] = [L_o]$. Substituting (4.413) into (4.412) yields

$$[Z_o] = \frac{[[C_o][C]]^{-1/2}}{v_o} \tag{4.414}$$

Solving Eq. (4.412) for $[L]$ and substituting into (4.413) yields:

$$[C_o] = \left(\frac{v}{v_o}\right)^2 [C] \tag{4.415}$$

Thus, it is evident that the electrical behavior of lossless n parallel-coupled transmission lines can be completely described in terms of the static capacitance matrices of the structure with and without dielectrics.

4.13.6.2 Propagation Velocities. Substituting equations (4.409)–(4.411) into (4.366) and (4.367), we obtain the following eigenvalue equations:

$$[L][C][V(z)] = \frac{1}{v^2}[V(z)] \tag{4.416}$$

and

$$[C][L][I(z)] = \frac{1}{v^2}[I(z)] \tag{4.417}$$

or, after using (4.413):

$$\frac{1}{v_o^2}[C_o]^{-1}[C][V(z)] = \frac{1}{v^2}[V(z)] \tag{4.418}$$

and

$$\frac{1}{v_o^2}[C][C_o]^{-1}[I(z)] = \frac{1}{v^2}[I(z)] \tag{4.419}$$

It is obvious that the eigenvalues of both (4.418) and (4.419) are the same, and they are the solutions of

$$\left|[L]\,[C] - \frac{1}{v^2}[U]\right| = 0 \tag{4.420}$$

or

$$\left|\frac{1}{v_o^2}\left[C_o\right]^{-1}[C] - \frac{1}{v^2}[U]\right| = 0 \tag{4.421}$$

The phase velocities for n propagating modes can be taken as the positive or negative square roots of the inverse of the eigenvalues $1/v_k^2$. The positive (negative) values correspond to waves propagating in the negative (positive) z direction. It is easy to show that these eigenvalues are real and positive. Therefore, the phase velocity v_k in mode k must be real as well. This confirms that waves propagating un-attenuated on lossless multiconductor transmission lines. Associated with these eigenvalues, there are voltage and current eigenvectors, $[V(z)]$ and $[I(z)]$, respectively. These eigenvectors can be found by substituting $1/v_k^2$ into the respective equations (4.418) and (4.419). Also, since $\frac{1}{v_o^2}[C_o]^{-1}[C]$, $\frac{1}{v^2}[C][C_o]^{-1}$ and their eigenvalues are real, the eigenvectors can be assumed real as well.

4.13.6.3 Mode Characteristic Impedances and Admittances. To derive the relations for the characteristic impedances and admittances of the n lines (associated with a ground plane) for the lossless case, we substitute (4.409)–(4.411) into (4.392)–(4.401), and obtain

$$\begin{aligned}
Z_{o1k} &= v_k(L_{11} + L_{12}R'_{2k} + \cdots + L_{1n}R'_{nk}) \\
R'_{2k}Z_{o2k} &= v_k(L_{21} + L_{22}R'_{2k} + \cdots + L_{2n}R'_{nk}) \\
&\vdots \\
R'_{nk}Z_{onk} &= v_k(L_{n1} + L_{n2}R'_{2k} + \cdots + L_{nn}R'_{nk})
\end{aligned} \tag{4.422}$$

$$\begin{aligned}
1 &= v_k(C_{11}Z_{o1k} + C_{12}Z_{o2k}R'_{2k} + \cdots + C_{1n}Z_{onk}R'_{nk}) \\
R'_{2k} &= v_k(C_{21}Z_{o1k} + C_{22}Z_{o2k}R'_{2k} + \cdots + C_{2n}Z_{onk}R'_{nk}) \\
&\vdots \\
R'_{nk} &= v_k(C_{n1}Z_{o1k} + C_{n2}Z_{o2k}R'_{2k} + \cdots + C_{nn}Z_{onk}R'_{nk})
\end{aligned} \tag{4.423}$$

$$\begin{aligned} Y_{o1k} &= v_k(C_{11} + C_{12}R_{2k} + \cdots + C_{1n}R_{nk}) \\ R_{2k}Y_{o2k} &= v_k(C_{21} + C_{22}R_{2k} + \cdots + C_{2n}R_{nk}) \\ &\vdots \\ R_{nk}Y_{onk} &= v_k(C_{n1} + C_{n2}R_{2k} + \cdots + C_{nn}R_{nk}) \end{aligned} \tag{4.424}$$

$$\begin{aligned} 1 &= v_k(L_{11}Y_{o1k} + L_{12}Y_{o2k}R_{2k} + \cdots + L_{1n}Y_{onk}R_{nk}) \\ R_{2k} &= v_k(L_{21}Y_{o1k} + L_{22}Y_{o2k}R_{2k} + \cdots + L_{2n}Y_{onk}R_{nk}) \\ &\vdots \\ R_{nk} &= v_k(L_{n1}Y_{o1k} + L_{n2}Y_{o2k}R_{2k} + \cdots + L_{nn}Y_{onk}R_{nk}) \end{aligned} \tag{4.425}$$

or, in closed-form expressions:

$$\begin{aligned} Z_{o1k} &= v_k(L_{11} + L_{12}R'_{2k} + \cdots + L_{1n}R'_{nk}) \\ Z_{ojk} &= \frac{v_k}{R'_{jk}}(L_{j1} + L_{j2}R'_{2k} + \cdots + L_{jn}R'_{nk}) \end{aligned} \tag{4.426}$$

assuming $R'_{jk} \neq 0$ and

$$\begin{aligned} Y_{o1k} &= v_k(C_{11} + C_{12}R_{2k} + \cdots + C_{1n}R_{nk}) \\ Y_{ojk} &= \frac{v_k}{R_{jk}}(C_{j1} + C_{j2}R_{2k} + \cdots + C_{jn}R_{nk}) \end{aligned} \tag{4.427}$$

assuming $R_{jk} \neq 0$.

4.13.6.4 Analysis of Lossless Coupled Lines Having [L][C] = [C][L].

In general, the matrix products $[L][C]$ and $[C][L]$ of lossless parallel-coupled transmission lines, are usually different. In practice, the conductors of parallel-coupled transmission lines, however, are not arranged arbitrarily. Instead, they are configured in a certain way to either achieve a particular function or result in a simple analysis – for instance, the conductors are arranged symmetrically. Electrically, the matrix product $[L][C]$ or $[C][L]$ would satisfy certain conditions corresponding to particular lossless parallel-coupled transmission lines. One of the most useful conditions is $[L][C] = [C][L]$, which is encountered in the well-known two symmetrical parallel-coupled lines. For these symmetrical transmission lines, we have $L_{11} = L_{22}$ and $C_{11} = C_{22}$, and hence

$$[L][C] = [C][L] \tag{4.428}$$

while, for two asymmetrical parallel-coupled lines:

$$[L][C] \neq [C][L] \tag{4.429}$$

In this section, the properties of lossless coupled transmission lines are investigated assuming the conductors are arranged in such a way that $[L][C]$ and $[C][L]$ are equal.

Since both matrices $[C_o]^{-1}$ and $[C]$ are real, symmetric, and commutative, the matrix product $[D] = \frac{1}{v_o^2}[C_o]^{-1}[C]$ must be real and symmetric, and hence can be diagonalized by a real orthogonal matrix [22]. Let $[P]$ be such a real orthogonal matrix, we can then derive:

$$\begin{aligned} [P]^{-1}[D][P] &= \frac{1}{v_o^2}[P]^{-1}[C_o]^{-1}[P][P]^{-1}[C][P] \\ &= \frac{1}{v_o^2}[C_o]_d^{-1}[C]_d = [D]_d \end{aligned} \tag{4.430}$$

where the subscript d indicates diagonal matrix. Equation (4.430) shows that $[D]$ is diagonalizable and has the same eigenvectors as $[C]$ and $[C_o]$.

Similarly, the characteristic impedance, velocity, and propagation-constant matrices can also be obtained in diagonal forms as

$$[Z_o] = \begin{bmatrix} Z_{o1} & 0 & \cdots & 0 \\ 0 & Z_{o2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \vdots & \cdots & Z_{on} \end{bmatrix} = \frac{[[C_o]_d[C]_d]^{-1/2}}{v_o} \tag{4.431}$$

$$[v] = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \vdots & \cdots & v_n \end{bmatrix} = v_o[[C_o]_d[C]_d^{-1}]^{1/2} \tag{4.432}$$

and

$$[\Gamma] = \begin{bmatrix} \gamma_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \vdots & \cdots & \gamma_n \end{bmatrix} = j\frac{\omega}{v_o}[[C_o]_d[C]_d]^{1/2} = j\omega[v]^{-1} \tag{4.433}$$

where Z_{ok}, v_k, and γ_k, $k = 1, 2, \ldots, n$, are the characteristic impedance, velocity, and propagation constant in mode k, respectively. Equation (4.431) shows that the characteristic impedances corresponding to the n conductors in each mode are equal. From (4.431)–(4.433), we can see that the electrical characteristics of n lossless parallel-coupled transmission lines can be completely described in terms of the eigenvalues of $[C]$ and $[C_o]$ of the parallel-coupled transmission lines.

4.13.6.5 Analysis of Two Parallel-Coupled Transmission Lines. As an example, the foregoing analysis for n parallel-coupled transmission lines is now applied to the analysis of two parallel-coupled transmission lines embedded in an inhomogeneous medium as shown in Figure 4.36. The transmission lines consist of two uniform conductors and a ground plane. The conductors may be identical or nonidentical. As compared to components employing coupled lines of identical conductors, coupled-line components using nonidentical conductors may have impedance-transformation characteristics.

Two Asymmetrical Coupled Lines

For the two parallel-coupled lines as shown in Figure 4.38, the per-unit-length series impedance and shunt admittance matrices can be written as

$$[Z] = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{12} & Z_{22} \end{bmatrix} \tag{4.434}$$

$$[Y] = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12} & Y_{22} \end{bmatrix} \tag{4.435}$$

The matrix product $[Z][Y]$ is of the form

$$[Z][Y] = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{4.436}$$

where

$$\begin{aligned} A &= Z_{11}Y_{11} + Z_{12}Y_{12} \\ B &= Z_{11}Y_{12} + Z_{12}Y_{22} \\ C &= Z_{12}Y_{11} + Z_{22}Y_{12} \\ D &= Z_{12}Y_{12} + Z_{22}Y_{22} \end{aligned} \tag{4.437}$$

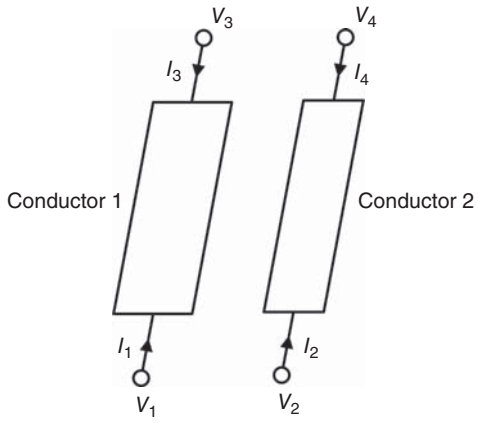


Figure 4.38. Two parallel-coupled transmission lines.

Solving the determinant equation (4.368) for the two parallel-coupled lines results in the following expression for the propagation constants for two modes, referred to here as mode x and mode y:

$$\gamma_{x,y}^2 = \frac{A+D}{2} \pm \frac{1}{2}[(A-D)^2 + 4BC]^{1/2} \tag{4.438}$$

The x mode and y mode are known as the c mode and π mode, respectively, in [23].

The voltage eigenvector matrix is

$$[V] = \begin{bmatrix} V_{1x} & V_{1y} \\ V_{2x} & V_{2y} \end{bmatrix} = \begin{bmatrix} V_{1x} & V_{1y} \\ R_{2x}V_{1x} & R_{2y}V_{1y} \end{bmatrix} \tag{4.439}$$

where

$$R_{2x(y)} \equiv \frac{V_{2x(y)}}{V_{1x(y)}} = \frac{1}{2B}\{(D-A) \pm [(D-A)^2 + 4BC]^{1/2}\} \tag{4.440}$$

is the ratio of the voltages on the two lines for mode $x(y)$ with the + and − sign corresponding to mode x and y, respectively. From (4.440), it is seen that

$$R_{2x}R_{2y} = -\frac{C}{B} \tag{4.441}$$

For the two coupled lines considered here, Eq. (4.399) is reduced to

$$\begin{aligned} \gamma_k Y_{o1k} &= Y_{11} + Y_{12}R_{2k} \\ \gamma_k R_{2k} Y_{o2k} &= Y_{12} + Y_{22}R_{2k} \end{aligned} \tag{4.442}$$

which can be solved to yield

$$Z_{o1k} = \frac{1}{Y_{o1k}} = \frac{\gamma_k}{Y_{11} + R_{2k}Y_{12}} \tag{4.443}$$

and

$$Z_{o2k} = \frac{1}{Y_{o2k}} = \frac{R_{2k}\gamma_k}{Y_{12} + R_{2k}Y_{22}} \tag{4.444}$$

where Z_{oik} and Y_{oik} ($i = 1, 2$ and $k = x, y$) are the characteristic impedance and admittance, respectively, of line i for mode k.

The port voltages and currents can be determined from (4.407) and (4.408), respectively, as

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ R_{2x} & R_{2x} & R_{2y} & R_{2y} \\ e^{-\gamma_x l} & e^{\gamma_x l} & e^{-\gamma_y l} & e^{\gamma_y l} \\ R_{2x}e^{-\gamma_x l} & R_{2x}e^{\gamma_x l} & R_{2y}e^{-\gamma_y l} & R_{2y}e^{\gamma_y l} \end{bmatrix} \begin{bmatrix} V_{1x}^+ \\ V_{1x}^- \\ V_{1y}^+ \\ V_{1y}^- \end{bmatrix} \tag{4.445}$$

and

$$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} = \begin{bmatrix} Y_{o1x} & -Y_{o1x} & Y_{o1y} & -Y_{o1y} \\ R_{2x}Y_{o2x} & -R_{2x}Y_{o2x} & R_{2y}Y_{o2y} & -R_{2y}Y_{o2y} \\ -Y_{o1x}e^{-\gamma_x l} & Y_{o1x}e^{\gamma_x l} & -Y_{o1y}e^{-\gamma_y l} & Y_{o1y}e^{\gamma_y l} \\ -R_{2x}Y_{o2x}e^{-\gamma_x l} & R_{2x}Y_{o2x}e^{\gamma_x l} & -R_{2y}Y_{o2y}e^{-\gamma_y l} & R_{2y}Y_{o2y}e^{\gamma_y l} \end{bmatrix} \begin{bmatrix} V_{1x}^+ \\ V_{1x}^- \\ V_{1y}^+ \\ V_{1y}^- \end{bmatrix} \tag{4.446}$$

where l is the length of the lines. Eliminating the amplitude coefficients V_{1x}^+, V_{1x}^-, V_{1y}^+, and V_{1y}^- leads to the impedance and admittance matrices of the coupled lines' four-port network whose elements are [23]:

$$\begin{aligned}
Z_{11} &= Z_{33} = \frac{Z_{o1x}\coth\gamma_x l}{(1-R_{2x}/R_{2y})} + \frac{Z_{o1y}\coth\gamma_y l}{(1-R_{2y}/R_{2x})} \\
Z_{12} &= Z_{21} = Z_{34} = Z_{43} = \frac{Z_{o1x}R_{2x}\coth\gamma_x l}{(1-R_{2x}/R_{2y})} + \frac{Z_{o1y}R_{2y}\coth\gamma_y l}{(1-R_{2y}/R_{2x})} \\
&= -\frac{Z_{o2x}\coth\gamma_x l}{R_{2y}(1-R_{2x}/R_{2y})} - \frac{Z_{o2y}\coth\gamma_y l}{R_{2x}(1-R_{2y}/R_{2x})} \\
Z_{14} &= Z_{41} = Z_{23} = Z_{32} = \frac{R_{2x}Z_{o1x}}{(1-R_{2x}/R_{2y})\sinh\gamma_x l} + \frac{R_{2y}Z_{o1y}}{(1-R_{2y}/R_{2x})\sinh\gamma_y l} \\
Z_{13} &= Z_{31} = \frac{Z_{o1x}}{(1-R_{2x}/R_{2y})\sinh\gamma_x l} + \frac{Z_{o1y}}{(1-R_{2y}/R_{2x})\sinh\gamma_y l} \\
Z_{22} &= Z_{44} = -\frac{R_{2x}Z_{o2x}\coth\gamma_x l}{R_{2y}(1-R_{2x}/R_{2y})} + \frac{R_{2y}Z_{o2y}\coth\gamma_y l}{R_{2x}(1-R_{2y}/R_{2x})} \\
&= \frac{R_{2x}^2 Z_{o1x}\coth\gamma_x l}{(1-R_{2x}/R_{2y})} + \frac{R_{2y}^2 Z_{o1y}\coth\gamma_y l}{(1-R_{2y}/R_{2x})} \\
Z_{24} &= Z_{42} = -\frac{R_{2x}^2 Z_{o1x}}{(1-R_{2x}/R_{2y})\sinh\gamma_x l} + \frac{R_{2y}^2 Z_{o1y}}{(1-R_{2y}/R_{2x})\sinh\gamma_y l}
\end{aligned} \tag{4.447}$$

and

$$\begin{aligned}
Y_{11} &= Y_{33} = \frac{Y_{o1x}\coth\gamma_x l}{(1-R_{2x}/R_{2y})} + \frac{Y_{o1y}\coth\gamma_y l}{(1-R_{2y}/R_{2x})} \\
Y_{12} &= Y_{21} = Y_{34} = Y_{43} = -\frac{Y_{o1x}\coth\gamma_x l}{R_{2y}(1-R_{2x}/R_{2y})} - \frac{Y_{o1y}\coth\gamma_y l}{R_{2x}(1-R_{2y}/R_{2x})} \\
Y_{14} &= Y_{41} = Y_{23} = Y_{32} = \frac{Y_{o1x}}{(R_{2y}-R_{2x})\sinh\gamma_x l} + \frac{Y_{o1y}}{(R_{2x}-R_{2y})\sinh\gamma_y l}
\end{aligned}$$

$$Y_{13} = Y_{31} = -\frac{Y_{o1x}}{(1 - R_{2x}/R_{2y})\sinh\gamma_x l} - \frac{Y_{o1y}}{(1 - R_{2y}/R_{2x})\sinh\gamma_y l}$$
$$Y_{22} = Y_{44} = -\frac{R_{2x}Y_{o2x}\coth\gamma_x l}{R_{2y}(1 - R_{2x}/R_{2y})} - \frac{R_{2y}Y_{o2y}\coth\gamma_y l}{R_{2x}(1 - R_{2y}/R_{2x})}$$
$$Y_{24} = Y_{42} = \frac{R_{2x}Y_{o2x}}{R_{2y}(1 - R_{2x}/R_{2y})\sinh\gamma_x l} + \frac{R_{2y}Y_{o2y}}{R_{2x}(1 - R_{2y}/R_{2x})\sinh\gamma_y l} \tag{4.448}$$

Under the lossless assumption, the two parallel-coupled lines shown in Figure 4.38 have the per-unit-length inductance matrix $[L]$ and capacitance matrices $[C]$ and $[C_o]$ of

$$[L] = \begin{bmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{bmatrix} \tag{4.449}$$

$$[C] = \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \tag{4.450}$$

$$[C_o] = \begin{bmatrix} C_{o11} & C_{o12} \\ C_{o12} & C_{o22} \end{bmatrix} \tag{4.451}$$

The two distinct velocities for the two modes can be found from (4.411), (4.413), (4.437), and (4.438) as

$$v_{x,y} = \left\{ \frac{d_1 + d_2}{2} \pm \frac{1}{2}\left[(d_1 - d_2)^2 + 4e_1e_2\right]^{1/2} \right\}^{-1/2} \tag{4.452}$$

where

$$d_1 = L_{11}C_{11} + L_{12}C_{12}$$
$$d_2 = L_{12}C_{12} + L_{22}C_{22}$$
$$e_1 = L_{11}C_{12} + L_{12}C_{22}$$
$$e_2 = L_{12}C_{11} + L_{22}C_{12} \tag{4.453}$$

or

$$v_{x,y} = \left\{ \frac{D_1 + D_2}{2} \pm \frac{1}{2}\left[(D_1 - D_2)^2 + 4E_1E_2\right]^{1/2} \right\}^{-1/2} \tag{4.454}$$

where

$$D_1 = \frac{(C_{11}C_{o22} - C_{12}C_{o12})}{v_o^2.|[C_o]|}$$
$$D_2 = \frac{(C_{22}C_{o11} - C_{12}C_{o12})}{v_o^2.|[C_o]|}$$
$$E_1 = \frac{(C_{12}C_{o22} - C_{22}C_{o12})}{v_o^2.|[C_o]|}$$
$$E_2 = \frac{(C_{12}C_{o11} - C_{11}C_{o12})}{v_o^2.|[C_o]|} \tag{4.455}$$

with

$$|[C_o]| = C_{o11}C_{o22} - C_{o12}^2 \tag{4.456}$$

The voltage ratio in (4.440) becomes

$$R_{2x,y} = \frac{1}{2e_1}\{(d_2 - d_1) \pm [(d_2 - d_1)^2 + 4e_1e_2]^{1/2}\}$$
$$= \frac{1}{2E_1}\{(D_2 - D_1) \pm [(D_2 - D_1)^2 + 4E_1E_2]^{1/2}\} \tag{4.457}$$

From (4.457), it is seen that

$$R_{2x}R_{2y} = -\frac{E_2}{E_1} \tag{4.458}$$

The system of equations (4.424) reduces to

$$Y_{o1k} = v_k(C_{11} + C_{12}R_{2k})$$
$$R_{2k}Y_{o2k} = v_k(C_{12} + C_{22}R_{2k}) \tag{4.459}$$

which can be solved to yield

$$Z_{o1k} = \frac{1}{v_k(C_{11} + R_{2k}C_{12})} \tag{4.460}$$

$$Z_{o2k} = \frac{1}{v_k(C_{22} + R_{2k}^{-1}C_{12}} \tag{4.461}$$

$$k = x, y$$

From (4.460) and (4.461), the total static capacitance per unit length C_{ik} of line i for mode k can be obtained as

$$C_{1k} = C_{11} + R_{2k}C_{12} \tag{4.462}$$

$$C_{2k} = C_{22} + R_{2k}^{-1}C_{12} \tag{4.463}$$

For the case of lossless coupled lines, $\gamma_{x,y} = j\beta_{x,y}$, and the hyperbolic functions can be replaced with trigonometric functions; that is, $\coth \gamma_{x,y}l = -j\cot\theta_{x,y}$ and $\sinh \gamma_{x,y}l = j\sin\theta_{x,y}$, where $\theta_{x,y} = \beta_{x,y}l$ are the electrical length of the lines for mode x or y. The impedance and admittance parameters for the two lossless coupled lines are found from (4.447) and (4.448) as

$$Z_{11} = Z_{33} = -j\left[\frac{Z_{o1x}\cot\theta_x}{(1 - R_{2x}/R_{2y})} + \frac{Z_{o1y}\cot\theta_y}{(1 - R_{2y}/R_{2x})}\right]$$

$$Z_{12} = Z_{21} = Z_{34} = Z_{43} = -j\left[\frac{Z_{o1x}R_{2x}\cot\theta_x}{(1 - R_{2x}/R_{2y})} + \frac{Z_{o1y}R_{2y}\cot\theta_y}{(1 - R_{2y}/R_{2x})}\right]$$

$$= j\left[\frac{Z_{o2x}\cot\theta_x}{R_{2y}(1 - R_{2x}/R_{2y})} + \frac{Z_{o2y}\cot\theta_y}{R_{2x}(1 - R_{2y}/R_{2x})}\right]$$

$$Z_{14} = Z_{41} = Z_{23} = Z_{32} = -j\left[\frac{R_{2x}Z_{o1x}}{(1 - R_{2x}/R_{2y})\sin\theta_x} + \frac{R_{2y}Z_{o1y}}{(1 - R_{2y}/R_{2x})\sin\theta_y}\right]$$

$$Z_{13} = Z_{31} = -j\left[\frac{Z_{o1x}}{(1 - R_{2x}/R_{2y})\sin\theta_x} + \frac{Z_{o1y}}{(1 - R_{2y}/R_{2x})\sin\theta_y}\right]$$

$$
\begin{aligned}
Z_{22} = Z_{31} &= -j\left[\frac{Z_{o1x}}{\left(1-R_{2x}/R_{2y}\right)\sin\theta_x} + \frac{Z_{o1y}}{(1-R_{2y}/R_{2x})\sin\theta_y}\right] \\
Z_{22} = Z_{44} &= j\left[\frac{R_{2x}Z_{o2x}\cot\theta_x}{R_{2y}\left(1-R_{2x}/R_{2y}\right)} + \frac{R_{2y}Z_{o2y}\cot\theta_y}{R_{2x}(1-R_{2y}/R_{2x})}\right] \\
&= -j\frac{R_{2x}^2 Z_{o1x}\cot\theta_x}{(1-R_{2x}/R_{2y})} + \frac{R_{2y}^2 Z_{o1y}\cot\theta_y}{(1-R_{2y}/R_{2x})} \\
Z_{24} = Z_{42} &= -j\left[\frac{R_{2x}^2 Z_{o1x}}{\left(1-R_{2x}/R_{2y}\right)\sin\theta_x} + \frac{R_{2y}^2 Z_{o1y}}{(1-R_{2y}/R_{2x})\sin\theta_y}\right]
\end{aligned}
\tag{4.464}
$$

and

$$
\begin{aligned}
Y_{11} = Y_{33} &= -j\left[\frac{Y_{o1x}\cot\theta_x}{\left(1-R_{2x}/R_{2y}\right)} + \frac{Y_{o1y}\cot\theta_y}{(1-R_{2y}/R_{2x})}\right] \\
Y_{12} = Y_{21} = Y_{34} = Y_{43} &= j\left[\frac{Y_{o1x}\cot\theta_x}{R_{2y}\left(1-R_{2x}/R_{2y}\right)} + \frac{Y_{o1y}\cot\theta_y}{R_{2x}(1-R_{2y}/R_{2x})}\right] \\
Y_{14} = Y_{41} = Y_{23} = Y_{32} &= -j\left[\frac{Y_{o1x}}{\left(R_{2y}-R_{2x}\right)\sin\theta_x} + \frac{Y_{o1y}}{(R_{2x}-R_{2y})\sin\theta_y}\right] \\
Y_{13} = Y_{31} &= j\left[\frac{Y_{o1x}}{\left(1-R_{2x}/R_{2y}\right)\sin\theta_x} + \frac{Y_{o1y}}{(1-R_{2y}/R_{2x})\sin\theta_y}\right] \\
Y_{22} = Y_{44} &= +j\left[\frac{R_{2x}Y_{o2x}\cot\theta_x}{R_{2y}\left(1-R_{2x}/R_{2y}\right)} + \frac{R_{2y}Y_{o2y}\cot\theta_y}{R_{2x}(1-R_{2y}/R_{2x})}\right] \\
Y_{24} = Y_{42} &= -j\left[\frac{R_{2x}Y_{o2x}}{R_{2y}\left(1-R_{2x}/R_{2y}\right)\sin\theta_x} + \frac{R_{2y}Y_{o2y}}{R_{2x}(1-R_{2y}/R_{2x})\sin\theta_y}\right]
\end{aligned}
\tag{4.465}
$$

The results obtained in this section are indeed generalized results of those for two parallel-coupled transmission lines where the even-and odd-mode analysis, as described in Section 8.2.2.2, is applied. The equations for the impedance and admittance matrices can be used to derive specific results for two parallel-coupled lines that have specific port terminations such as those used in the well-known parallel-coupled band-pass filters discussed in Section 8.5.6.2.

Two Symmetrical Coupled Lines

For two symmetrical coupled lines, the two conductors in Figure 4.38 are identical, leading to

$$
\begin{aligned}
Z_{11} &= Z_{22} \\
Y_{11} &= Y_{22}
\end{aligned}
\tag{4.466}
$$

and the voltage ratio becomes

$$
R_{2x,y} = \pm 1 \tag{4.467}
$$

which is the well-known result. The resulting two propagation modes are known as the even (e) mode and odd (o) mode. The two propagation constants can be found from (4.438) as

$$
\gamma_{e,o} = \left[(Y_{11} \pm Y_{12})(Z_{11} \pm Z_{12})\right]^{1/2} \tag{4.468}
$$

where the subscripts e and o denote the even and odd modes, respectively. The even- and odd-mode characteristic impedances can be easily derived as

$$Z_{o1e} = Z_{o2e} = Z_{oe} = \frac{\gamma_e}{Y_{11} + Y_{12}} \tag{4.469}$$

and

$$Z_{o1o} = Z_{o2o} = Z_{oo} = \frac{\gamma_o}{Y_{11} - Y_{12}} \tag{4.470}$$

respectively. The elements of the impedance and admittance matrices become

$$\begin{aligned}
Z_{11} = Z_{22} = Z_{33} = Z_{44} &= \frac{1}{2}(Z_{oe}\coth\gamma_e l + Z_{oo}\coth\gamma_o l) \\
Z_{12} = Z_{21} = Z_{34} = Z_{43} &= \frac{1}{2}(Z_{oe}\coth\gamma_e l - Z_{oo}\coth\gamma_o l) \\
Z_{14} = Z_{41} = Z_{23} = Z_{32} &= \frac{1}{2}(Z_{oe}\mathrm{csch}\gamma_e l - Z_{oo}\mathrm{csch}\gamma_o l) \\
Z_{13} = Z_{31} = Z_{24} = Z_{42} &= \frac{1}{2}(Z_{oe}\mathrm{csch}\gamma_e l + Z_{oo}\mathrm{csch}\gamma_o l)
\end{aligned} \tag{4.471}$$

and

$$\begin{aligned}
Y_{11} = Y_{22} = Y_{33} = Y_{44} &= \frac{1}{2}(Y_{oe}\coth\gamma_e 1 + Y_{oo}\coth\gamma_o 1) \\
Y_{12} = Y_{21} = Y_{34} = Y_{43} &= \frac{1}{2}(Y_{oe}\coth\gamma_e 1 - Y_{oo}\coth\gamma_o 1) \\
Y_{14} = Y_{41} = Y_{23} = Y_{32} &= -\frac{1}{2}(Y_{oe}\mathrm{csch}\gamma_e 1 - Y_{oo}\mathrm{csch}\gamma_o 1) \\
Y_{13} = Y_{31} = Y_{24} = Y_{42} &= -\frac{1}{2}(Y_{oe}\mathrm{csch}\gamma_e 1 + Y_{oo}\mathrm{csch}\gamma_o 1)
\end{aligned} \tag{4.472}$$

For two symmetrical lossless coupled lines,

$$\begin{aligned}
L_{11} &= L_{22} \\
C_{11} &= C_{22} \\
C_{o11} &= C_{o22}
\end{aligned} \tag{4.473}$$

and the even- and odd-mode phase velocities and characteristic impedances become

$$v_{e,o} = (C_{e,o}L_{e,o})^{1/2} = v_o\left(\frac{C_{oe,o}}{C_{e,o}}\right)^{1/2} \tag{4.474}$$

$$Z_{oe,o} = \frac{1}{v_{e,o}C_{e,o}} \tag{4.475}$$

where

$$C_{e,o} = C_{11} \pm C_{12} \tag{4.476}$$

and

$$L_{e,o} = L_{11} \pm L_{12} \tag{4.477}$$

are the even (odd) mode static capacitance and inductance per unit length, and $C_{oe,o}$ is $C_{e,o}$ with air dielectric.

The resulting expressions for the four-port coupled-line parameters are the same as those in [24] for an inhomogeneous medium. They are:

$$
\begin{aligned}
Z_{11} = Z_{22} = Z_{33} = Z_{44} &= -j\frac{1}{2}(Y_{oe}\cot\theta_e + Z_{oo}\cot\theta_o) \\
Z_{12} = Z_{21} = Z_{34} = Z_{43} &= -j\frac{1}{2}(Z_{oe}\cot\theta_e - Z_{oo}\cot\theta_o) \\
Z_{14} = Z_{41} = Z_{23} = Z_{32} &= -j\frac{1}{2}(Z_{oe}\csc\theta_e - Z_{oo}\csc\theta_o) \\
Z_{13} = Z_{31} = Z_{24} = Z_{42} &= -j\frac{1}{2}(Z_{oe}\csc\theta_e + Z_{oo}\csc\theta_o)
\end{aligned} \tag{4.478}
$$

and

$$
\begin{aligned}
Y_{11} = Y_{22} = Y_{33} = Y_{44} &= -j\frac{1}{2}(Y_{oe}\cot\theta_e + Y_{oo}\cot\theta_o) \\
Y_{12} = Y_{21} = Y_{34} = Y_{43} &= -j\frac{1}{2}(Y_{oe}\cot\theta_e - Y_{oo}\cot\theta_o) \\
Y_{14} = Y_{41} = Y_{23} = Y_{32} &= -j\frac{1}{2}(Y_{oe}\csc\theta_e - Y_{oo}\csc\theta_o) \\
Y_{13} = Y_{31} = Y_{24} = Y_{42} &= -j\frac{1}{2}(Y_{oe}\csc\theta_e + Y_{oo}\csc\theta_o)
\end{aligned} \tag{4.479}
$$

where $\theta_{e,o} = \beta_{e,o}l$ is the electrical length of the lines for the even (odd) mode.

REFERENCES

1. C. R. Paul, K. W. Whites, and S. A. Nasar, Introduction to Electromagnetic Fields, Third Edition, McGraw-Hill, New York, 1998, Chapter 7.
2. D. D. Grieg and H. F. Englemann, "Microstrip – A New Transmission Technique for the Kilomegacycle Range," *Proc. IRE*, Vol. 40, pp. 1644–1650, Dec. 1952.
3. T. C. Edwards, Foundations for Microstrip Circuit Design, John Wiley & Sons, Inc., New York, 1984, pp. 44–45.
4. K. C. Gupta, R. Garg, and I. J. Bahl, Microstrip Lines and Slotlines, pp. 89–90, Artech House, Dedham, Massachusetts, 1979.
5. E. O. Hammerstad and O. Jensen, "Accurate Models for Microstrip Computer-Aided Design," *IEEE MTT-S Digest*, pp. 407–409, 1980.
6. R. A. Pucel, D. J. Masse, and C. P. Hartwig, "Losses in Microstrip," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-16, pp. 342–350, Jun. 1968.
7. R. A. Pucel, D. J. Masse, and C. P. Hartwig, "Correction to Losses in Microstrip," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-16, p. 1064, Dec. 1968.
8. J. D. Welch and H. J. Pratt, "Losses in Microstrip Transmission Systems for Integrated Microwave Circuits," *NEREM Rec.*, Vol. 8, pp. 100–101, 1966.
9. C. P. Wen, "Coplanar Waveguide: A Surface Strip Transmission Line Suitable for Non-Reciprocal Gyromagnetic Device Application," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-17, pp. 1087–1090, Dec. 1969.
10. G. Ghione and C. Naldi, "Analytical Formulas for Coplanar Lines in Hybrid and Monolithic MICs," *Electron. Lett.*, Vol. 20, No. 4, pp. 179–181, Feb. 1984.
11. G. Ghione and C. Naldi, "Parameters of Coplanar Waveguides with Lower Ground Plane," *Electron. Lett.*, Vol. 19, No. 18, pp. 734–735, Sept. 1983.
12. W. Hilberg, "From Approximation to Exact Relations for Characteristic Impedances," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-13, pp. 29–38, Jan. 1965.

13. K. C. Gupta, R. Garg, and I. J. Bahl, Microstrip Lines and Slotlines, Artech House, Dedham, MA, pp. 277–280, 1979.
14. K. C. Gupta, R. Garg, and R. Chadha, Computer-Aided Design of Microwave Circuits, p. 72, Artech House, Dedham, Massachusetts, 1981.
15. S. B. Cohn, "Problems in Strip Transmission Lines," *IRE Trans. Microwave Theory Tech.*, Vol. MTT-2, pp. 52–55, Jul. 1954.
16. H. A. Wheeler, "Transmission Line Properties of a Stripline Between Parallel Planes," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-26, pp. 866–876, Nov. 1978.
17. R. Garg and K. C. Gupta, "Expressions for Wavelength and Impedance of Slotline," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-24, p. 532, Aug. 1976.
18. S. B. Cohn, "Slot Line on a Dielectric Substrate," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-17, pp. 768–778, Oct. 1969.
19. R. Janaswamy and D. H. Schaubert, "Characteristic Impedance of a Wide Slotline on Low-Permittivity Substrates," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-34, pp. 900–902, Aug. 1986.
20. C. H. Ho, L. Fan, and K. Chang, "Broad-Band Uniplanar Hybrid-Ring and Branch-Line Couplers," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-41, pp. 2116–2125, Dec. 1993.
21. S. A. Schelkunoff, "Conversion of Maxwell's Equations into Generalized Telegraphists' Equations," *Bell Syst. Tech. J.*, Vol. 34, pp. 995–1043, Sept. 1955.
22. F. Ayres, Jr., Matrices, McGraw-Hill, New York, New York, 1962.
23. V. K. Tripathi, "Asymmetric Coupled Transmission lines in an Inhomogeneous Medium," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-23, pp. 734–739, Sept. 1975.
24. G. I. Zysman and A. K. Johnson, "Coupled Transmission Line Networks in an Inhomogeneous Dielectric Medium," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-17, pp. 753–759, Oct. 1969.

PROBLEMS

4.1 As discussed in Section 4.1 and illustrated in Figure 4.1, the signal entering an interconnect is $v_S(t) = V_S \cos \omega t$. Assume $V_S = 10$ V. Plot $v_S(t)$ at $t = t_o$, where t_o is the time it takes for the signal to reach the load, as a function of ℓ / λ, where λ is the wavelength and ℓ is the length of the interconnect. Based on the results, what do you think the limit of an interconnect length with respect to wavelength, above which an interconnect should be treated as a transmission line, is for most engineering applications?

4.2 Consider a cascade of N different transmission lines. Prove that the total propagation constant is the sum of the propagation constants of individual lines.

4.3 Consider a lossy transmission line terminated by a load impedance. Prove that, when the transmission line is very long, the input impedance to the line is approximately equal to the characteristic impedance. This result indicates that a load impedance different from the characteristic impedance is matched to a lossy line when the line's length is very large.

4.4 Derive the per-unit-length R, L, G and C of a parallel-plate transmission line shown in Figure P4.1. Assume the dielectric constant of the substrate between the conducting plates is given as $\hat{\varepsilon} = \varepsilon' - \varepsilon''$, the plates have conductivity σ, and W is very large compared to d. These parameters can also be used to model metal–insulator–metal (MIM) capacitors discussed in Chapter 3.

4.5 Show that the magnitude of the reflection coefficient in a lossless transmission line terminated with a purely reactive load is always 1.

4.6 Consider an (inhomogeneous) transmission line with the effective relative dielectric constant $\varepsilon_{\text{reff}} = a\varepsilon_r + b\omega$, where ε_r is the dielectric's relative dielectric constant, ω is the frequency, and a and b are constants. Derive expressions for the phase and group velocities of the quasi-TEM wave propagating along the line.

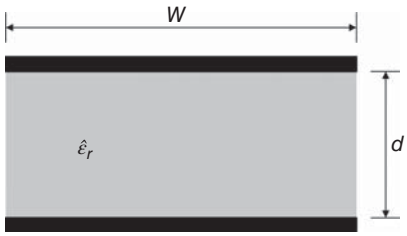


Figure P4.1.

4.7 Consider a 10-m long lossy coaxial cable having $R = 2.25\ \Omega/\text{m}$, $L = 1\ ^{-}\text{H/m}$, $C = 100\ \text{pF/m}$, and $G \simeq 0$ at 2 GHz. Assume the load impedance is 10 Ω. Determine the following parameters at 2 GHz:

a) Attenuation constant.
b) Attenuation per 1 m of the cable length in decibels (dB).
c) Reflection coefficient at the load.
d) Reflection coefficient at the input of the line (looking toward the load).

4.8 Consider a lossless quarter-wavelength transmission line with characteristic impedance Z_o terminated with a real impedance R_L. Determine the input impedance of the line and comment on possible use of quarter-wavelength transmission lines in circuit design.

4.9 Consider a 100-Ω microstrip line having an effective relative dielectric constant of 2.2.

a) Determine the length of the microstrip line with one end open-circuited such as it behaves as a capacitor of 1 pF at 30 GHz.
b) Determine the length of the microstrip line with one end short-circuited such as it behaves as an inductor of 1 nH at 30 GHz.

4.10 Consider an RF pulse train as shown in Figure P4.2, with each pulse being a half of the cosine waveform, propagating into the input of the microstrip line given in Problem 4.24 with $h = 8\ \mu\text{m}$ and $W = 15\ \mu\text{m}$. Assume the microstrip line is terminated with a load equal to the real part of its characteristic impedance at 50 GHz and is used in an RFIC operating across 0.1–100 GHz.

a) Compute and plot the voltage waveform of the signal arriving at the load (output signal) on the same figure of the input signal's voltage waveform for 500-μm and 3-mm lengths. Comment and provide rationale on the results concerning attenuation, delayed time and distortion.
b) Assume the microstrip line used is lossless and dispersion-less with a frequency-independent effective dielectric constant equal to its quasi-static value. Compute and plot the output signal waveform on the same figure of the input signal waveform for both 500-μm and 3-mm lengths. Comment and provide rationale on the results and compare to those in Part a.

4.11 Repeat Part a of Problem 4.10 for a microstrip line with the conducting strip and ground plane on the top and bottom of a silicon substrate, respectively. The strip and ground plane are copper with 1-μm thickness and conductivity of 5.8×10^7 mhos/m and has a width of 10 and 100 μm, respectively. The silicon substrate $h = 100\ \mu\text{m}$, $\varepsilon_r = 11.7$, and $\rho = 0.5\ \Omega\text{-cm}$. The microstrip line is still terminated with

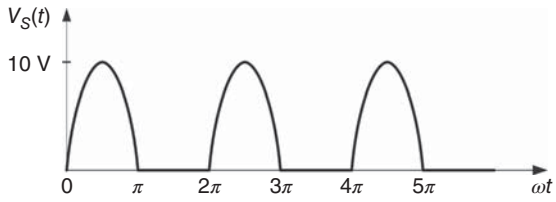


Figure P4.2.

a load equal to the real part of its characteristic impedance at 50 GHz and is used in an RFIC operating across 0.1–100 GHz.

4.12 Repeat Part a of Problem 4.10 for the conductor-backed asymmetric CPW of Problem 4.30 with $G = 10$ μm and $S = 20$ μm. Assume the CPW is terminated with a load equal to the real part of its characteristic impedance at 2 GHz and has a length of 2 mm, and the RF pulse train is at 2 GHz.

4.13 Consider a CPW (transmission line) having $R = 2\ \Omega/\text{m}$, $L = 2.5\ \mu\text{H/m}$, $C = 250\ \text{pF/m}$, and $G = 1\ \mu\text{S/m}$ at 2.45 GHz. Determine the following parameters at 2.45 GHz:

a) Characteristic impedance Z_o and wavelength λ_g.
b) Attenuation constant in dB/m.
c) Reflection coefficient at the load $Z_L = (40 - j20)\ \Omega$.
d) Reflection coefficient at the input of the transmission line, whose length is 2 mm, terminated with the load $Z_L = (40 - j20)\ \Omega$.

4.14 We want to modify the transmission line considered in Problem 4.13 so that it is "distortion-less" by periodically loading it with inductors at 1-mm intervals. Calculate the value of the added inductors and the resultant value for the characteristic impedance of the distortion-less transmission line. We assume the inductors added have negligible resistance.

4.15 Repeat Problem 4.14 using lossless inductors loaded at 500-μm intervals. Compare and comment on the results to those in Problem 4.14.

4.16 We consider a CPS with an effective relative dielectric constant of $\hat{\varepsilon}_{\text{reff}} = 6.8 - j0.003$ at 40 GHz. The conductor of the transmission line is copper having conductivity $\sigma = 5.8 \times 10^7\ \text{S/m}$. Calculate the following parameters at 40 GHz:

a) Dielectric attenuation constant.
b) Phase constant.
c) Skin depth of copper.
d) Surface resistance R_s.
e) Surface inductance L_s.

4.17 The total capacitance per unit length of the microstrip line shown in Figure P4.3 can be obtained using a conformal mapping technique, assuming $t = 0$, as

$$C = 2\varepsilon_o\varepsilon_r \frac{K(k')}{K(k)} + 2\varepsilon_o \frac{K(k'_o)}{K(k_o)}$$

where $K(k)$ is the complete elliptic integral of the first kind and

$$k = \text{sech}\left(\frac{\pi a}{2h}\right)$$

$$k' = \sqrt{1 - k^2}$$

Figure P4.3.

$$k_o = k(h = h_o)$$

$$k'_o = k'(h = h_o)$$

a) Derive the following equations:

$$\varepsilon_{\text{reff}} = \frac{\varepsilon_r \frac{K(k')}{K(k)} + \frac{K(k'_o)}{K(k_o)}}{\frac{K(k')}{K(k)} + \frac{K(k'_o)}{K(k_o)}}$$

$$Z_o = \frac{60\pi}{\sqrt{\varepsilon_{\text{reff}}}\left[\frac{K(k')}{K(k)} + \frac{K(k'_o)}{K(k_o)}\right]}$$

b) What is $\varepsilon_{\text{reff}}$ when $h = h_o$?

4.18 Design a synthetic transmission line having characteristic impedance Z_o of 50 Ω and cut-off frequency f_c of 15 GHz. Plot the input impedance Z_{in} (both real and imaginary parts or magnitude and phase) and transmission coefficient (e.g., S_{21}), reflection coefficient (e.g., S_{11}) (both magnitude and phase in dB and degrees, respectively) from DC-20 GHz for $N = 3, 8, 10, 20$ and 60, assuming 50-Ω system ($Z_L = 50\,\Omega$). The operating bandwidth of the designed synthetic transmission line can be determined based on several criteria; list these possible criteria along with your rationales for the bandwidths and provide the corresponding bandwidths that you determine. What do you think is the most suitable criterion for determining the bandwidth? (Note: You can write your own program or use any commercially available CAD program for plotting these parameters.)

4.19 Repeat Problem 4.18 for $Z_o = 50\ \Omega$ and $f_c = 30$ GHz. Plot Z_{in}, S_{21} and S_{11} from DC-40 GHz, assuming 50-Ω system. On the basis of the results of Problems 4.18 and 4.19, also comment on possible approximation of the upper operating frequency with respect to the cutoff frequency considering the number of sections in synthetic transmission lines.

4.20 A synthetic transmission line can be realized with series lumped-element inductors with inductance L and shunt capacitors with capacitance C, semiconductor devices, or a combination of both lumped elements and semiconductor devices – for example, using series inductors and shunt reversed-bias Schottky diodes simulating voltage-variable capacitors. In practice, inductors, capacitors, and semiconductor devices have their own parasitics, generally consisting of resistors, inductors, and capacitors, which may severely degrade the transmission-line performance, especially at high RF frequencies. Assume we design a 10-section 50-Ω synthetic transmission line using lumped-element series inductors L and shunt capacitors C, and theoretically obtain $L = 50$ ¯H and $C = 20$ nF. In order to realize this transmission line, we use (real) inductors and capacitors. The inductors have $L = 50$ ¯H with parasitic (series) resistance of 3 Ω and (shunt) capacitance of 3 pF. The capacitors have $C = 20$ nF with parasitic resistance of 2 Ω and inductance of 100 pH. One such section of the designed synthetic transmission line is shown in Figure P4.4.

a) Plot the transmission coefficient S_{21} (magnitude and phase in dB and degrees, respectively) from DC to $5f_c$, where f_c is the cutoff frequency, assuming no parasitics for all the series inductors L and shunt capacitors C.

b) Repeat Part a, but considering all the parasitics for L and C. Compare the results to those in Part a and explain the difference. Comment on the use of the designed synthetic transmission line with all the parasitics considered.

4.21 Repeat Problem 4.20 but with different resistive parasitics ($R1 = R2 = 10\ \Omega$) for a single section of the synthetic transmission line. Comment on the results and provide your rationale as needed.

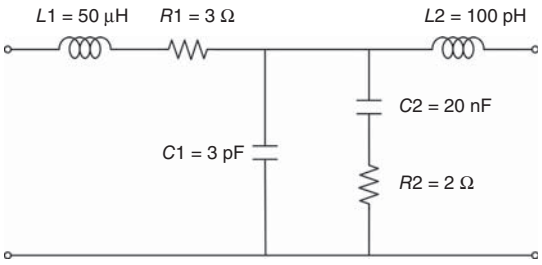


Figure P4.4.

4.22 a) Design a synthetic transmission line having characteristic impedance Z_o of 100 Ω and a cut-off frequency f_c of 20 GHz to achieve a total time delay of 0.10 ns using ideal lumped elements. Calculate the phase velocity and the equivalent physical length of each (LC) segment. Verify the designed time delay by calculating the voltage responses of this synthetic transmission line in time domain; that is, the input and output voltages.

b) Assume the constituent inductors and capacitors have the same parasitics as those in Problem 4.20, calculate the time delay from the voltage responses when considering all the parasitics. Is this calculated time delay meaningful at all frequencies up to $f_c = 20$ GHz? Provide your rationale.

c) Assume that the resistance, inductance, capacitance, and conductance of each segment used in Part b are distributed uniformly across each segment, determine the propagation constant and phase velocity. Calculate and plot the time needed for the signal to traverse the designed transmission line as a function of frequency.

4.23 We want to redesign the synthetic transmission line in Problem 4.21 to achieve distortion-less. Can this be done? If not, then provide your rationale. If yes, then design this distortion-less synthetic transmission line. Is your designed synthetic transmission line valid at a single frequency or across a frequency range? If it can be used only for a single frequency, then redesign it to work over a frequency bandwidth since, for transmission lines to be useful, they need to pass signals over a certain bandwidth. What is the bandwidth of your designed transmission line?

4.24 Consider a microstrip line used in RFIC. The microstrip line, as shown in Figure P4.5, is fabricated on silicon substrate with SiO_2, having relative dielectric constant ε_r of 3.9 and loss tangent $\tan\delta$ of 0.0002, as its dielectric medium and copper, having thickness of 1 μm and conductivity σ of 5.8×10^7 mhos/m, as its top and bottom conductors. The bottom conductor is assumed to be infinitely large and thick and serves as a shield for the microstrip line from the Si substrate. Using closed-form equations provided in this chapter, calculate and plot the following parameters versus the microstrip width W of 1 μm and from 10 to 500 μm in 10-μm steps for SiO_2 thickness h of 5, 8, and 15 μm at 10 GHz: characteristic impedance (ohm), effective relative dielectric constant, conductor attenuation constant (dB/mm), dielectric attenuation constant (dB/mm), total attenuation constant (dB/mm), wavelength

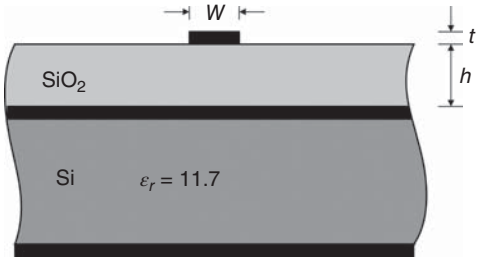


Figure P4.5.

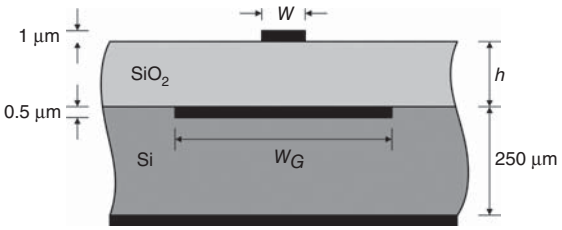


Figure P4.6.

(mm). Compare and comment on the characteristic impedance range and loss between $h = 5$, 8 and 15 μm.

4.25 Consider the microstrip line shown in Figure P4.6. The top conductor and the SiO_2 layer have the same properties as those in Problem 4.24. The bottom conductor, however, is different from that in Problem 4.24; it has a finite width of $W_G = 5$ W and a finite metallization thickness of 0.5 μm. The Si substrate has $\varepsilon_r = 11.7$ and resistivity $\rho = 0.5$ Ω-cm. Use an EM simulation program to do the following. You would need to consider the Si substrate in this problem.

a) Calculate and plot the following parameters versus frequency from 0 to 20 GHz in 2-GHz steps for the microstrip width of 10, 40, 70, and 100 μm for each of the SiO_2 thickness of 5, 8, and 15 μm: real and imaginary parts of characteristic impedance (ohm), real and imaginary parts of effective relative dielectric constant, conductor attenuation constant (dB/mm), dielectric attenuation constant (dB/mm), total attenuation constant (dB/mm), and wavelength (mm). Discuss the results.

b) Plot the parameters in Part a as a function of W from 10 to 100 μm in 10-μm steps for h of 8 μm at 10 GHz, and compare and comment on the results with those in Problem 4.24.

4.26 Consider a CPS used in RFIC. The CPS, as shown in Figure P4.7, is fabricated on silicon substrate with SiO_2, having relative dielectric constant ε_r of 3.9 and loss tangent $\tan\delta$ of 0.0002, as its dielectric medium and copper, having thickness of t and conductivity σ of 5.8×10^7 mhos/m, as the strip conductor. The width of the each of the two strips is $W = 10$ μm. Neglect the strip's metallization thickness and effects of the Si substrate (i.e., consider the CPS as there is only one dielectric below the conducting strips). Using closed-form equations provided in this chapter, calculate and plot the following parameters versus the strip gap S from 1 to 60 μm in 1-μm steps at 5 and 15 GHz: characteristic impedance (ohm), effective relative dielectric constant, conductor attenuation constant (dB/mm), dielectric attenuation constant (dB/mm), total attenuation constant (dB/mm), and wavelength (mm). Comment on the results.

4.27 Repeat Problem 4.26 with the strip's metallization thickness t of 0.5, 1, and 2 μm. Compare and comment on the results between different thickness and between Problems 4.26 and 4.27.

4.28 Repeat Problem 4.27 but using an EM simulation program. The Si substrate has $\varepsilon_r = 11.7$ and resistivity $\rho = 0.5$ Ω-cm. Using an EM simulation program and consider the Si substrate, calculate and plot the parameters versus frequency from 0 to 20 GHz in 2-GHz steps for S from 5 to 20 μm in 5-μm steps. You

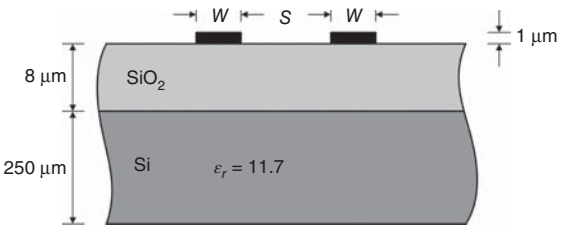


Figure P4.7.

would need to plot both the real and imaginary parts of Z_o and $\varepsilon_{\text{reff}}$. Plot the parameters as a function of S from 5 to 20 μm in 5-μm steps at 5 and 15 GHz, and compare and comment on the results with those in Problem 4.27.

4.29 In a conventional microstrip line, the field in the air for a small strip width (W) should be smaller than that for a wider width. That means that real part of the effective dielectric constant $\text{Re}(\varepsilon_{\text{reff}})$ for small W should be smaller than that for large W. In the limit of zero width, $\text{Re}(\varepsilon_{\text{reff}})$ should approach $(\varepsilon_r + 1)/2$, and in the other limit of infinite width, $\text{Re}(\varepsilon_{\text{reff}})$ should approach ε_r. The situation is, however, different as can be seen in the calculated results of $\text{Re}(\varepsilon_{\text{reff}})$ shown in Figure 4.8 for a possible CMOS microstrip line, which show that $\text{Re}(\varepsilon_{\text{reff}})$ reduces as W/h is increased when $W/h < 1$ approximately, while $\text{Re}(\varepsilon_{\text{reff}})$ increases as W/h is increased for $W/h > 1$ approximately. Explain why $\text{Re}(\varepsilon_{\text{reff}})$ behaves like that for such a microstrip structure.

4.30 Consider a conductor-backed asymmetric CPW shown in Figure P4.8. The top and bottom ground strips are connected through via-holes along the center of the top ground strip.

a) Using an EM simulator, calculate and plot the real and imaginary parts of characteristic impedance and effective dielectric constant, conductor attenuation constant (dB/mm), dielectric attenuation constant (dB/mm), total attenuation constant (dB/mm), and wavelength (mm) versus G from 5 to 15 μm in 2.5-μm steps for $S = 20$ and 40 μm, $W_S = 10$ μm, $W_{G1} = 50$ μm at 10 GHz. Plot the same parameters on the same plot. Discuss the results.

b) Using an EM simulator, calculate and plot the above parameters for $G = 10$ μm, $S = 20$ μm versus frequency from 0 to 16 GHz in steps of 2 GHz. Comment on the results.

4.31 There are two kinds of coupling between adjacent transmission lines in RFICs: one is electric coupling caused by the electric fields between the transmission lines and one is magnetic coupling caused by the magnetic fields. These coupling phenomena can be explained theoretically and can be visualized through field distributions among the transmission lines. Consider two separate (single) microstrip lines, running in parallel in RFIC, as shown in Figure P4.9, on silicon substrate with SiO_2, having relative dielectric constant ε_r of 3.9 and loss tangent $\tan\delta$ of 0.0002, as its dielectric medium and copper, having conductivity σ of 5.8×10^7 mhos/m, as its top and bottom conductors. The bottom conductor's width is finite. Using an EM simulator to plot separately the electric field and magnetic field distributions between the microstrip lines at 20 GHz and the coupling S_{41} (dB) between these two lines from 1 to 30 GHz (in 1-GHz steps) for $S = 0.21W$, $0.5W$, W, $2W$, and $3W$. Assume ports 2 and 3 are terminated with a load resistance equal to the real part (R_o) of the characteristic impedance of each line. Examine the field distributions and comment on the mechanism causing the electric and magnetic couplings as well as their behavior and contribution to the overall coupling magnitude. What would you do to reduce these couplings to improve your RFIC circuit both in performance and size?

4.32 Derive the per-unit-length R, L, G, and C of a parallel-plate transmission line as shown in Figure P4.1. Assume the dielectric between the plates is given as $\hat{\varepsilon} = \varepsilon' - \varepsilon''$, the conducting plates have conductivity σ, and W is very large compared to d. These parameters can be used to model MIM capacitors discussed in Chapter 3.

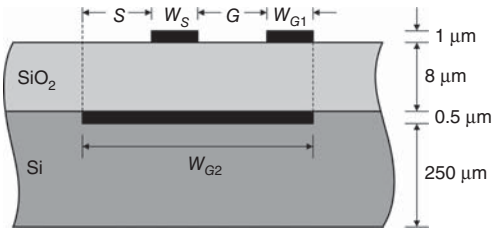


Figure P4.8. $W_{G2} = W_{G1} + W_S + G + S$.

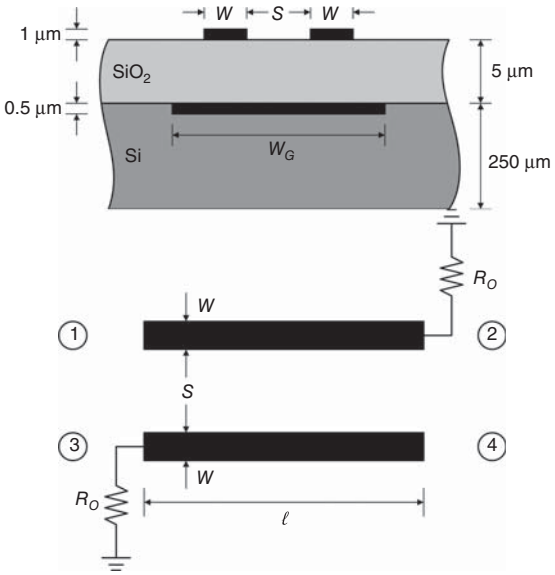


Figure P4.9. $W = 10$ μm, $\ell = 200$ μm. $W_G = 6W + S$.

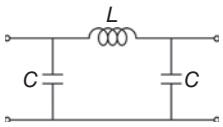


Figure P4.10.

4.33 A lossless transmission line, whose length is less than a quarter-wavelength, can be approximately modeled by a π-network as shown in Figure P4.10. Derive the following expressions for the network's elements:

$$C \simeq \frac{1}{\omega Z_o} \tan\left(\frac{\omega \ell}{2v_p}\right)$$

$$L \simeq \frac{Z_o}{\omega} \sin\left(\frac{\omega \ell}{v_p}\right)$$

where Z_o is the characteristic impedance, ω is the radian frequency, ℓ is the transmission line's length, and v_p is the phase velocity.

4.34 Derive Eqs. (4.356) and (4.357).

4.35 Derive the impedance matrix of n parallel-coupled transmission lines using (4.407)–(4.409).

4.36 Derive the admittance matrix of n parallel-coupled transmission lines using (4.407)–(4.409).

4.37 Derive the chain matrix of n parallel-coupled transmission lines using (4.407)–(4.409).

4.38 Derive (4.431)–(4.433).

4.39 Derive (4.438).

4.40 Derive (4.447).

4.41 Derive (4.448).

4.42 Derive (4.452) and (4.454).

APPENDIX 4: TRANSMISSION-LINE EQUATIONS DERIVED FROM MAXWELL'S EQUATIONS

In Section 4.2, we derive the transmission-line equations governing the voltage and current along transmission lines based on the circuit theory by considering a very short section of transmission lines. As expected, we can also derive these equations using Maxwell equations as follows.

To simplify the analysis without loss of generality, we consider a coaxial transmission line as shown in Figure A4.1. The length of the transmission line is along the z direction and assumed to be infinitely long. We also assume that the conductors are perfect, hence supporting the TEM mode of propagation, and the dielectric medium between the conductors is lossless. Due to the angular symmetry of the coaxial line, the fields is independent of the angle ϕ. This, along with the properties of the TEM mode, leads to

$$\frac{\partial E}{\partial \phi} = \frac{\partial H}{\partial \phi} = 0 \tag{A4.1}$$

$$E_z = H_z = 0 \tag{A4.2}$$

The fields in the coaxial line satisfy the following Maxwell's equations:

$$\nabla \times E = -\mu \frac{\partial H}{\partial t} \tag{A4.3}$$

and

$$\nabla \times H = \varepsilon \frac{\partial E}{\partial t} \tag{A4.4}$$

where μ and ε are the permeability and permittivity of the dielectric medium between the two conductors, and the following boundary conditions on the conductors:

$$E_z = E_\phi = 0 \quad \text{at } r = a \text{ and } b \tag{A4.5}$$

$$H_r = 0 \quad \text{at } r = a \text{ and } b \tag{A4.6}$$

Expanding Eq. (A4.3), we obtain

$$\frac{\partial E_z}{r\partial \phi} - \frac{\partial E_\phi}{\partial z} = -\mu \frac{\partial H_r}{\partial t} \tag{A4.7}$$

$$\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = -\mu \frac{\partial H_\phi}{\partial t} \tag{A4.8}$$

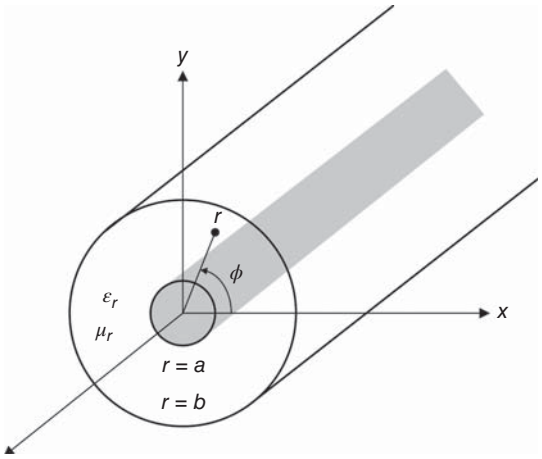


Figure A4.1. Coaxial line.

$$\frac{1}{r}\left[\frac{\partial}{\partial r}\left(rE_\phi\right) - \frac{\partial E_r}{\partial \phi}\right] = -\mu\frac{\partial H_z}{\partial t} \tag{A4.9}$$

Similarly, from (A4.4), we have

$$\frac{\partial H_z}{r\partial \phi} - \frac{\partial H_\phi}{\partial z} = \varepsilon\frac{\partial E_r}{\partial t} \tag{A4.10}$$

$$\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = \varepsilon\frac{\partial E_\phi}{\partial t} \tag{A4.11}$$

$$\frac{1}{r}\left[\frac{\partial}{\partial r}\left(rH_\phi\right) - \frac{\partial H_r}{\partial \phi}\right] = \varepsilon\frac{\partial E_z}{\partial t} \tag{A4.12}$$

Applying (A4.1) and (A4.2) to (A4.9) and (A4.12), we get

$$\frac{\partial}{\partial r}(rE_\phi) = 0 \tag{A4.13}$$

and

$$\frac{\partial}{\partial r}(rH_\phi) = 0 \tag{A4.14}$$

respectively. Equations (A4.13) and (A4.14) imply that rE_ϕ and rH_ϕ are not a function of r. Furthermore, since E_ϕ and H_ϕ are not a function of ϕ due the angular symmetry of the coaxial line, as mentioned earlier, they can be described in the following forms:

$$E_\phi = \frac{1}{r}f(z,t) \tag{A4.15}$$

$$H_\phi = \frac{1}{r}g(z,t) \tag{A4.16}$$

where $f(z,t)$ and $g(z,t)$ are functions of z and t only. The boundary condition (A4.5) requires that $f(z) = 0$, leading to $E_\phi = 0$ as well. Substituting $E_\phi = 0$ into (A4.7) and recognizing that $E_z = 0$, we then have H_r equal to zero or be independent of t. Since H_r is in general a function of t, we then conclude that $H_r = 0$. Substituting (A4.16) and (A4.2) into (A4.8), we can describe E_r as

$$E_r = \frac{1}{r}h(z,t) \tag{A4.17}$$

where $h(z,t)$ is a function of z and t. (A4.8) and (A4.11) now reduce to

$$\frac{\partial E_r}{\partial z} = -\mu\frac{\partial H_\phi}{\partial t} \tag{A4.18}$$

$$\frac{\partial H_\phi}{\partial z} = -\varepsilon\frac{\partial E_r}{\partial t} \tag{A4.19}$$

Substituting (A4.16) and (A4.17) into (A4.18) and (A4.19), we obtain

$$\frac{\partial h(z,t)}{\partial z} + \mu\frac{\partial g(z,t)}{\partial t} = 0 \tag{A4.20}$$

$$\frac{\partial g(z,t)}{\partial z} + \varepsilon \frac{\partial h(z,t)}{\partial t} = 0 \tag{A4.21}$$

We now define the voltage between the two conductors as

$$\begin{aligned} V(z,t) &= \int_a^b E_r(r,z,t)dr \\ &= \int_a^b \frac{1}{r} h(z,t)dr \\ &= h(z,t)\ln\left(\frac{b}{a}\right) \end{aligned} \tag{A4.22}$$

The current density on the inner conductor is given from the boundary condition as

$$\begin{aligned} J_s &= \hat{a}_r \times \vec{H} \\ &= H_\phi(a,z,t)\hat{a}_z \end{aligned} \tag{A4.23}$$

from which, the total current flowing on the inner conductor can be derived as

$$\begin{aligned} I(z,t) &= \int_0^{2\pi} |\vec{J}_s(a,z,t)| a d\phi \\ &= 2\pi a H_\phi(a,z,t) \\ &= 2\pi g(z,t) \end{aligned} \tag{A4.24}$$

Note that the current flow on the outer conductor has the same magnitude but with opposite direction with that on the inner conductor. Substituting $h(z,t)$ and $g(z,t)$ from (A4.22) and (A4.24), respectively, into (A4.20) and (A4.21), we get

$$\frac{\partial V(z,t)}{\partial z} + \frac{\mu \ln(b/a)}{2\pi} \frac{\partial I(z,t)}{\partial t} = 0 \tag{A4.25}$$

$$\frac{\partial I(z,t)}{\partial z} + \frac{2\pi\varepsilon}{\ln(b/a)} \frac{\partial V(z,t)}{\partial t} = 0 \tag{A4.26}$$

The inductance L and capacitance C per unit length of the coaxial transmission line are obtained from (4.118) and (4.119) as

$$L = \frac{\mu \ln(b/a)}{2\pi} \tag{A4.27}$$

$$C = \frac{2\pi\varepsilon}{\ln(b/a)} \tag{A4.28}$$

Substituting these in (A4.25) and (A4.26) gives

$$\frac{\partial V(z,t)}{\partial z} + L \frac{\partial I(z,t)}{\partial t} = 0 \tag{A4.29}$$

$$\frac{\partial I(z,t)}{\partial z} + C \frac{\partial V(z,t)}{\partial t} = 0 \tag{A4.30}$$

which are indeed the telegraphist's equations in (4.9) and (4.10). The transmission-line equations (4.11) and (4.12) can then be obtained from (A4.29) and (A4.30). The same telegraphist's and transmission-line equations as given in (4.3), (4.4) and (4.7), (4.8), respectively, can also be derived for any transmission lines with imperfect dielectric and conductors. When the conductors are imperfect, however, the equivalence between using lumped-element and field analyses is not exact. Nevertheless, for practical transmission lines with good conductors, the two approaches give results that are approximately close to each other.

CHAPTER 5

RESONATORS

Resonators are common and important components in radio frequency integrated circuits (RFICs). They are used in various circuits such as oscillators and filters. There are, in principle, two types of resonators: lumped-element and distributed resonators. In general, lumped-element resonators have a smaller size, but lower operating frequency and quality factor (Q) as compared to distributed resonators, which are made up of transmission lines or, in general, waveguide structures.[1] Advances in complementary metal oxide silicon (CMOS) technologies have made feasible good lumped-element resonators for RFICs. The compactness, particularly, makes lumped-element resonators popular in RFICs operating in the microwave regime. Lumped-element resonators also make their way into the millimeter-wave region. On the other hand, distributed resonators are attractive for RFIC design at millimeter-wave frequencies, particularly in the high millimeter-wave end, where lumped-element resonators either are not feasible or have lower quality than their distributed counterparts in current CMOS processes. In this chapter, we will present the analysis and design of both lumped-element and distributed resonators that can be implemented in the CMOS processes.

5.1 FUNDAMENTALS OF RESONATORS

Resonators or resonant circuits, as the name implies, are used in principle to store the energy of signals. As signals contain both electric and magnetic energies, a resonator is electrically represented by a combination of an inductor and a capacitor (for ideal resonators) together with a resistor (for nonideal resonators), which account for the magnetic energy, electric energy, and loss in the resonator, respectively. A resonator is also sometimes called "tank circuit" in oscillators. A resonator type is dictated by its equivalent electrical representation as series or parallel resonator. As such, a resonator – whether lumped, distributed, or a combination of lumped and distributed structures – may be classified as a series or parallel resonator depending on its electrical equivalent circuit – not the actual constituent elements themselves or the resonator's configuration. Also, a general circuit consisting of multiple inductors, capacitors, and distributed elements may behave as a parallel or series resonator at a certain frequency depending on its equivalent-circuit representation at

[1]Transmission lines are basically one class of waveguides or wave-guiding structures.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

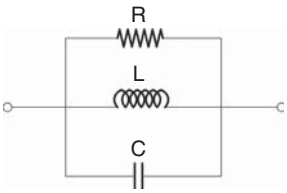


Figure 5.1. Parallel resonator consisting of R, L, and C in parallel.

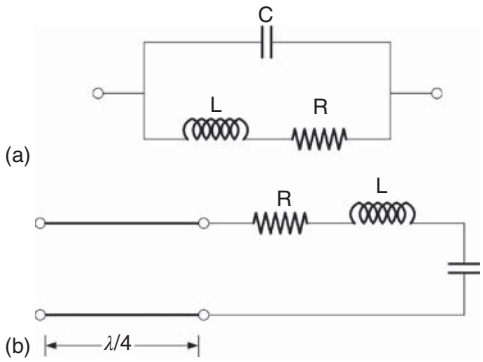


Figure 5.2. R, L, C lumped-element circuit (a) and combined lumped elements and transmission line (b) acting as parallel resonators.

that frequency. It is noted that practical inductors and capacitors contain parasitic resistance, inductance, and capacitance and can also be modeled as series or parallel RLC networks which behave as resonators at certain frequencies.

5.1.1 Parallel Resonators

Parallel resonators refer to lumped-element circuits, distributed circuits, or a combination of lumped and distributed elements, that can electrically be represented by R, L, and C in parallel or a combination of R, L, and C, which resembles parallel R, L, and C. It is noted that the resonant action is created by L and C, while R degrading the resonator's performance. Figure 5.1 shows the electrical equivalent circuit of parallel resonators. Figure 5.2 shows a lumped-element network and a combined network, consisting of lumped elements and a transmission line (distributed element), that behave as parallel resonators over a narrow frequency band around the resonant frequency.

Consider a parallel resonator as shown in Figure 5.3. The input impedance Z_{in} can be derived as [1]

$$Z_{\text{in}} = R_{\text{in}} + jX_{\text{in}} = \frac{P_L + 2j\omega(W_m - W_e)}{\frac{1}{2}II^*} \tag{5.1}$$

where P_L represents the time-average power loss (in the resistor), W_e is the average electric energy stored in the electric field given as

$$W_e = \frac{1}{4}|V|^2C \tag{5.2}$$

and W_m is the average magnetic energy stored in the magnetic field:

$$W_m = \frac{1}{4}LI_LI_L^* = \frac{1}{4}L\left|\frac{V}{\omega L}\right|^2 = \frac{1}{4\omega^2 L}|V|^2 \tag{5.3}$$

For good resonators, R is large and, hence, the input impedance behaves approximately as inductance or capacitance, depending on the frequency. There is a frequency at which X_{in} is equal to zero, resulting in a

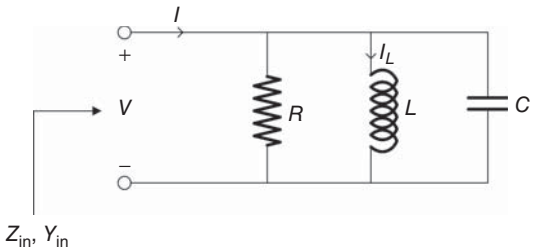


Figure 5.3. A parallel resonator.

resonant circuit. The resonance of the resonator thus corresponds to $X_{\text{in}} = 0$ or $X_{\text{in}} = R_{\text{in}}$, at which $W_e = W_m$. The resonant frequency can then be derived as

$$\omega_r = \frac{1}{\sqrt{LC}} \tag{5.4}$$

The same resonant frequency can also be derived by letting the imaginary part of the input admittance $Y_{\text{in}} = G_{\text{in}} + jB_{\text{in}}$ equal to zero. It is noted that for parallel resonators, which are not made up of actual R, L, and C connected in parallel, such as distributed resonators, L and C are used as their equivalent representatives derived from the resonator's parameters.

Now consider a small frequency range $\Delta\omega$ around the resonant frequency ω_r. Assume $\omega_r \gg \Delta\omega$. Using McLaurin series, the frequency $\omega = \omega_r + \Delta\omega$ can be approximately written as

$$\omega = \omega_r + \Delta\omega \simeq \frac{\omega_r}{1 - \dfrac{\Delta\omega}{\omega_r}} \tag{5.5}$$

and, hence, the input impedance can be derived in the vicinity of resonance as

$$Z_{\text{in}} \simeq \frac{\omega_r^2 RL}{\omega_r^2 L + j2R\Delta\omega} \tag{5.6}$$

It is easily seen that at frequencies below and above the resonant frequency, a parallel resonator behaves essentially like an inductor and capacitor, respectively.

5.1.2 Series Resonators

Series resonators refer to lumped-element circuits, distributed circuits, or a combination of lumped and distributed elements, that can electrically be represented by R, L, and C in series or a combination of R, L, and C, which approximates series R, L, and C. Figure 5.4 shows the electrical equivalent circuit of series resonators.

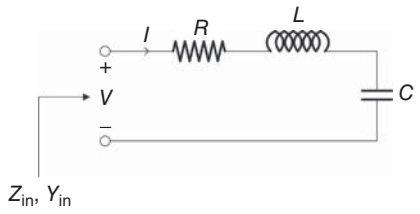


Figure 5.4. Series resonator consisting of R, L, and C in series.

As for parallel resonators, the input impedance Z_{in} can also be derived as

$$Z_{\text{in}} = R + jX_{\text{in}} = \frac{P_L + 2j\omega(W_m - W_e)}{\frac{1}{2}II^*} \tag{5.7}$$

At resonance, $Z_{\text{in}} = R$ and $W_e = W_m$, leading to the same resonant frequency as in (5.4):

$$\omega_r = \frac{1}{\sqrt{LC}} \tag{5.8}$$

Consider a small frequency range $\Delta\omega$ around the resonant frequency $\omega_r \gg \Delta\omega$. At $\omega = \omega_r + \Delta\omega$, the input impedance can be derived as

$$Z_{\text{in}} = R + j\omega L\left(\frac{\omega^2 - \omega_r^2}{\omega^2}\right) \tag{5.9}$$

which can be approximated using

$$\begin{aligned}\omega^2 - \omega_r^2 &= (\omega + \omega_r)(\omega - \omega_r) = (2\omega - \Delta\omega)\Delta\omega \\ &\simeq 2\omega\Delta\omega\end{aligned}$$

as

$$Z_{\text{in}} \simeq R + j2L\Delta\omega \tag{5.10}$$

At frequencies lower and higher than the resonant frequency, a series resonator functions as a capacitor and inductor, respectively.

5.2 QUALITY FACTOR

Practical resonators – whether lumped- or distributed-element – always have loss, which can be substantial at radio frequency (RF), particularly for those on RFICs, and degrades the resonator's quality. Quality factor (Q) of resonators is a figure of merit characterizing the loss of the resonator, and hence its usefulness as a circuit element. It is essentially the most important "metric" specifying how good a resonator is in storing the energy (both electric and magnetic) in the resonator and, effectively, dictates how a resonator would contribute to the performance of RFICs containing it.

Ideally, a resonator is used to store energy indefinitely with respect to time. However, the resonator's loss reduces this storage time and lowers the resonator's quality. Essentially, there are two important phenomena in resonators: one is energy storage and another one is energy dissipation. The quality of storing energy with respect to power dissipation or power removal gives what is known as quality factor. As the energy is being stored in a resonator, the energy reduces as time passes. As such, Q of resonators is conventionally defined as the ratio between the total time-average stored (electric and magnetic) energy and the energy loss per second (or time-average power loss or dissipated) within the resonators as[2]

$$\begin{aligned}Q &= \frac{\omega_r(\text{time-average energy stored})}{\text{energy loss per second}} \\ &= \frac{\omega_r W}{P_L}\end{aligned} \tag{5.11}$$

where W is the total average stored energy and P_L represents the power loss. The total energy is interchanged between the electric and magnetic energies stored in the electric and magnetic fields, respectively. At the

[2]This is the "unloaded" or "intrinsic" Q of resonators, which is different from the "loaded" Q.

instant when the electric field is maximum, the magnetic field is zero, and vice versa. As such, the total energy can be obtained as

$$W = W_e + W_m = 2W_e = 2W_m \tag{5.12}$$

where W_e and W_m are the time-average stored electric and magnetic energies, respectively.

It can be proved that the resistance representing the loss of resonators and their quality factor are proportional to each other and depend on the dissipated power. The ratio between this resistance and quality factor is, therefore, independent of the losses of the resonators and depends only on the dimensions and configurations of the resonators' geometry.

5.2.1 Parallel Resonators

For parallel resonators, as shown in Figure 5.3, the average electric energy stored in capacitors can be obtained as

$$W_e = \frac{1}{4} C V_{\text{rms}}^2 \tag{5.13}$$

and the power loss is

$$P_L = \frac{1}{2} \frac{V_{\text{rms}}^2}{R} \tag{5.14}$$

where V_{rms} represents the (rms) voltage across the capacitor. Q of parallel resonators can then be derived from (5.11), (5.13), and (5.14) as

$$Q = \omega_r RC = \frac{R}{\omega_r L} = R\sqrt{\frac{C}{L}} \tag{5.15}$$

Equation (5.15) indicates that Q of parallel resonators increases as resistance increases, which is expected, because the higher the resistance the closer the resonator to an ideal one. Q can also be increased by increasing the ratio between C and L. We obtain from (5.15)

$$\frac{R}{Q} = \sqrt{\frac{L}{C}} \tag{5.16}$$

which is independent of the resonator's loss due to the fact that L/C depends only on the dimensions and physical configuration of the resonator.

The input impedance Z_{in} given in Eq. (5.6) can be rewritten, using (5.15), as

$$Z_{\text{in}} \simeq \frac{R}{1 + j2Q\frac{\Delta\omega}{\omega_r}} = \frac{R}{1 + j2\Delta\omega RC} \tag{5.17}$$

which shows that the maximum magnitude of Z_{in} is equal to R at resonance. Figure 5.5 shows a sketch of the input impedance magnitude versus $\Delta\omega/\omega_r$. Taking the magnitude of Z_{in} and equating it to $0.707R$ leads to

$$0.707\left[1 + 4Q^2\left(\frac{\Delta\omega}{\omega_r}\right)^2\right]^{1/2} = 1 \tag{5.18}$$

from which, we can obtain

$$Q = \frac{\omega_r}{2\Delta\omega} = \frac{f_r}{2\Delta f} \tag{5.19}$$

where f_r and $2\Delta f$ are the resonant frequency and 3-dB bandwidth of the input impedance, respectively. As can be seen in (5.19), Q is equal to the reciprocal of the 3-dB fractional bandwidth of the input impedance

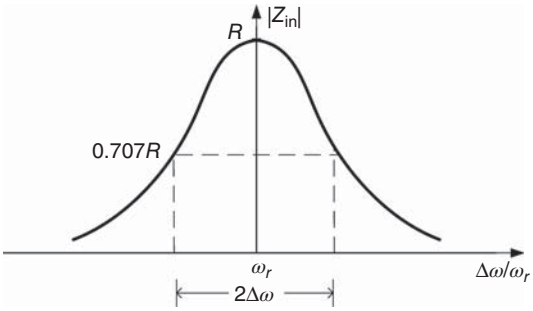


Figure 5.5. Sketch of Z_{in} versus $\Delta\omega/\omega_r$.

around the resonant frequency. Equation (5.19) suggests that the Q of a parallel resonator can be determined as the ratio between the resonant frequency and 3-dB bandwidth obtained from the calculated or measured magnitude of the input impedance versus frequency. Q is thus inversely proportional to the bandwidth, implying that a narrow-band performance corresponds to a high Q. High Q resonators thus result in high frequency-resolution, which are desirable in circuits such as filters and oscillators, and are needed for narrow-band circuits. It is noted that the foregoing input impedance is the self-impedance of the one-port resonator. For resonators configured as a two-port network, Eq. (5.19) still applies provided that one port of resonators is terminated and the corresponding input impedance is used. In this case, however, the determined Q is not the unloaded Q of the original resonator but is the Q of the resultant (one-port) network consisting of the original resonator terminated with an impedance at one port. This Q is indeed the loaded Q of the original resonator with the loading being the terminating impedance.

Now consider a parallel resonator represented by its input or self admittance Y_{in} or impedance Z_{in} and connect this resonator across a transmission line with characteristic admittance Y_o or impedance Z_o. The scattering parameters of the resultant two-port network can be derived as

$$S_{11} = S_{22} = -\frac{Y_{\text{in}}}{Y_{\text{in}} + 2Y_o} \tag{5.20}$$

and

$$S_{21} = S_{12} = \frac{2Y_o}{Y_{\text{in}} + 2Y_o} \tag{5.21}$$

From (5.20) and (5.21), we can obtain

$$Z_{\text{in}} = Z_o \frac{1 - S_{11}}{2S_{11}} = Z_o \frac{S_{21}}{2(1 - S_{21})} \tag{5.22}$$

which implies that the quality factor of the resonator, given in (5.19), can also be determined from S_{11} or S_{21} (both magnitude and phase) of the two-port circuit consisting of the transmission line and the resonator connecting across it. It is particularly noted that the return loss ($|S_{11}|^2$) or insertion loss ($|S_{21}|^2$) versus frequency of this circuit is not sufficient for accurate determination of the unloaded Q, although either one can be used to assess it qualitatively.

The quality factor can also be estimated from the time-domain or transient response of resonators. Consider a parallel resonator shown in Figure 5.3. The frequency response of this resonator, using Laplace transform, is given as

$$H(s = j\omega) = \frac{s}{C(s - s_1)(s - s_2)} \tag{5.23}$$

where

$$s_{1(2)} = -\frac{1}{2RC} \pm \frac{1}{2}\sqrt{\frac{L - 4R^2C}{R^2C^2L}} \tag{5.24}$$

with the + and − sign corresponding to s_1 and s_2, respectively. The impulse response can be obtained from (5.23) as

$$h(t) = \ell^{-1}[H(s)] = \ell^{-1}\left[\frac{s}{C(s-s_1)(s-s_2)}\right] \tag{5.25}$$

Now we consider three cases:

1. $L - 4R^2C = 0$
   Under this condition, we get
   $$H(s) = \frac{s}{C\left(s + \frac{1}{2RC}\right)^2} \tag{5.26}$$
   and
   $$h(t) = \frac{1}{C}\left(1 - \frac{t}{2RC}\right)e^{-t/2RC} \tag{5.27}$$
   from (5.23)–(5.25).
2. $L - 4R^2C > 0$
   Under this condition, the impulse response can be derived as
   $$h(t) = \frac{1}{2A}e^{-\frac{1}{2RC}t}\left[A\left(e^{\frac{A}{2}t} + e^{-\frac{A}{2}t}\right) - \frac{1}{RC}\left(e^{\frac{A}{2}t} - e^{-\frac{A}{2}t}\right)\right] \tag{5.28}$$
   where $A = \sqrt{\frac{L-4R^2C}{R^2C^2L}}$.
3. $L - 4R^2C < 0$
   In this case, the impulse response is obtained as
   $$h(t) = -\frac{1}{B}e^{-\frac{1}{2RC}t}\left[\frac{1}{RC}\sin\left(\frac{B}{2}t\right) - B\cos\left(\frac{B}{2}t\right)\right] \tag{5.29}$$
   where $B = -jA$.

We can now see that the voltage across the terminals of a parallel resonator shown in Figure 5.3, or the electric field contained in a signal confined within the resonator, decays from its initial value according to the damping factor $\delta = 1/2RC$, where $2RC$ represents the time constant, as

$$v(t) = V_o e^{-\delta t} = V_o e^{-\frac{1}{2RC}t} = V_o e^{-\frac{\pi f_r}{Q}t} \tag{5.30}$$

making use of $Q = 2\pi f_r RC$, where V_o is the initial value of the voltage. The voltages at two successive times t_1 and t_2 are written from (5.30) as

$$v(t_1) = V_o e^{-\frac{\pi f_r}{Q}t_1} \tag{5.31}$$

and

$$v(t_2) = V_o e^{-\frac{\pi f_r}{Q}t_2} \tag{5.32}$$

Taking the ratio of these voltages gives

$$\frac{v(t_1)}{v(t_2)} = e^{-\frac{\pi f_r}{Q}(t_1 - t_2)} \tag{5.33}$$

from which, we obtain

$$Q = \frac{\pi f_r (t_2 - t_1)}{\ln \dfrac{v(t_1)}{v(t_2)}} \tag{5.34}$$

Let Δt be the time duration during which the voltage amplitude or, in general, the time-domain response of the resonator's voltage, reduces to $1/e$ or 36.8% of its initial value, then the quality factor can be determined from (5.34) as

$$Q = \pi f_r \Delta t \tag{5.35}$$

In the above expression, $f_r \Delta t$ is equal to the number of RF periods or cycles during the duration Δt. Therefore, it can be said that the voltage of the resonator reduces to 36.8% of the initial value in Q/π cycles. The time interval Δt, as used here, is, in fact, equivalent to the skin depth of material, which is defined as the distance from the material's surface, over which the electric or magnetic field decays to 36.8% of its initial value at the surface. As can be expected, the energy stored within resonators decays to negligible value after several Δt's, at which time the resonators render their effectiveness. It is particularly noted that when the quality factor is significantly low, the decay happens so quickly that the decaying time is too short to be measured accurately, leading to inaccurate Q determination. The time-domain response is thus only useful when Q is sufficiently large that results in an accurately measurable decaying time interval. In circuit environments, the resonator is actually connected with external elements and so, in these environments, Q needs to be replaced with the loaded quality factor Q_L described later.

Similar to the electric field and voltage, the energy stored in the resonator decays over time, due to the resonator's loss, as

$$W = W_o e^{-2\delta t} \tag{5.36}$$

where W_o is the initial stored energy. The power loss can be derived as

$$P_L = -\frac{dW}{dt} = 2\delta W \tag{5.37}$$

from which,

$$\delta = \frac{P_L}{2W} = \frac{\omega_r}{2Q} \tag{5.38}$$

upon using (5.11). The unloaded quality factor can thus be used to describe the decaying rate of the stored energy, using (5.36) and (5.38), as

$$W = W_o e^{-\frac{\omega_r}{Q} t} \tag{5.39}$$

5.2.2 Series Resonators

For series resonators, as shown in Figure 5.4, the average magnetic energy stored in inductors can be obtained as

$$W_m = \frac{1}{4} L I_{\text{rms}}^2 \tag{5.40}$$

and the power loss is

$$P_L = \frac{1}{2} R I_{\text{rms}}^2 \tag{5.41}$$

where I_{rms} represents the (rms) current flowing through the resistor. Q of series resonators can be derived from (5.11), (5.40), and (5.41) as

$$Q = \frac{\omega_r L}{R} = \frac{1}{\omega_r RC} = \frac{1}{R}\sqrt{\frac{L}{C}} \tag{5.42}$$

The quality factor thus increases as the resistance reduces as expected. It can also be increased by increasing the ratio between L and C.

The input impedance Z_{in} given in Eq. (5.10) can be rewritten, using (5.42), as

$$Z_{\text{in}} \simeq R + j2RQ_o\frac{\Delta\omega}{\omega_r} \tag{5.43}$$

which shows that the minimum magnitude of Z_{in} is equal to R at resonant frequency. Figure 5.6 shows a sketch of the input impedance's magnitude versus $\Delta\omega/\omega_r$. Equating the magnitude of Z_{in} to $R/0.707$ leads to the same equation for Q of parallel resonators:

$$Q = \frac{\omega_r}{2\Delta\omega} = \frac{f_r}{2\Delta f} \tag{5.44}$$

where f_r and $2\Delta f$ are the resonant frequency and 3-dB bandwidth of the series resonator, respectively.

Similar to parallel resonators, the scattering parameters for a circuit consisting of a series resonator, represented by its impedance Z_{in}, connecting in series between two transmission lines of characteristic impedance Z_o can be derived as

$$S_{11} = S_{22} = \frac{Z_{\text{in}}}{Z_{\text{in}} + 2Z_o} \tag{5.45}$$

and

$$S_{21} = S_{12} = \frac{2Z_o}{Z_{\text{in}} + 2Z_o} \tag{5.46}$$

from which, the resonator's impedance is obtained as

$$Z_{\text{in}} = 2Z_o\frac{S_{11}}{1 + S_{11}} = 2Z_o\frac{1 - S_{21}}{S_{21}} \tag{5.47}$$

Equation (5.47) indicates that both the magnitude and phase of S_{11} or S_{21} for the circuit consisting of two transmission lines and series resonator are needed to determine accurately the unloaded Q of the resonator using (5.44). Similar to parallel resonators, Q of series resonators can also be estimated from Eq. (5.35) by calculating or measuring the resonant frequency and the time-domain response of the resonators, and the rate of energy decay can be determined from Eq. (5.39).

Practical resonators also contain parasitic capacitances and inductances, which can be substantial at RF frequencies. These parasitics, along with the main resistance, capacitance, and inductance of a resonator, can change the nature of resonance and make the resonator acts as a series resonant circuit at a certain frequency and a parallel resonant circuit at another frequency. These resonant frequencies and the corresponding quality factors can be determined by deriving the equivalent R, L, and C from the models of practical resonators.

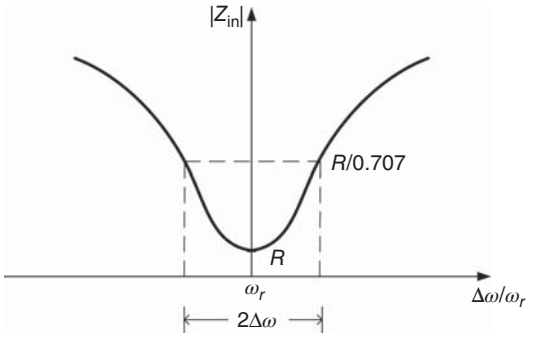


Figure 5.6. Sketch of Z_{in} versus $\Delta\omega/\omega_r$.

5.2.3 Unloaded Quality Factor

The foregoing quality factor is the "intrinsic" Q of resonators, which is often referred to as "unloaded" Q. Consider a typical resonator in CMOS processes which is made up of on-chip lumped-element inductor and capacitor or of (distributed) transmission line. The loss associated with lumped-element inductor, capacitor, and transmission line, in general, consists of losses due to metallization (conductor loss), dielectrics and Si substrate (dielectric loss), and radiation (radiation loss). Accordingly, we can divide the total unload Q into three different quality-factor components, which are defined as the intrinsic quality factors corresponding to conductor loss (Q_c), dielectric loss (Q_d), and radiation loss (Q_r), as

$$Q_c = \frac{\omega_r W}{P_c} \tag{5.48}$$

$$Q_d = \frac{\omega_r W}{P_d} \tag{5.49}$$

and

$$Q_r = \frac{\omega_r W}{P_r} \tag{5.50}$$

where P_c, P_d, and P_r are the power losses due to conductor, dielectric, and radiation. It is noted these equations for Q_c, Q_d, and Q_r are defined assuming the dielectric is perfect and no radiation, conductor is perfect and no radiation, and conductor and dielectric are perfect, respectively.

Assume the resonator is surrounded by a single dielectric characterized by its complex relative dielectric constant $\widehat{\varepsilon}_r = \varepsilon_r' + j\varepsilon_r''$, Q_d can be derived, assuming perfect conductor and no radiation, as

$$Q_d = \frac{\omega_r W}{P_d} = \frac{\omega_r \varepsilon_0 \varepsilon_r' \int |E|^2 dV}{\omega_r \varepsilon_0 \varepsilon_r'' \int |E|^2 dV}$$

$$= \frac{\varepsilon_r'}{\varepsilon_r'} = \frac{1}{\tan\delta} \tag{5.51}$$

where $\tan\delta$ represents the loss tangent of dielectric. Substituting $P_L = P_c + P_d + P_r$ into Eq. (5.11) and using (5.48)–(5.50) give

$$\frac{1}{Q} = \frac{1}{Q_c} + \frac{1}{Q_d} + \frac{1}{Q_r} \tag{5.52}$$

For properly designed resonators, Q_r is typically much greater than Q_c and Q_d and may be neglected. For on-chip resonators influenced primarily by SiO_2 dielectrics, Q_d is also much larger than Q_c, leaving Q_c as the dominant quality factor. On the other hand, for on-chip resonators significantly exposed to Si substrate, Q_d may be substantially lower than Q_c and hence dominates the overall quality factor.

5.2.4 Loaded Quality Factor

In practice, resonators are always connected with other (external) circuits or elements – for example, an LC resonator (or tank) is connected to a metal oxide semiconductor field effect transistor (MOSFET) in a voltage-control oscillator (VCO), a transmission-line resonator used in an RF band-pass filter, or a resonator connecting to a network analyzer for measuring the S-parameters. The connection is generally considered a coupling mechanism or network, which may be in any form, such as physically through a transmission line or electrically via a gap as shown in Figure 5.7. Within such a "circuit" environment, the quality factor of a resonator is affected not only by the intrinsic quality of the resonator, but also the quality of the external circuit

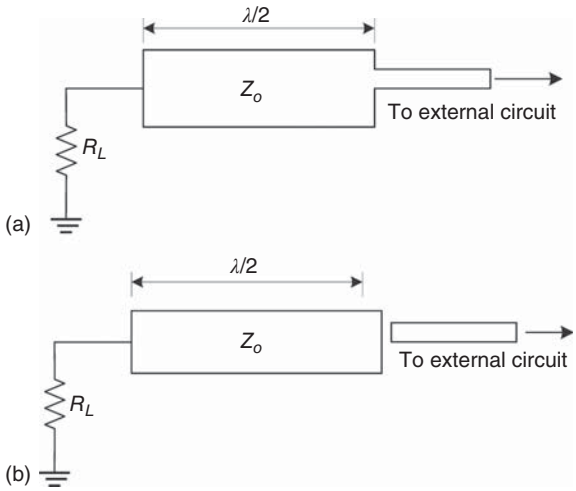


Figure 5.7. A half-wavelength resonator connecting to external circuits via a transmission-line (a) or gap (b) coupling network. The load resistance R_L can be any value, including zero and infinity, except Z_o.

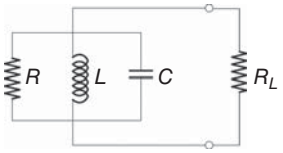


Figure 5.8. A parallel resonator loaded with a resistor of resistance R_L.

that is connected to the resonator. Consequently, there exists an effective (circuit) quality factor, commonly known as "loaded" quality factor for resonators.

Consider a parallel RLC resonator connected to a load represented by a resistance[3] R_L, as shown in Figure 5.8. To examine the effect of the load resistance on the resonator, we first let R to be infinitely large and determine the quality factor of the new resonator consisting of R_L, L, and C as

$$Q_e = \frac{\omega_r W}{P_e} = \frac{R_L}{\omega_r L} = \omega_r RC \tag{5.53}$$

where P_e represents the power loss due to the external circuit or load. Q_e is referred to as "external" Q. It is noted that the resonant frequency is still ω_r as of the original resonator. Now considering both R and R_L, the quality factor of the loaded resonator shown in Figure 5.8 can be derived as

$$Q_L = \frac{\omega_r W}{P_t} = \frac{RR_L(R+R_L)}{\omega_r L} = \frac{\omega_r CRR_L}{R+R_L} \tag{5.54}$$

where $P_t = P_L + P_e$ is the total power loss in the circuit. This quality factor is known as "loaded" Q and, from (5.15), (5.53), and (5.54), relates to Q and Q_e as

$$\frac{1}{Q_L} = \frac{1}{Q} + \frac{1}{Q_e} \tag{5.55}$$

which can be rewritten, using (5.53), as

$$\frac{1}{Q_L} = \frac{1}{Q} + \frac{P_e/P_L}{\omega_r W/P_L} = \frac{1}{Q} + \frac{K}{Q} \tag{5.56}$$

[3]In general, the load impedance representing an external circuit is complex. However, only real impedance is used for the external circuit in order to maintain the same resonant frequency as that of the intrinsic resonator. This allows the results to be examined simply without loss of generality.

where

$$K \equiv \frac{P_e}{P_L} \tag{5.57}$$

stands for the (coupling) coefficient of the coupling network between the resonator and external circuit. The unloaded quality factor can thus be obtained from the loaded quality factor, from (5.56), as

$$Q = (1 + K)Q_L \tag{5.58}$$

and is related to the external Q, using (5.11), (5.53), and (5.57), as

$$Q = KQ_e \tag{5.59}$$

It is noted that Q_L can be determined from Eq. (5.19) or (5.44) using the resonant frequency and 3-dB bandwidth of the magnitude of the input impedance of the loaded resonator (i.e., impedance looking into the composite one-port network of R, L, C, and R_L in parallel); that is, in the same way used for the unloaded Q of the RLC resonator. In the special case of no coupling loss ($K = 0$), Q_L is identical to Q as seen from (5.58).

As an example of evaluating loaded Q, we consider a parallel RLC resonator connected with a transmission line and a source, as shown in Figure 5.9(a). This arrangement is similar to that encountered in measuring the resonator's S-parameters using a network analyzer. Figure 5.9(b) shows the equivalent circuit, which is divided into internal (resonator) and external sections. The coupling coefficient can be derived, using (7.15), (5.54), and (5.56), as

$$K = \frac{Q}{Q_L} - 1 = \frac{R}{Z_o} \tag{5.60}$$

Table 5.1 summarizes the results for different values of K.

It should be noted that for the critical coupling case, the total power loss (P_t) for the loaded resonator is equal to twice of either the power loss in the resonator itself (P_L) or the external network (P_e).

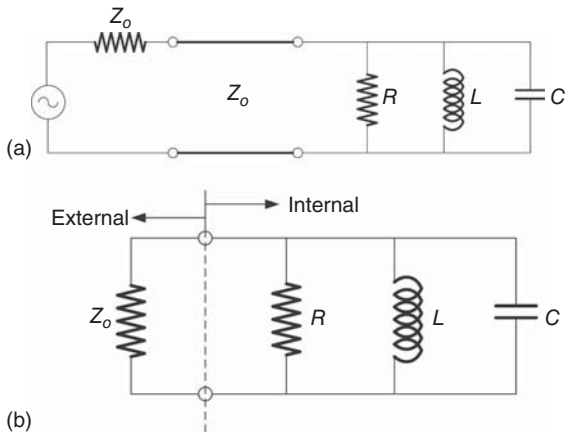


Figure 5.9. (a, b) A parallel resonator embedded in a circuit environment.

TABLE 5.1. Quality Factor for Different Values of the Coupling Coefficient

K	R	Resonator coupling	Quality factor
$K = 1$	$R = Z_o$	Critically coupled	$Q = 2Q_L$
$K < 1$	$R < Z_o$	Under-coupled	$Q < 2Q_L$
$K > 1$	$R > Z_o$	Over-coupled	$Q > 2Q_L$

5.2.5 Evaluation of and Relation between Unloaded and Loaded Quality Factors

In measurement of resonators, the coupling network between a resonator and external elements needs to be characterized accurately in order to determine the accurate unloaded quality factor for the resonator. When the coupling is very light ($K \to 0$), the measured loaded Q approaches the unloaded Q and thus can be used for unloaded Q without the need of characterizing the coupling network. For instance, the loaded Q of a half-wavelength microstrip resonator measured with a gap producing around −20-dB coupling at each end is close to the unloaded Q.

As can be recognized, the coupling from resonators to external circuits or elements in RFICs is important, as it affects the loaded Q and hence the overall circuit performance. Trade-off between different circuit specifications with respect to the resonator's loaded Q, therefore, should be done in RFIC design. It is noted that unloaded Q is the highest Q obtained from a resonator and circuits should be designed and laid out so that the resonator can achieve a loaded Q approaching its unloaded counterpart. For instance, in oscillators, generally the stronger the coupling to the resonator, the higher the generated power and phase noise. So for applications where phase noise of oscillators is critical, the resonator should be lightly coupled (under-coupled) to external circuit elements. Figure 5.10 illustrates an oscillator configured from a feedback amplifier employing a (distributed) transmission-line resonator between the gate and drain. The coupling networks are used to control the couplings between the resonator to external elements, which affect the output power and phase noise of the oscillator. In order to achieve low phase noise, the coupling networks need to be configured to produce small couplings. Assuming the coupling networks are formed by gaps between the transmission-line resonator and other transmission lines, small couplings can be obtained by using large or off-set gaps. Figure 5.11 illustrates an off-set gap-coupling configuration between a transmission-line resonator and a transmission line. The off-set coupling can be further reduced by using multilayer available in RFIC structures such that the two ends of a gap lie on two different metal layers separated by one or more dielectric layers.

It is particularly noted the unloaded quality factor of a resonator is for the resonator itself – not terminated with any impedance. However, the quality factor of a resonator determined from the input impedance or S-parameters of a two-port network within which the resonator is embedded, such as that seen in measurement using network analyzers, where both the input and output ports of a resonator are connected to the network analyzer, is the loaded quality factor.

We consider a general resonator, configured as a two-port element, connected to a source and load via coupling networks and transmission lines as shown in Figure 5.12(a). This connection specifically exemplifies the two-port measurement of resonators using a network analyzer. The coupling networks at the input and output ports serve as coupling mechanism needed for measuring the resonator's characteristics. An example of these coupling networks is gaps in transmission lines. In general, a coupling network transfers the impedance or admittance of the resonator at one side into another impedance or admittance at the other side or transfers the generator or load at one side into another generator or load at the other side. A coupling network can be represented, in general, by resistances, accounting for losses due to dielectrics, substrates and conductors,

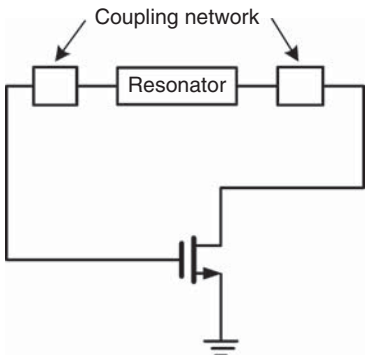


Figure 5.10. A feedback oscillator employing a distributed resonator.

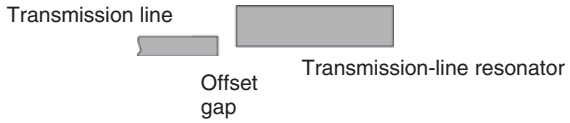


Figure 5.11. Offset coupling-gap between a transmission-line resonator and transmission line.

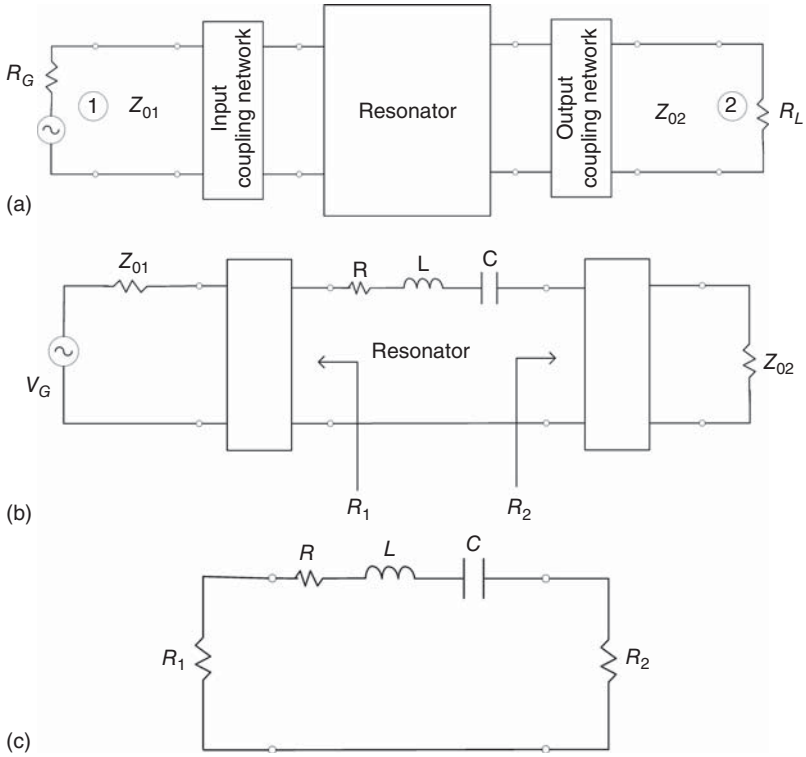


Figure 5.12. A resonator connected to generator and load in a two-port environment (a) and its equivalent circuits (b) and (c).

and inductances and capacitances representing the field distributions within the coupling network. Without loss of generality, we assume the resonator is a series resonator represented by R, L, and C, the coupling networks act as ideal impedance or admittance inverters, and the source and load impedances are equal to the characteristic impedances of the corresponding connecting transmission lines, leading to the equivalent circuit in Figure 5.12(b). Transferring the source and load across the respective impedance inverter gives the equivalent circuit shown in Figure 5.12(c).

As will be seen in Section 5.5, an impedance or admittance transforms an impedance or admittance connected at one end of the inverter into another impedance or admittance at the other inverter's end, respectively. The two typical coupling networks used for printed-circuit resonators are gap and connecting junction between a transmission line and the resonator. A gap, as can be seen in Section 5.3.2.3, can be accurately represented by a π-network consisting of two shunt capacitances and one series capacitance. This network, as explained in Section 5.4.3, behaves as an admittance inverter. A junction between a transmission line and printed-circuit resonator represents a discontinuity in the width of the transmission line and can be modeled as a T-network shown in Figure 5.13, where the series inductances and shunt capacitance can be accurately determined using full-wave electromagnetic (EM) methods. This network approximately represents an impedance inverter.

The unloaded Q of the RLC resonator in Figure 5.12 is obtained from (5.42) as

$$Q_o = \frac{\omega_r L}{R} \tag{5.61}$$

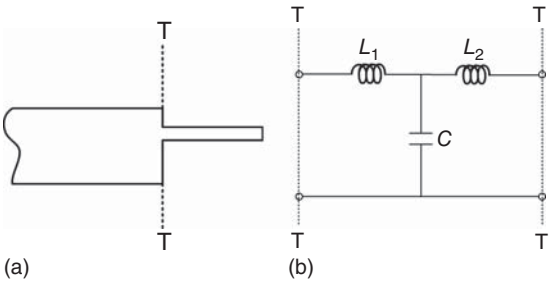


Figure 5.13. A step discontinuity (a) and its equivalent circuit (b) behaving as an impedance inverter.

The loaded Q can also be determined from (5.42) and Figure 5.12(c) as

$$Q_L = \frac{\omega_r L}{R + n_1^2 Z_{o1} + n_2^2 Z_{o2}} \tag{5.62}$$

where n_1 and n_2 are the ratios of the ideal transformers representing the input and output coupling networks, respectively. The external Q's due to loading at the input and output ports are derived from (5.53) as

$$Q_{ei} = \frac{\omega_r L}{n_i^2 Z_{oi}} \tag{5.63}$$

where $i = 1$ and 2 corresponding to the input and output ports, respectively. The coupling coefficients K_1 and K_2 corresponding to the input and output coupling networks, respectively, can be determined from (5.57), using (5.61) and (5.63), as

$$K_i = \frac{\dfrac{\omega_r W}{Q_{ei}}}{\dfrac{\omega_r W}{Q}} = \frac{R_i}{R} \tag{5.64}$$

where R_i $(i = 1, 2)$ is the transformed impedance in Figure 5.12(b) and (c). These impedances can be determined from the terminated coupling networks and relate to the coupling coefficients as

$$R_i = K_i R \tag{5.65}$$

The unloaded Q is related to the loaded Q, making use of (5.61), (5.62), and (5.64), as

$$Q = Q_L(1 + K_1 + K_2) \tag{5.66}$$

The insertion loss from the input port 1 to the output port 2 can be derived, under matched conditions at the source and load, as

$$\text{Insertion loss} \triangleq |S_{21}(\omega)|^2 = \frac{4K_1K_2}{(1 + K_1 + K_2)^2 + \left(2Q\dfrac{\Delta\omega}{\omega_r}\right)^2} \tag{5.67}$$

At resonant frequency, the insertion loss reaches a minimum value given by

$$|S_{21}(\omega_r)|^2 = \frac{4K_1K_2}{(1 + K_1 + K_2)^2} \tag{5.68}$$

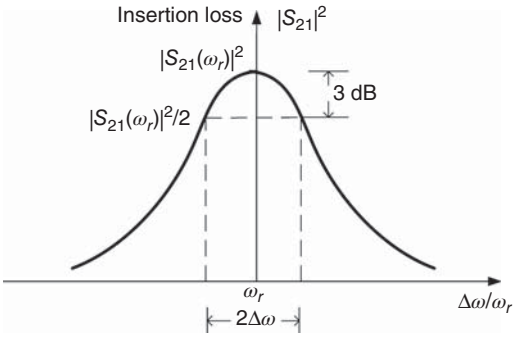


Figure 5.14. Sketch of insertion loss versus $\Delta\omega/\omega_r$.

which, upon substituted into (5.67) and utilizing (5.66), results in

$$|S_{21}(\omega)|^2 = \frac{|S_{21}(\omega_r)|^2}{1 + \left(2Q_L \dfrac{\Delta\omega}{\omega_r}\right)^2} \tag{5.69}$$

Figure 5.14 shows a sketch of the insertion loss versus $\Delta\omega/\omega_r$.

Equating the insertion loss in (5.69) to $|S_{21}(\omega_r)|^2/2$, which corresponds to half-power or 3-dB points on the insertion loss curve, leads to

$$Q_L = \frac{\omega_r}{2\Delta\omega} = \frac{f_r}{2\Delta f} \tag{5.70}$$

where f_r and $2\Delta f$ are the resonant frequency and 3-dB bandwidth of the insertion loss, respectively. Replacing the series resonator in Figure 5.12 with a parallel resonator and following the same approach, we can also derive the same equations (5.66) and (5.70). As the coupling is increased, which corresponds to increasing K_i, the insertion loss approaches toward the minimum value at resonance and the loaded Q reduces and, in the limit of K_i approaching infinity, the resonance vanishes. This is expected as the composite network, consisting of the resonator and loading circuits, acts as an all-pass network with insertion loss being a constant versus frequency. On the other hand, as the coupling is reduced to zero, the insertion loss increases faster as frequency moves farther from the resonant frequency, and the loaded Q approaches the unloaded Q. The loading circuits essentially modify the original resonator to create a new resonator consisting of the original resonators and two coupling networks. The original resonator's loaded Q is indeed the unloaded Q of the new resonator. Due to the loading network with its own resistance, inductance, and capacitance, it is expected that the resonant frequency of the new resonator would be different from that of the original resonator. This change, however, is not significant if the loading network was properly designed as typically done in RFIC design. To minimize possible effects from the loading circuits in evaluating the unloaded Q of a resonator, it is best to use coupling networks producing as small coupling coefficient as possible to the limit that the combined resonator and coupling networks still produces an insertion loss that can be obtained with sufficient accuracy within a 3-dB bandwidth. A sufficiently large gap between the feeding transmission line and the input of a transmission-line resonator, for instance, can help achieve this.

The accuracy of determining the unloaded Q of resonators, as can be seen from (5.66), depends significantly on accurate determination of the coupling coefficients. In practice, same coupling network is typically used at the input and output of resonators. In this case, the insertion loss at resonance given in (5.68) becomes

$$|S_{21}(\omega_r)|^2 \triangleq L_r = \frac{4K^2}{(1+2K)^2} \tag{5.71}$$

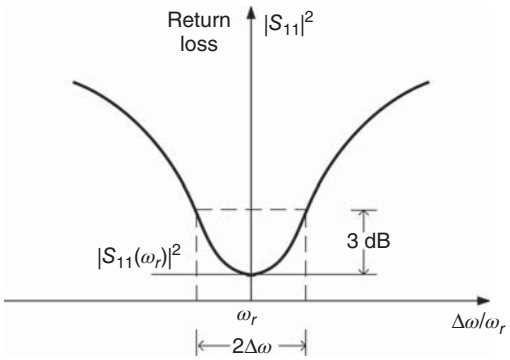


Figure 5.15. Sketch of return loss versus $\Delta\omega/\omega_r$.

from which, we can determine the coupling coefficient K from the insertion loss at the resonant frequency as

$$K = \frac{L_r}{2(1-L_r)} \pm \sqrt{\left[\frac{L_r}{2\,(1-L_r)}\right]^2 + \frac{1}{4(1-L_r)}} \tag{5.72}$$

This equation produces two values for K, which are smaller and greater than 1. The choice of a particular value depends on the expected magnitude of the coupling.

Using return loss versus frequency, we can also prove that the relation between the unloaded and loaded Q's of a parallel or series resonator is given in (5.66) and the loaded Q can be determined from (5.70) with f_r and $2\Delta f$ now representing the resonant frequency and 3-dB bandwidth, respectively, seen in the plot of the return loss at port 1 or 2 versus frequency as shown in Figure 5.15.

The loaded Q of a resonator can thus be determined as the ratio between the resonant frequency and 3-dB bandwidth obtained from the insertion loss of the composite circuit, as shown in Figure 5.12(a), in which the resonator is embedded, from which the unloaded Q of the resonator can be determined with the knowledge of the coupling coefficients. Accurate evaluation of the coupling coefficients, either through calculations or through measurements, is critical for accurate determination of the unloaded Q from the loaded Q. This may pose difficulty for some particular coupling networks. It is noted that the coupling coefficients, as given in (5.64), assume ideal transformers and known resistance for a resonator. This poses additional problems of determining accurately the transformers associated with the coupling networks and the resonator's resistance. It is also noted that different models for a coupling network give different formulas for the coupling coefficient. For instance, for a series RLC resonator connecting directly with a transmission line of characteristic impedance Z_o, the coupling coefficient is Z_o/R. There are thus several uncertainties in determining the unloaded Q using the insertion loss data and proper care needs to be exercised for accurate determination. As pointed out earlier and seen in (5.66), the unloaded Q approaches the loaded Q when the couplings are very small. Therefore, two-port evaluation of unloaded Q should be done using coupling networks with sufficiently small coupling coefficients. This can be implemented relatively easy in circuits and thus ensuring accurate measurement of unloaded Q.

When the two coupling networks are different, evaluation of coupling coefficients can be carried individually using two different circuits. For instance, the input coupling coefficient can be determined by connecting the input of the resonator to the input coupling network and terminating the output with known impedance R_L, as shown in Figure 5.16(a). The coupling coefficient of the coupling network can be derived, from Figure 5.16(b), as

$$K = \frac{R_1}{R + R_L} \tag{5.73}$$

Without loss of generality, we assume the coupling network function as an impedance inverter (see Section 5.5) which converts the series "Resonator 2" connected at one end into a parallel resonator at the other end.

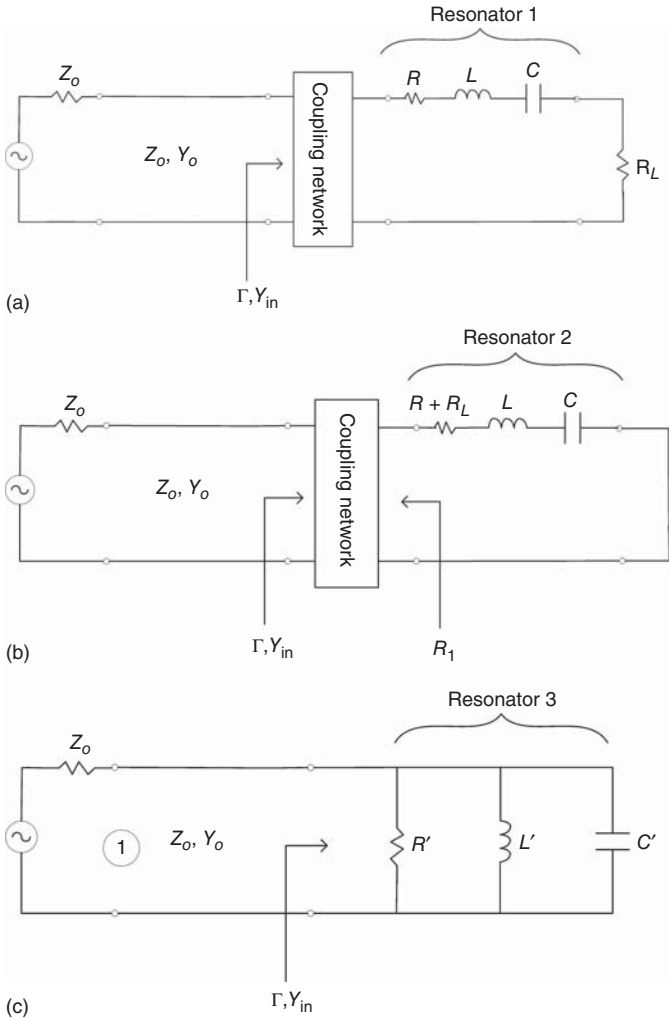


Figure 5.16. A terminated series resonator connecting to a transmission line via a coupling network (a) and its equivalent circuits (b) and (c) used for determining the coupling coefficient.

This results in a new resonator, namely Resonator 3, as shown in Figure 5.16(c). Elements R', L', and C' of Resonator 3 can be extracted from Y_{in} in Figure 5.16(b).

The reflection coefficient at Resonator 3's input can be derived as

$$\Gamma = \frac{Y_o - Y_{\text{in}}}{Y_o - Y_{\text{in}}} \simeq \frac{\omega_r(K' - 1) - j2Q_L\Delta\omega}{\omega_r(K' + 1) + j2Q_L\Delta\omega} \tag{5.74}$$

upon using (5.6) and (5.15), where K' is the coupling coefficient characterizing the coupling between Resonator 3 and the transmission line given, from (5.60), as

$$K' = \frac{R'}{Z_o} \tag{5.75}$$

It is noted that the unloaded quality factor of Resonator 3 is the same as the loaded quality factor of Resonator 1 or loaded quality factor of Resonator 2. Assume the source impedance is equal to the transmission line's characteristic impedance as shown in Figure 5.16, the power absorbed by Resonator 3, which is the combined network of Resonator 1, load resistance R_L and coupling network, is equal to the difference between the incident power at port 1 and the reflected power at Resonator 3's input. For typical low loss coupling

networks, this power represents mainly the power dissipated on R and R_L. The absorbed power normalized to the incident power is considered to be the loss caused by Resonator 3 and can be derived as

$$L_a = (1 - |\Gamma|^2) \simeq \frac{L_{ar}}{1 + \left(\frac{2Q_L \Delta\omega/\omega_r}{K'+1}\right)^2} \tag{5.76}$$

where Q_L is the unloaded quality factor of Resonator 1 and

$$L_{ar} = \frac{4K'}{(K'+1)^2} \tag{5.77}$$

is the loss at the resonant frequency. Solving (5.76) and (5.77) for K' gives

$$K' = \frac{1 + |\Gamma_r|^2}{1 - |\Gamma_r|^2} \pm \sqrt{\left(\frac{1 + |\Gamma_r|^2}{1 - |\Gamma_r|^2}\right)^2 - 1} \tag{5.78}$$

where Γ_r represents the reflection coefficient at the resonance. K' has two values, greater and smaller than 1, depending on the coupling network employed. Once K' is determined, the coupling coefficient K of the coupling network can be calculated using (5.73), (5.75), and the parameters of the coupling network.

The loaded Q_L of Resonator 1 is related to its unloaded Q as

$$Q = (1 + K + K_L)Q_L \tag{5.79}$$

where K_L is the coefficient representing the coupling between Resonator 1 and the load R_L given as

$$K_L = \frac{R_L}{R} \tag{5.80}$$

When $R_L = 0\ \Omega$, Resonator 1 essentially behaves as a one-port resonator and corresponding equations can be obtained from the foregoing equations.

When a resonator is connected with only one external circuit, such as a parallel RLC resonator connected directly with a transmission line of characteristic impedance Z_o similar to that shown in Figure 5.16(c), it is essentially embedded in a one-port environment. In this case, there is only one coupling coefficient involved, and the relation between the loaded and unloaded Q's is

$$Q = Q_L(1 + K) \tag{5.81}$$

where K is given as

$$K = \frac{R}{Z_o} \tag{5.82}$$

and Q_L can be determined from (5.70) using the resonant frequency and 3-dB bandwidth obtained from the return loss versus frequency of the one-port network.

As an example, Figure 5.17 shows the layout and calculated input impedance of a series resonator consisting of a 0.629-nH spiral inductor and a 49-fF metal–insulator–metal (MIM) capacitor in a 0.18-μm CMOS process. The resonant frequency and unloaded quality factor can be determined from Figure 5.17(b) as 28.52 GHz and 12.4, respectively. Figure 5.18 shows the calculated input impedance of a parallel resonator consisting of a 0.213-nH spiral inductor and a 146-fF MIM capacitor in the same process, showing a resonant frequency and unloaded quality factor of 28.42 GHz and 22.2, respectively.

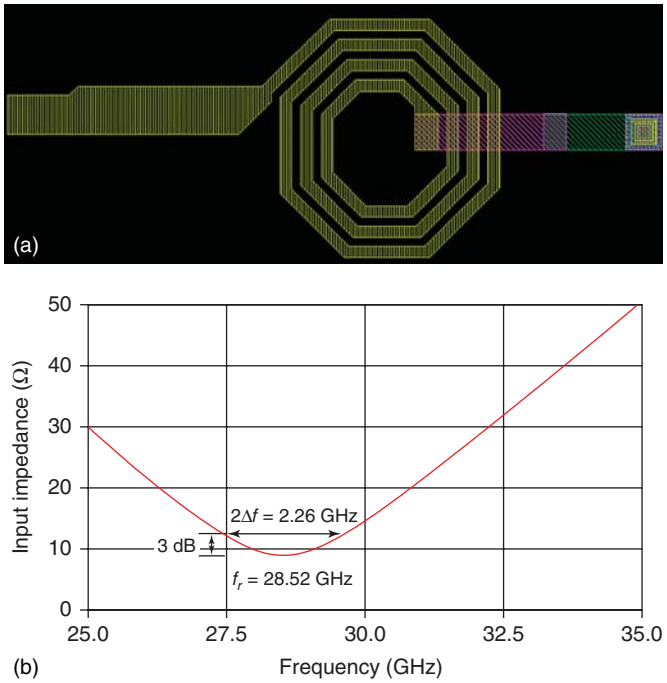


Figure 5.17. Layout (a) and calculated input impedance (b) of a series resonator in a 0.18-μm CMOS process.

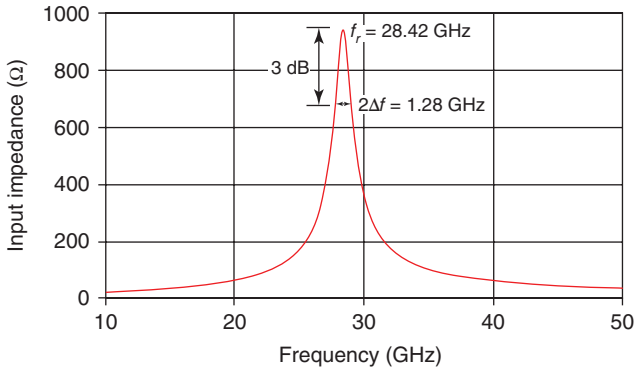


Figure 5.18. Calculated input impedance of a parallel resonator in a 0.18-μm CMOS process.

5.3 DISTRIBUTED RESONATORS

As frequency is increased, good on-chip lumped elements such as inductors and capacitors become more difficult to be realized – both in quality and inductance and capacitance values. Specifically, the loss is increased causing low Q, and the inductance and capacitance become limited. For instance, consider two-parallel-plate MIM capacitors typically employed in RFICs, large capacitance is achieved by increasing the relative dielectric constant of the dielectric between the metal plates and/or increasing the metal plates' area. SiO_2 is typically used as the dielectric and so the only design option is using large plate area. This, however, causes problems such as increased fringing capacitances along the edges, increased coupling to nearby elements, making the capacitor resembling a distributed structure (parallel-plate transmission line) or becoming a resonant or radiation structure. These difficulties are particularly more pronounced in the high millimeter-wave frequency range. As such, it may be difficult to design high quality, high resonant-frequency lumped-element resonators in the high millimeter-wave end, and distributed resonators become necessary for RFICs at these frequencies.

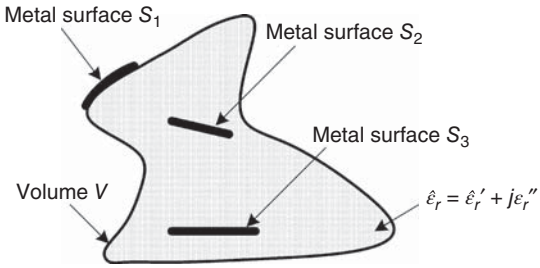


Figure 5.19. General configuration for distributed resonators.

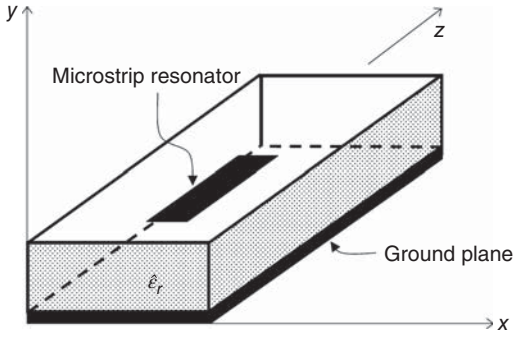


Figure 5.20. A distributed microstrip resonator.

5.3.1 Quality-Factor Characteristics

We consider a general distributed resonator made up of metals and dielectric having a complex relative dielectric constant $\varepsilon_r = \varepsilon_r' + j\varepsilon_r''$ as shown in Figure 5.19. Figure 5.20 shows a microstrip line with two conductors – one acting as the signal line and the other as the ground plane – and a dielectric substrate, as an example of distributed resonators. The total time-average energy stored within the structure can be obtained as

$$W = 2W_e = 2W_m = \frac{1}{2}\varepsilon_o\varepsilon_r \iiint_V |\vec{E}|^2 \, dV \tag{5.83}$$

where V stands for the volume enclosing the resonator and E represents the electric field.

Assume the metals are good conductors having low loss, the power loss due to the metals can be approximately determined by employing a perturbation procedure, which assumes the current flow on the metal is the same as that flowing on a (lossless) perfect metal. The power loss due to the metals is then approximately given as

$$P_c \simeq \sum_{n=1}^{N} \frac{R_{sn}}{2} \iint_{S_n} |\vec{J}_{sn}|^2 dS = \sum_{n=1}^{N} \frac{R_{sn}}{2} \iint_{S_n} |\vec{H}_{t,n}|^2 dS \tag{5.84}$$

where N is the number of metals, R_{sn} is the surface resistivity of metal n, $\vec{J}_{sn} = \hat{n} \times \vec{H}_n = \vec{H}_{n,t}$, with H_n and $H_{n,t}$ being the magnetic field and tangential magnetic field along the perfect metal surface, is the surface current density flow on the perfect metal n. The power loss due to dielectric can be derived as

$$P_d = \frac{1}{2}\omega_r\varepsilon_o\varepsilon_r'' \iiint_V |\vec{E}|^2 \, dV \tag{5.85}$$

The quality factor due to dielectric loss can hence be obtained from (5.49), (5.83), and (5.85) as

$$Q_d = \frac{\omega_r W}{P_d} = \frac{\varepsilon_r'}{\varepsilon_r'} = \frac{1}{\tan\delta} \tag{5.86}$$

For good resonators, the radiation loss is negligibly small and so, for given dielectrics, the unloaded quality factor Q of resonators is thus proportional to Q_c, or

$$Q \propto \frac{\text{energy stored}}{\text{energy dissipated within metals}} \propto \frac{\text{volume of resonator}}{\text{surface area of metals}} \tag{5.87}$$

which shows that Q can be increased by increasing the volume of resonators and/or reducing the metal surface areas. Equation (5.87) underlines the principle for enhancing the quality factor. Similar equation can also be derived for lumped-element resonators. It is worth mentioning that the volume here refers to the volume of the "entire" resonators, not necessarily relating to the metal surfaces directly. As such, it may be possible to increase the volume while reducing the metal areas. Since the volume and surface areas, or consequently the dimensions of constituent elements, of resonators are related to resonant frequency, these requirements are more realizable at high frequencies, considering the size constraints for RFICs. It should be noted Eq. (5.87) also applies to individual lumped inductors and capacitors as well as distributed elements. For instance, the Q of spiral inductors can be increased by using large diameters (i.e., the outer conducting traces) while minimizing the number of traces (i.e., inner traces.)

5.3.2 Transmission-Line Resonators

In general, transmission-line resonators of various lengths may be formed with ends terminated by a proper impedance such as short, open, or resistance. Consider a transmission-line resonator with termination at one end, the requirement for resonance to occur is the input impedance or admittance is real at the resonant frequency and complex at other frequencies. Commonly used transmission-line resonators are short- and open-circuited quarter- and half-wavelength transmission lines. Another type is transmission-line rings, which are also used widely. Conventional implementations of these resonators produce relatively large resonators, resulting in large RFICs. Implementing these transmission-line resonators using multiple metal and dielectric layers available in CMOS structures, however, can help reduce the size and hence cost significantly.

It is particularly noted that, due to high loss in Si substrate and finite physical size of on-chip transmission lines (and even lumped-element inductors and capacitors), very high Q may not be possible with on-chip transmission-line (and lumped-element) resonators. This limited Q, together with a relatively large size, which is undesirable for RFICs, makes on-chip resonators less attractive than their off-chip counterparts for some applications – for example, typical oscillators in cell phones employ off-chip quarter-wavelength resonators fabricated on piezoelectric materials, such as Barium Titanate ($BaTiO_3$) having relative dielectric constant of around 1070, which results in very compact, high Q resonators.

Transmission-line resonators are typically designed using transmission lines configured so that their losses are as low as possible. For low loss transmission lines, it is recalled that the attenuation constant α, phase constant β, phase velocity v_p, and characteristic impedance Z_o are approximately given in Eqs. (4.70)–(4.73) as

$$\alpha \simeq \frac{1}{2}\sqrt{L'C'}\left(\frac{R'}{L'} + \frac{G'}{C'}\right) \tag{5.88}$$

$$\beta \simeq \omega\sqrt{L'C'} \tag{5.89}$$

$$v_p \simeq \frac{1}{\sqrt{L'C'}} \tag{5.90}$$

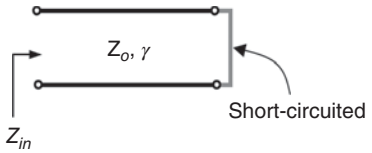


Figure 5.21. Short-circuited transmission line.

and

$$Z_o \simeq \sqrt{\frac{L'}{C'}} \tag{5.91}$$

where R', L', G', and C' are the transmission line's resistance, inductance, conductance, and capacitance per unit length, respectively. These equations will be used in subsequent analyses.

5.3.2.1 Short-Circuited Transmission-Line Resonators. Consider a transmission line characterized by characteristic impedance Z_o, physical length ℓ, and propagation constant $\gamma = \alpha + j\beta$, where α and β are the attenuation and phase constants, respectively. Assume one end of the transmission line is short-circuited as shown in Figure 5.21. The input impedance of the short-circuited transmission line is given by

$$Z_{\text{in}} = Z_o \tanh(\alpha + j\beta)\ell = Z_o \frac{\tanh(\alpha\ell) + j\tan(\beta\ell)}{1 + j\tanh(\alpha\ell)\tan(\beta\ell)} \tag{5.92}$$

As we are interested in the behavior of the transmission line at a (single) resonant frequency ω_r, we consider a very narrow frequency range $\Delta\omega$ around ω_r such that $\Delta\omega \ll \omega_r$. At frequency $\omega = \omega_r + \Delta\omega$ within the neighborhood of ω_r, we can write the electrical length θ of the transmission line as

$$\theta = \beta\ell = \frac{\omega\ell}{v_p} = \frac{\omega_r\ell}{v_p} + \frac{\Delta\omega\ell}{v_p} = 2\pi\frac{\ell}{\lambda} + 2\pi\frac{\ell}{\lambda}\frac{\Delta\omega}{\omega_r} \tag{5.93}$$

where v_p is the phase velocity and λ is the wavelength. We further assume that the transmission line has low loss, hence $\alpha\ell \ll 1$, and so

$$\tanh(\alpha\ell) \simeq \alpha\ell \tag{5.94}$$

Short-Circuited Half-Wavelength Resonators

For resonators whose length is half-wavelength at the resonant frequency, the electrical length can be obtained from (5.93) as

$$\theta = \pi + \pi\frac{\Delta\omega}{\omega_r} \tag{5.95}$$

from which, we can write, in the vicinity of ω_r,

$$\tan\theta = \tan\left(\pi + \pi\frac{\Delta\omega}{\omega_r}\right) = \tan\left(\pi\frac{\Delta\omega}{\omega_r}\right) \simeq \pi\frac{\Delta\omega}{\omega_r} \tag{5.96}$$

The input impedance of a half-wavelength short-circuited low loss transmission line can now be expressed from (5.92), using (5.94) and (5.96), as

$$Z_{\text{in}} \simeq Z_o \frac{\alpha\ell + j\pi\dfrac{\Delta\omega}{\omega_r}}{1 + j\pi\alpha\ell\dfrac{\Delta\omega}{\omega_r}} \simeq Z_o\left(\alpha\ell + j\pi\frac{\Delta\omega}{\omega_r}\right) \tag{5.97}$$

At resonant frequency,

$$\ell = \frac{\lambda}{2} = v_p \frac{\pi}{\omega_r} \tag{5.98}$$

from which we obtain, using (5.90),

$$\frac{\pi}{\omega_r} \simeq \ell \sqrt{L'C'} \tag{5.99}$$

Equation (5.97) can now be rewritten, using (5.88), (5.91), and (5.99), as

$$Z_{\text{in}} \simeq \frac{1}{2}\ell \left(R' + \frac{G'L'}{C'} \right) + j\ell L' \Delta\omega \tag{5.100}$$

Comparing (5.97) and (5.100) to (5.10) shows that low loss half-wavelength short-circuited resonators behave as series resonators, whose elements are

$$R \simeq Z_o \alpha \ell \simeq \frac{1}{2}\ell \left(R' + \frac{G'L'}{C'} \right) \tag{5.101}$$

$$L \simeq \frac{\pi Z_o}{2\omega_r} = \left(\frac{\pi}{2} \right)^2 \ell L' \tag{5.102}$$

and

$$C = \frac{1}{\omega_r^2 L} \simeq \left(\frac{2}{\pi} \right)^2 \ell C' \tag{5.103}$$

obtained with the help of (5.8) and (5.98) assuming $L = \ell L'$ and $C = \ell C'$. When the dielectric loss can be neglected, such as in the case of RFIC transmission-line resonators surrounded by low loss SiO_2 dielectrics and isolated from lossy Si substrate, a half-wavelength short-circuited resonators approximately behave as a series resonator with resistance and inductance equal to half of the total resistance and inductance, respectively, and capacitance equal to 0.2 of the total capacitance of the transmission line.

The quality factor of low loss half-wavelength short-circuited resonators can be determined from (5.42), (5.101), and (5.102) as

$$Q \simeq \frac{\pi}{2\alpha\ell} \simeq \omega_r \frac{L'C'}{R'C' + G'L'} \simeq \frac{\beta}{2\alpha} \tag{5.104}$$

The conductor (α_c) and dielectric (α_d) attenuation constants for different transmission lines are given in Chapter 4 for transmission lines. Equation (5.104), as will be seen later, is also applicable to other transmission-line resonators. It shows that the quality factor is increased as the phase constant β and attenuation constant α increases and decreases, respectively. Like the effect of α, which has been discussed previously, the effect of β is also expected since larger β translates to slower signal speed, resulting in longer time for energy to be stored in resonators and hence larger Q. Using structures and/or materials allowing signals to travel with slower speed, such as slow-wave structures, would indeed help improve the quality factor not only of resonators, but also of individual lumped and distributed elements.

The quality factor of a microstrip short-circuited half-wavelength resonator, as shown in Figure 5.22, due to radiation can be approximated from [2]

$$Q_r \simeq \frac{15 Z_o \varepsilon_{re}^2}{32\eta_a \left(\frac{h}{\lambda_a} \right)^2 \left(1 + \frac{2}{7\varepsilon_{re}} \right)} \tag{5.105}$$

where ε_{re} and λ_a are the effective relative dielectric constant and free-space wavelength at resonant frequency, Z_o is the characteristic impedance, and $\eta_a = 377\,\Omega$ is the intrinsic impedance of air.

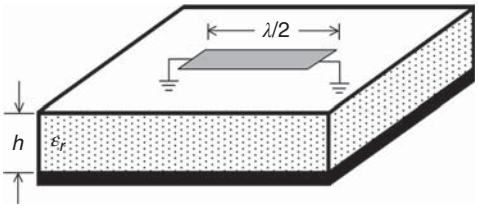


Figure 5.22. A microstrip short-circuited half-wavelength resonator.

It is noted that the input impedance of half-wavelength short-circuited transmission lines, as seen in (5.97) or (5.100), becomes resistive capacitance or resistive inductance, respectively, at frequencies smaller or greater than the resonant frequency, which correspond to a length greater or smaller than a half-wavelength. It is recalled that series resonators behave as a lossy capacitor (RC) or lossy inductor (RL) network at frequencies smaller or greater than the resonant frequency, respectively. It is thus deemed that half-wavelength short-circuited resonators function as series resonators, as proved earlier.

Short-Circuited Quarter-Wavelength Resonators

For resonators whose length is a quarter-wavelength at the resonant frequency, the electrical length can be obtained from (5.93) as

$$\theta = \frac{\pi}{2} + \frac{\pi}{2}\frac{\Delta\omega}{\omega_r} \tag{5.106}$$

Within a small frequency range $\Delta\omega$ around ω_r, we can write

$$\tan\theta = \tan\left(\frac{\pi}{2} + \frac{\pi}{2}\frac{\Delta\omega}{\omega_r}\right) = -\frac{1}{\tan\left(\frac{\pi}{2}\frac{\Delta\omega}{\omega_r}\right)} \simeq -\frac{2}{\pi}\frac{\omega_r}{\Delta\omega} \tag{5.107}$$

The input impedance of a low loss quarter-wavelength short-circuited transmission line can now be expressed from (5.92), using (5.94) and (5.107), as

$$Z_{\text{in}} \simeq \frac{Z_o}{\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_r}} \tag{5.108}$$

where $\ell = \lambda/4$ at ω_r, which can be rewritten as

$$Z_{\text{in}} \simeq \frac{1}{\frac{1}{2}\ell\left(G' + \frac{R'C'}{L'}\right) + j\frac{\ell C'}{2}\Delta\omega} \tag{5.109}$$

Comparing (5.108) and (5.109) to (5.17) reveals that low loss quarter-wavelength short-circuited resonators behave as parallel resonators, whose elements are given by

$$R \simeq \frac{Z_o}{\alpha\ell} \simeq \frac{1}{\frac{1}{2}\ell\left(G' + \frac{R'C'}{L'}\right)} \tag{5.110}$$

$$C \simeq \frac{\pi}{4\omega_r Z_o} \simeq \left(\frac{\pi}{4}\right)^2 \ell C' \tag{5.111}$$

and

$$L = \frac{1}{\omega_r^2 C} \simeq \left(\frac{4}{\pi}\right)^2 \ell L' \tag{5.112}$$

The quality factor of low loss quarter-wavelength short-circuited resonators can be determined from (5.15), (5.110), and (5.111) as

$$Q \simeq \frac{\pi}{4\alpha\ell} \simeq 8\omega_r \frac{L'C'}{R'C' + G'L'} \simeq \frac{\beta}{2\alpha} \tag{5.113}$$

which is the same as that for low loss half-wavelength short-circuited resonators. The resonance occurs at multiple frequencies of the (fundamental) resonant frequency ω_r corresponding to $\ell = (2n+1)\lambda/4$, where $n = 0, 1, 2, \ldots$; that is, at frequencies where the length is a multiple of a quarter-wavelength.

The reactance of the input impedance of quarter-wavelength short-circuited transmission lines, as given in (5.108), is inductive or capacitive, respectively, at frequencies below or above the resonance, which is the same as parallel resonators. Quarter-wavelength short-circuited resonators thus behave as parallel resonators, as we have seen previously.

5.3.2.2 Open-Circuited Transmission-Line Resonators. Consider an open-circuited transmission line characterized by a characteristic impedance Z_o and propagation constant $\gamma = \alpha + j\beta$ as shown schematically in Figure 5.23. Assume the transmission line's loss is small and its length is a multiple of a half-wavelength ($\ell = n\lambda/2$ with $n = 1, 2, \ldots$). The input impedance is obtained as

$$Z_{\text{in}} = Z_o \text{cotanh}(\alpha + j\beta)\ell = Z_o \frac{1 + j\tanh(\alpha\ell)\tan(\beta\ell)}{\tanh(\alpha\ell) + j\tan(\beta\ell)} \tag{5.114}$$

which, upon using (5.88), (5.91), (5.94), (5.96), and (5.99), becomes

$$Z_{\text{in}} \simeq Z_o \frac{1 + j\pi\alpha\ell\dfrac{\Delta\omega}{\omega_r}}{\alpha\ell + j\pi\dfrac{\Delta\omega}{\omega_r}} \simeq \frac{Z_o}{\alpha\ell + j\pi\dfrac{\Delta\omega}{\omega_r}} \simeq \frac{1}{\dfrac{1}{2}\ell\left(G' + \dfrac{R'C'}{L'}\right) + j\ell C'\Delta\omega} \tag{5.115}$$

within a narrow range of $\Delta\omega$ around the resonant frequency ω_r. Comparing (5.115) with the input impedance of a parallel lumped-element resonator given in (5.17) shows that a low loss half-wavelength open-circuited resonator behaves as a parallel resonator whose resistance, capacitance, and inductance are

$$R \simeq \frac{Z_o}{\alpha\ell} \simeq \frac{1}{\dfrac{1}{2}\ell\left(G' + \dfrac{R'C'}{L'}\right)} \tag{5.116}$$

$$C \simeq \frac{\pi}{2\omega_r Z_o} = \left(\frac{\pi}{2}\right)^2 \ell C' \tag{5.117}$$

and

$$L = \frac{1}{\omega_r^2 C} \simeq \left(\frac{2}{\pi}\right)^2 \ell L' \tag{5.118}$$

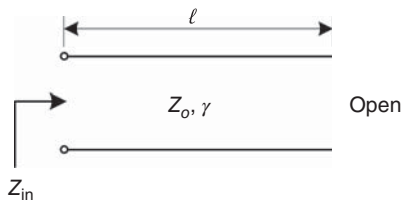


Figure 5.23. An open-circuited transmission line.

respectively, where $\ell = \lambda/2$ at ω_r. Equations (5.115) and (5.17) also suggest that the quality factor of a low loss half-wavelength open-circuited resonator is given as

$$Q \simeq \frac{\pi}{2\alpha\ell} \simeq \omega_r \frac{L'C'}{R'C' + G'L'} \simeq \frac{\beta}{2\alpha} \tag{5.119}$$

which is the same as that for a short-circuited counterpart, assuming same radiation loss.

Similarly, for low loss quarter-wavelength open-circuited resonators, whose lengths are an odd multiple of a quarter-wavelength ($\ell = (2n+1)\lambda/4$ with $n = 0, 1, 2, \ldots$), we can derive

$$Z_{\text{in}} \simeq Z_o \left(\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_r}\right) \simeq \frac{1}{2}\ell\left(R' + \frac{G'L'}{C'}\right) + j\frac{1}{2}\ell L'\Delta\omega \tag{5.120}$$

$$R \simeq Z_o\alpha\ell \simeq \frac{1}{2}\ell\left(R' + \frac{G'L'}{C'}\right) \tag{5.121}$$

$$L \simeq \frac{\pi Z_o}{4\omega_r} \simeq \left(\frac{\pi}{4}\right)^2 \ell L' \tag{5.122}$$

$$C = \frac{1}{\omega_r^2 L} \simeq \left(\frac{4}{\pi}\right)^2 \ell C' \tag{5.123}$$

and

$$Q \simeq \frac{\pi}{4\alpha\ell} \simeq 8\omega_r \frac{L'C'}{R'C' + G'L'} \simeq \frac{\beta}{2\alpha} \tag{5.124}$$

where $\ell = \lambda/4$ at ω_r. It is noted that the Q's of half-wavelength open- and short-circuited resonators are equal, assuming they have the same radiation losses.

The radiation quality factor of microstrip open-circuited half-wavelength resonators, as shown in Figure 5.24, can be estimated as [2]

$$Q_r \simeq \frac{3Z_o\varepsilon_{re}}{16\eta_a(h/\lambda_a)^2} \tag{5.125}$$

It is particularly noted that, as the currents and voltages along transmission lines are periodic functions with respect to location, multiple resonances are possible for transmission-line resonators. For half-wavelength resonators, the resonance would occur at harmonics of the (fundamental) resonant frequency ω_r corresponding to $\ell = n\lambda/2$, where $n = 1, 2, 3, \ldots$; that is, at frequencies where the length is a multiple of a half-wavelength. For quarter-wavelength resonators, however, the resonance only occurs at odd harmonics of the resonant frequency corresponding to $\ell = n\lambda/4$, where $n = 1, 3, \ldots$ Particularly for quarter-wavelength resonators, multiple resonances also occur at even harmonics, at which the resonators become half-wavelength resonators. Multi-resonance may cause some problems in circuits if they are not properly designed – for instance, possibly simultaneous oscillations in oscillators.

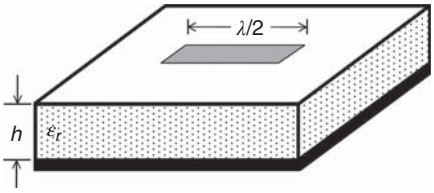


Figure 5.24. A microstrip open-circuited half-wavelength resonator.

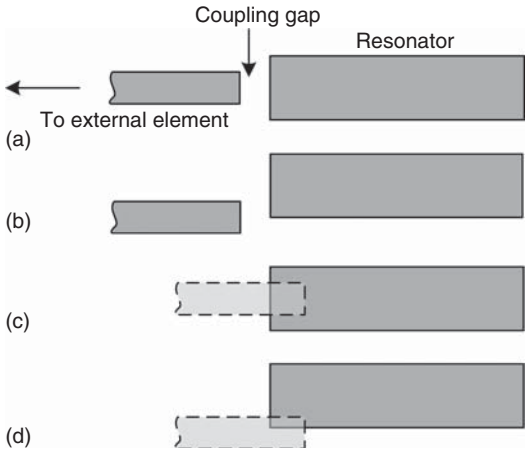


Figure 5.25. Gap couplings for transmission-line resonators. (a) coplanar symmetrical gap, (b) coplanar asymmetrical gap, (c) broadside symmetrical gap, and (d) broadside asymmetrical gap.

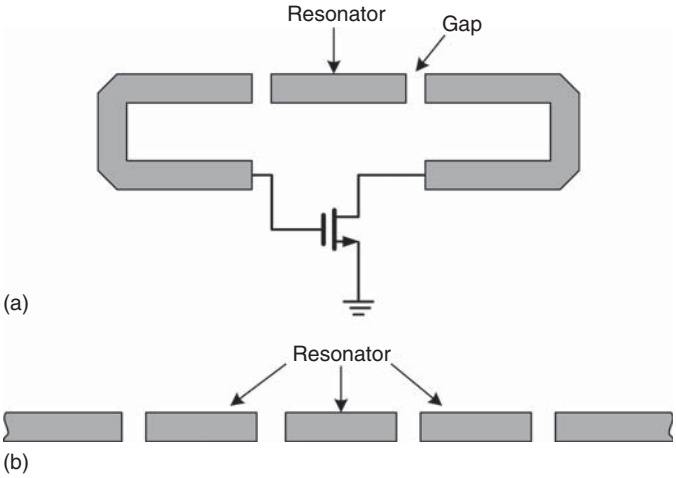


Figure 5.26. (a) A feedback MOSFET oscillator. (b) An end-coupled band-pass filter.

5.3.2.3 Gap-Coupling to Transmission-Line Resonators. One of the most commonly used coupling mechanisms between transmission-line resonators and external circuits or elements is gap, either on the same surface of the signal line (coplanar end-coupling) or on different surfaces via one or more dielectric layers (broadside-coupling). Figure 5.25 shows some gap-coupling configurations. Figure 5.26 illustrates a simple feedback MOSFET oscillator implementing a planar transmission-line resonator between the MOSFET's gate and drain and a band-pass filter employing half-wavelength open-circuited planar resonators. It is thus deemed useful to analyze the gap coupling for resonators.

Consider a gap between two transmission lines along with its equivalent circuit, consisting of a (series) coupling capacitance C_g and a (shunt) fringing capacitance C_p, as shown in Figure 5.27. This structure may be viewed as two parallel-coupled transmission lines along the vertical direction. The charge equations for the considered structure can be written as

$$Q_1 = (C_1 + C_p)V_1 + C_g(V_1 - V_2) \tag{5.126}$$

and

$$Q_2 = (C_1 + C_p)V_2 + C_g(V_2 - V_1) \tag{5.127}$$

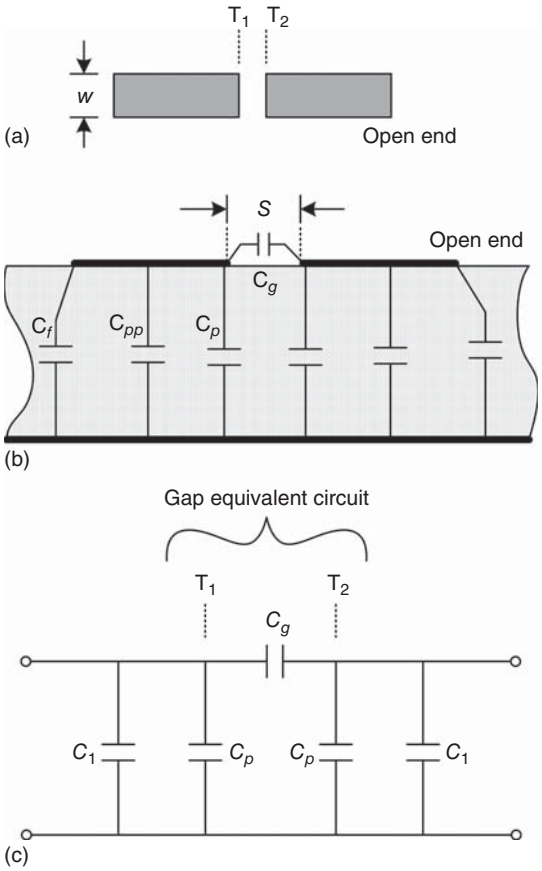


Figure 5.27. Top (a) and side (b) view of a transmission line gap and its corresponding equivalent circuit at reference planes T_1 and T_2 (c). Microstrip line is used in (b) to illustrate the structure, although any transmission line can be used.

where Q_i and V_i $(i = 1, 2)$ are the charge and potential on line i, respectively; and C_1 represents the total capacitance of each transmission line given as

$$C_1 = W(C_{pp} + C_f) \tag{5.128}$$

with C_{pp} representing the parallel-plate capacitance of the transmission line and C_f being the fringing capacitance at the open end. Imposing the potential condition $V_1 = V_2$ for the even-mode to (5.126) and (5.127), we obtain

$$\frac{Q_1}{V_1} = C_1 + C_p \tag{5.129}$$

Comparing with

$$\frac{Q_1}{V_1} \equiv WC_e = W(C_{pp} + C_f + C_{fe}) \tag{5.130}$$

where C_e and C_{fe} are the total even-mode capacitance and even-mode fringing capacitance per unit length of the two coupled lines, respectively, gives

$$C_p = WC_{fe} \tag{5.131}$$

after making use of (5.129). Applying the odd-mode excitation ($V_1 = V_2$), we can similarly derive

$$\frac{Q_1}{V_1} = W(C_{pp} + C_f + C_p + 2C_g) \tag{5.132}$$

$$\frac{Q_1}{V_1} \equiv WC_o = W(C_{pp} + C_f + C_{fo}) \tag{5.133}$$

from which, we can obtain

$$C_g = \frac{1}{2}W(C_{fo} - C_{fe}) \tag{5.134}$$

where C_o and C_{fo} are the total odd-mode capacitance and odd-mode fringing capacitance per unit length of the two coupled lines, respectively. C_{fo} and C_{fe} can be determined using closed-form equations for several planar transmission lines such as microstrip lines. They can also be calculated accurately using EM numerical methods or simulators. Closed-form equations of the equivalent gap capacitances C_g and C_p for various planar transmission lines can also be derived using techniques such as conformal mapping or fitting to numerical results. For asymmetric gaps having different widths at each end of the gap, approximate closed-form equations for the corresponding fringing capacitances C_{p1} and C_{p2} at the ends and the coupling capacitance C_g can also be developed.

The total capacitance C_{op} at the open end of the transmission line can be obtained, letting the gap width approach infinity, as

$$C_{op} = WC_{fe}|_{S\to\infty} \tag{5.135}$$

The fringing capacitance at the open end of open-circuited transmission-line resonators and those along the ends of the gap used as a coupling mechanism effectively increase the resonator's length, resulting in a lower resonant frequency. In practice, this effect is normally considered in circuit design by two procedures. One is to incorporate the fringing capacitances into the resonator's model and vary the length of the resonator to obtain the desired resonant frequency. The other is to estimate the fringing capacitances and their equivalent extra lengths and externally foreshorten the resonator to achieve the desired resonant frequency. Figure 5.28 illustrates the effects of open end and gap in increasing the length of open-circuited transmission-line resonators.

The extended length at each end of open-circuited transmission-line resonators can be approximated as

$$\Delta\ell_o \simeq v_p C_{op} Z_o \tag{5.136}$$

and the resultant resonant frequency,$\Delta\ell_o$ taking into account the two open ends, is

$$f_r \simeq \begin{cases} \dfrac{c}{2\sqrt{\varepsilon_{re}}(\ell + 2\Delta\ell_o)}, & \text{half-wavelength resonators} \\ \dfrac{c}{4\sqrt{\varepsilon_{re}}(\ell + 2\Delta\ell_o)}, & \text{quarter-wavelength resonators} \end{cases} \tag{5.137}$$

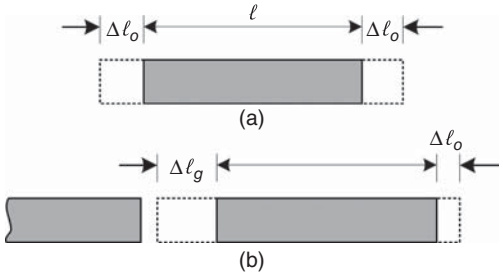


Figure 5.28. Extended lengths due two open ends (a) and an open end and gap (b).

where $c = 3 \times 10^8$ m/s and ε_{re} is the effective relative dielectric constant of the transmission line. At the gap, the approximate extended length is

$$\Delta\ell_g \simeq v_p C_p Z_o \tag{5.138}$$

which, together with $\Delta\ell_o$, lowers the resonant frequency.

5.3.2.4 Temperature Effects. In practical operations, particularly those exposed to wide ranges of temperatures, the resonant frequency of resonators, in general, and transmission line resonators, in particular, changes with temperature. Consider a transmission-line resonator having an electrical length of $\theta = s\pi$ at resonance. The corresponding resonant frequency is

$$f_r = \frac{sc}{2\ell\sqrt{\varepsilon_{re}}} \tag{5.139}$$

Taking derivative of f_r with respect to temperature T gives

$$\frac{df_r}{dT} = \frac{\partial f_r}{\partial \ell} \cdot \frac{\partial \ell}{\partial T} + \frac{\partial f_r}{\partial \varepsilon_{re}} \cdot \frac{\partial \varepsilon_{re}}{\partial T} \tag{5.140}$$

Substituting

$$\frac{\partial f_r}{\partial \ell} = -\frac{sc}{2\ell^2\sqrt{\varepsilon_{re}}} \tag{5.141}$$

$$\frac{\partial f_r}{\partial \varepsilon_{re}} = -\frac{sc}{4\ell(\varepsilon_{re})^{3/2}} \tag{5.142}$$

$$\frac{\partial \ell_r}{\partial T} = \ell T_\ell \tag{5.143}$$

and

$$\frac{\partial \varepsilon_{re}}{\partial T} = \frac{\varepsilon_{re}^2}{\varepsilon_r} T_\varepsilon \tag{5.144}$$

into (5.140) gives

$$\frac{df_r}{dT} = -\frac{sc}{2\ell\sqrt{\varepsilon_{re}}} T_\ell - \frac{sc\sqrt{\varepsilon_{re}}}{4\ell\varepsilon_r} T_\varepsilon \quad (\text{Hz} \cdot \text{ppm}/^\text{o}\text{C}) \tag{5.145}$$

where T_ℓ is the thermal expansion coefficient of the substrate (ppm/$^\text{o}$C) with *ppm* standing for *part-per-million per degree-celcius*; T_ε represents the temperature coefficient of the relative dielectric constant (ppm/$^\text{o}$C). The total frequency change in MHz/$^\text{o}$C can then be calculated from

$$\Delta f_r = \frac{df_r}{dT} \times 10^{12} \tag{5.146}$$

5.3.3 Waveguide Cavity Resonators

Waveguides are traditionally implemented using solid metallic walls and not applicable to printed circuits. As mentioned in Chapter 4 for Transmission Lines, advances in printed circuits, in general, and RFICs, in particular, allow waveguides to be realized compactly in multilayer CMOS structures. For instance, a millimeter-wave CMOS cavity resonator based on rectangular waveguide was developed in a CMOS process [3]. CMOS-based waveguide cavity resonators provide an alternate to transmission-line resonators for RFICs, especially those operating in the high end of the millimeter-wave frequency range, where the size is more suitable for RFICs. As such, studies of waveguide cavity resonators, typically only encountered in EM and microwave areas, are deemed to be useful and hence needed for RFIC design.

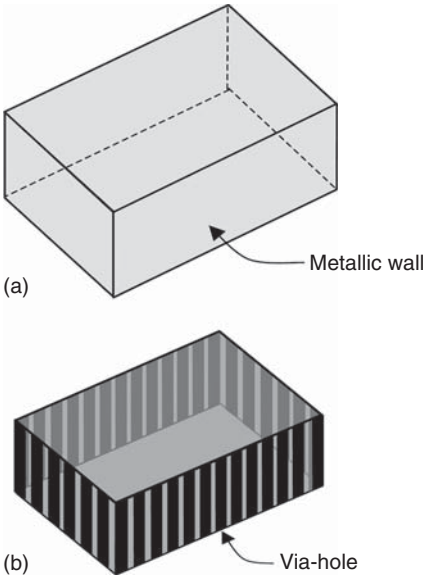


Figure 5.29. (a) Conventional rectangular cavity resonator using metallic walls. (b) Equivalent CMOS rectangular cavity resonator.

5.3.3.1 Rectangular Cavity Resonator. Figure 5.29 shows a conventional rectangular cavity resonator, consisting of a rectangular waveguide terminated with metallic walls at two ends, and an equivalent structure in CMOS technology. Figure 5.30 shows possible coupling mechanisms into a CMOS waveguide cavity resonator. Arrays of via-holes are used to simulate conventional metallic side-walls and interconnections. Horizontal walls may be realized by single or multiple metal strips in parallel, which may be interconnected by air-bridges, depending on the wall size. In typical CMOS processes, the maximum width of metal strips is in the order of several tens of a micron. However, if a pad layer is added onto a metal layer, then practically any width can be used for the metal strips within the pad coverage, making it very convenient to realize the top and bottom walls. Advanced CMOS technologies allow rectangular-shaped via-holes of very small dimensions (in the order of few tenths of a micron for each side of the via) to be placed very close to each other (in the order of few tenths of a micron from edge to edge), thus simulating well metallic walls.

We begin the analysis by placing two metallic walls at two ends ($z = 0, d$) of a rectangular waveguide having width a and height b and enclosing a dielectric with permittivity ϵ and permeability μ, as shown in Figure 5.31. Without loss of generality, we assume the conducting walls and dielectric are perfect materials and the structure only supports propagating modes (i.e., evanescent modes are nonexistent). Due to the end-walls, both the transverse electric (TE) and transverse magnetic (TM) modes consist of waves propagating in both positive and negative z-directions, creating standing waves within the rectangular cavity resonator. The analysis procedure is to determine the total electric and magnetic fields in cavity resonators using those for waveguides and apply the boundary conditions at the two end-walls of the cavity resonators. The results of boundary conditions will be used to determine the final fields and characteristics of the cavity resonators.

TM Modes

TM modes are characterized by zero longitudinal magnetic field; that is, $H_z(x, y, z) = 0$. The total longitudinal electric field component can be derived from those for rectangular waveguides as

$$E_z(x, y, z) = E_z^+ + E_z^- = \sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}(K^+e^{-\gamma_{mn}z} + K^-e^{\gamma_{mn}z}) \tag{5.147}$$

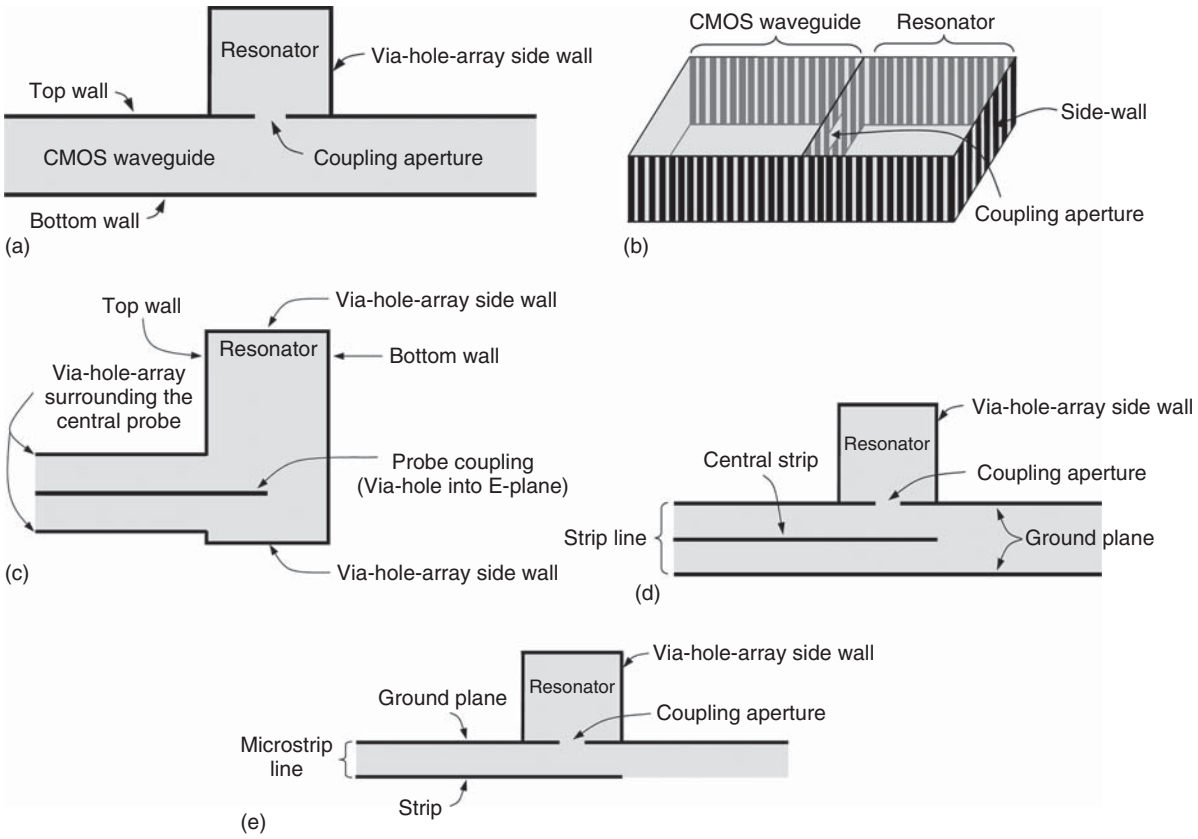


Figure 5.30. Coupling mechanisms to a CMOS rectangular cavity resonator from a CMOS waveguide (a) and (b), coaxial-like transmission line (c), stripline (d), and microstrip line (e).

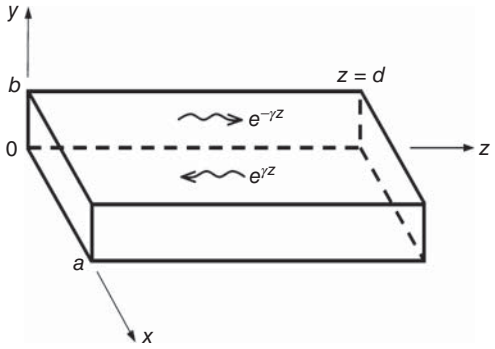


Figure 5.31. A rectangular cavity resonator.

where E_z^+ and E_z^- are the longitudinal electric fields of the forward (+) and backward (−) traveling waves in rectangular waveguides; K^+ and K^- are arbitrary constants corresponding to these waves, which can be determined from signals exciting the resonator; $m, n = 1, 2, 3, \ldots$; and $\gamma_{mn} = j\beta_{mn}$ is the propagation constant, with β_{mn} being the phase constant, given by

$$\gamma_{mn}^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - \omega^2\varepsilon\mu \tag{5.148}$$

The transverse field components can be derived from

$$
\begin{aligned}
\overset{+}{E}_y^{-}(x,y) &= \frac{1}{\gamma_{mn}^2+\omega^2\varepsilon\mu}\left[j\omega\mu\frac{\partial \overset{+}{H}_z^{-}(x,y)}{\partial x} \mp \gamma_{mn}^2\frac{\partial \overset{+}{E}_z^{-}(x,y)}{\partial y}\right] \\
\overset{+}{E}_x^{-}(x,y) &= -\frac{1}{\gamma_{mn}^2+\omega^2\varepsilon\mu}\left[j\omega\mu\frac{\partial \overset{+}{H}_z^{-}(x,y)}{\partial y} \pm \gamma_{mn}^2\frac{\partial \overset{+}{E}_z^{-}(x,y)}{\partial x}\right] \\
\overset{+}{H}_y^{-}(x,y) &= -\frac{1}{\gamma_{mn}^2+\omega^2\varepsilon\mu}\left[j\omega\varepsilon\frac{\partial \overset{+}{E}_z^{-}(x,y)}{\partial x} \pm \gamma_{mn}^2\frac{\partial \overset{+}{H}_z^{-}(x,y)}{\partial y}\right] \\
\overset{+}{H}_x^{-}(x,y) &= \frac{1}{\gamma_{mn}^2+\omega^2\varepsilon\mu}\left[j\omega\varepsilon\frac{\partial \overset{+}{E}_z^{-}(x,y)}{\partial y} \mp \gamma_{mn}^2\frac{\partial \overset{+}{H}_z^{-}(x,y)}{\partial x}\right]
\end{aligned}
\tag{5.149}
$$

where the superscripts + and − indicate the forward (+) and backward (−) traveling waves, respectively.

The total field components can be obtained, using (5.147)–(5.149), as

$$
\begin{aligned}
E_x(x,y,z) &= -\frac{M}{M^2+N^2}\gamma_{mn}\cos Mx\sin Ny(K^+e^{-\gamma z}-K^-e^{\gamma z}) \\
E_y(x,y,z) &= -\frac{N}{M^2+N^2}\gamma_{mn}\sin Mx\cos Ny(K^+e^{-\gamma z}-K^-e^{\gamma z}) \\
E_z(x,y,z) &= \sin Mx\sin Ny(K^+e^{-\gamma z}+K^-e^{\gamma z}) \\
H_x(x,y,z) &= \frac{j\omega\varepsilon N}{M^2+N^2}\sin Mx\cos Ny(K^+e^{-\gamma z}+K^-e^{\gamma z}) \\
H_y(x,y,z) &= -\frac{j\omega\varepsilon M}{M^2+N^2}\cos Mx\sin Ny(K^+e^{-\gamma z}+K^-e^{\gamma z}) \\
H_z(x,y,z) &= 0
\end{aligned}
\tag{5.150}
$$

where

$$
\begin{aligned}
M &= \frac{m\pi}{a}, \quad m = 1,2,3,\ \ldots \\
N &= \frac{n\pi}{b}, \quad n = 1,2,3,\ \ldots
\end{aligned}
\tag{5.151}
$$

These fields already satisfy the boundary conditions along the resonator's walls, except those at the ends. Applying the boundary conditions of $E_x = E_y = 0$ at the two end-walls, located at $z = 0$ and d, yields

$$
K^+ = K^- \equiv K \tag{5.152}
$$

and

$$\gamma_{mn} = j\beta_{mn} \equiv jP = j\frac{p\pi}{d} \tag{5.153}$$

respectively, where $p = 0, 1, 2, \ldots$

Substituting (5.152) and (5.153) into (5.150) leads to

$$\begin{aligned}
E_x(x,y,z) &= -2K\frac{MP}{M^2+N^2}\cos Mx \sin Ny \sin Pz \\
E_y(x,y,z) &= -2K\frac{NP}{M^2+N^2}\sin Mx \cos Ny \sin Pz \\
E_z(x,y,z) &= 2K \sin Mx \sin Ny \cos Pz \\
H_x(x,y,z) &= j2\omega\varepsilon K\frac{N}{M^2+N^2}\sin Mx \cos Ny \cos Pz \\
H_y(x,y,z) &= -j2\omega\varepsilon K\frac{M}{M^2+N^2}\cos Mx \sin Ny \cos Pz
\end{aligned} \tag{5.154}$$

There are no propagation terms $e^{\pm\gamma_{mn}z}$ in the field expressions in (5.154), indicating that only standing waves exist within the rectangular cavity resonators as expected. Each set of (m, n, p) corresponds to a mode with its own field distributions, thus leading to an infinite number of TM_{mnp} modes with $m = n = 1, 2, 3 \ldots$ and $p = 0, 1, 2, \ldots$ The time-domain fields can be obtained by multiplying the corresponding fields with $e^{j\omega t}$ and taking the real part, resulting in sine or cosine of ωt, signifying the signal's oscillation phenomenon in the resonators.

It is recalled that the propagation constant of a TM_{mn} wave propagating in a rectangular waveguide, given in Eq. (5.148), is obtained by enforcing the boundary conditions of electric fields along its four conducting walls. By imposing additional boundary conditions at the two conducting end-walls of a rectangular cavity resonator, the propagation constant becomes that derived in (5.153). So in order for the fields existing in the resonator to satisfy all the required boundary conditions, the following equation must be held:

$$\left(\frac{p\pi}{d}\right)^2 = \omega^2\varepsilon\mu - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2 \tag{5.155}$$

which implies that the fields inside the resonator for a particular TM_{mnp} mode can exist only at a specific resonant frequency f_{mnp} given by

$$f_{mnp} = \frac{1}{2\pi\sqrt{\varepsilon\mu}}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{d}\right)^2} \tag{5.156}$$

or

$$f_{mnp} = \frac{v}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 + \left(\frac{p}{d}\right)^2} \tag{5.157}$$

using $v = 1/\sqrt{\varepsilon\mu}$. From (5.153), the rectangular cavity resonator's length can be obtained as

$$d = \frac{p\pi}{\beta_{mn}} = \frac{p\lambda_{mn}}{2} \tag{5.158}$$

where $p = 1, 2, 3, \ldots$ and λ_{mn} is the wavelength of the TM_{mn} wave propagating in the corresponding rectangular waveguide at the resonant frequency.

TE Modes

For TE modes, the longitudinal electric field $E_z(x, y, z) = 0$. Following the same analysis for TM modes, we can derive the following equations for the field distributions:

$$\begin{aligned}
E_x(x, y, z) &= -2\omega\mu K \frac{N}{M^2 + N^2} \cos Mx \sin Ny \sin Pz \\
E_y(x, y, z) &= 2\omega\mu K \frac{M}{M^2 + N^2} \sin Mx \cos Ny \sin Pz \\
E_z(x, y, z) &= 0 \\
H_x(x, y, z) &= -j2K \frac{MP}{M^2 + N^2} \sin Mx \cos Ny \cos Pz \\
H_y(x, y, z) &= -j2K \frac{NP}{M^2 + N^2} \cos Mx \sin Ny \cos Pz \\
H_z(x, y, z) &= j2K \cos Mx \cos Ny \sin Pz
\end{aligned} \tag{5.159}$$

where

$$M = \frac{m\pi}{a}, \quad m = 0, 1, 2, \ldots$$

$$N = \frac{n\pi}{b}, \quad n = 0, 1, 2, \ldots$$

$$P = \frac{p\pi}{d}, \quad p = 1, 2, 3, \ldots \tag{5.160}$$

Note that m and n cannot be zero simultaneously. The field distributions, as given in (5.159), are unique for each TE_{mnp} mode and represent standing waves corresponding to possible infinite TE_{mnp} modes within the cavity resonator. The time-domain responses of the fields can be plotted as function of time to show the oscillation within the resonator. The resonant frequencies corresponding to TE_{mnp} modes are found to be the same as those for TE_{mnp} modes, as given in (5.156) and (5.157), at which there exist unique field distributions inside the resonator, as given in (5.159).

TE_{101} Mode. If $a \geq d > b$ or $a > b > d$, which is typical for rectangular cavity resonators, the resonant frequency f_{101} has the lowest value, and the corresponding TE_{101} mode is referred to as the *dominant mode*. The fields and resonant frequency for TE_{101} mode can be obtained from (5.159) and (5.156), respectively, as

$$\begin{aligned}
E_y(x, y, z) &= \frac{2\omega\mu a}{\pi} K \sin\left(\frac{\pi x}{a}\right) \sin\left(\frac{\pi z}{d}\right) \\
E_x(x, y, z) &= E_z(x, y, z) = 0 \\
H_x(x, y, z) &= -j\frac{2a}{d} K \sin\left(\frac{\pi x}{a}\right) \cos\left(\frac{\pi z}{d}\right) \\
H_y(x, y, z) &= 0 \\
H_z(x, y, z) &= j2K \cos\left(\frac{\pi x}{a}\right) \sin\left(\frac{\pi z}{d}\right)
\end{aligned} \tag{5.161}$$

and

$$f_{101} = \frac{1}{2\sqrt{\varepsilon\mu}} \sqrt{\frac{1}{a^2} + \frac{1}{d^2}} \tag{5.162}$$

Now assume that the conducting walls are made of a (good) low loss conductor having finite conductivity σ and the enclosed dielectric has $\varepsilon = \varepsilon' + j\varepsilon''$ with $\varepsilon' = \varepsilon_o\varepsilon_r$. The time-average energy stored in the cavity resonator for the TE_{101} mode can be obtained from (5.83) and (5.161) as

$$W = \frac{\varepsilon'}{2}\int_0^a \int_0^b \int_0^d |E_y^2| dxdydz = \frac{\mu abd}{2}|K|^2\left[1 + \left(\frac{a}{d}\right)^2\right] \tag{5.163}$$

Since the conductor loss is small, the power loss due to conducting walls can be approximately determined by employing a perturbation procedure, which assumes the current flow on the conductor is the same as that flowing on a (lossless) perfect conductor. This approximate power loss can thus be obtained from (5.84) as

$$P_c \simeq \sum_{n=1}^{6} \frac{R_{sn}}{2}\iint_{S_n} |\vec{H}_{t,n}|^2 \, dS \tag{5.164}$$

where $R_{sn} = R_s = 1/\sigma\delta_s$ is the surface resistivity of the conducting walls, with δ_s being the skin depth of the conductor. Taking the integrals all over the walls gives

$$\begin{aligned} P_c \simeq \frac{R_s}{2}\Bigg\{ & 2\int_0^b \int_0^a |H_x(z=0)|^2 dxdy + 2\int_0^d \int_0^b |H_z(x=0)|^2 dydz \\ & + 2\int_0^d \int_0^a \left[|H_x(y=0)|^2 + |H_z(y=0)|^2\right] dxdz \Bigg\} \end{aligned} \tag{5.165}$$

which becomes

$$P_c \simeq R_s|K|^2 \frac{2b(a^3+d^3) + ad(a^2+d^2)}{d^2} \tag{5.166}$$

The power loss due to dielectric can be derived from (5.85) using (5.161) as

$$P_d = \frac{1}{2}\omega_r\varepsilon_o\varepsilon_r''\iiint_V |\vec{E}|^2 \, dV = \frac{\omega\varepsilon_o\varepsilon_r'' abd}{8}|E_o|^2 \tag{5.167}$$

where $E_o = \frac{2\omega\mu a}{\pi}K$.

The quality factor due to conductor loss can be obtained from (5.48), (5.163), and (5.166) as

$$Q_c \simeq \frac{f_{101}}{R_s}\frac{\pi\mu abd(a^2+d^2)}{2b(a^3+d^3) + ad(a^2+d^2)} \tag{5.168}$$

which is proportional to the square-root of the resonant frequency. The quality factor due to the loss of dielectric enclosed by perfectly conducting walls is obtained from (5.49), (5.163), and (5.167) as

$$Q_d = \frac{\varepsilon'}{\varepsilon''} = \frac{1}{\tan\delta} \tag{5.169}$$

The total Q of the cavity resonator, considering losses from both conducting walls and dielectric, is obtained, from (5.52) with $Q_r \to \infty$, as

$$Q = \frac{Q_c Q_d}{Q_c + Q_d} \tag{5.170}$$

As can be expected from a large volume contained within the cavity resonator and typically employed good conductor and dielectric (usually, air), the Q of rectangular cavity resonators is much higher than those

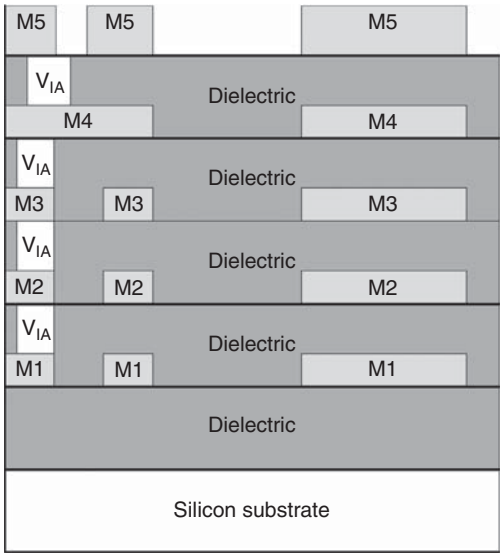


Figure 5.32. Cross section of a typical CMOS structure showing metal layers from the bottom-most M1 to the top-most M5 and via-hole interconnects.

made by lumped elements or transmission lines. In CMOS technology, however, the height of cavity resonators is dictated by the thickness of the dielectric layers between different metal layers and is small. This, in conjunction with very thin metal layers, causes limited Q for CMOS-based cavity resonators, as can be seen in the following example of a millimeter-wave CMOS rectangular cavity resonator.

An Example of CMOS Rectangular Cavity Resonators. A rectangular cavity resonator was developed at millimeter-wave frequencies using a commercial standard 0.25-μm CMOS process [3]. Figure 5.32 shows the cross section of a typical CMOS configuration showing possible metal layers and via-hole interconnects. Figure 5.33 illustrates the developed CMOS cavity resonator, which is based on a conventional rectangular waveguide cavity with the addition of a capacitive load. The top and bottom metal walls of the cavity are on metal layers M4 and M1, as seen in Figure 5.32, respectively. The side-walls of the cavity extending from the top to the bottom walls are formed using lattices of periodic metallic via-holes having minimum distance between them. The connections between M1 and M2, M2 and M3, and M3 and M4 needed for the side-walls are made using three different groups of via-holes, considering available structure of the employed process. Two narrow metal frames on M2 and M3 are also used to connect all the via-holes together to further confine the fields within the cavity and enhance the electrical isolation between the interior and exterior of the cavity. It is noted that the size of the via-holes and distance between them in typical CMOS processes are extremely small with respect to a millimeter-wave wavelength. The via-hole lattices, together with the metal frames, can thus create side-walls resembling well solid metallic walls electrically and hence can be used to replace them. Two 50-Ω microstrip lines with meandered open stubs on M5 couple the energy to and from the cavity, through two H-shaped coupling slots formed on the top metal wall of the cavity, and are used as coupling and interface mechanisms between the resonator and other CMOS on-chip elements. A capacitive post is inserted into the middle of the cavity and formed by a metallic patch on M3 and lattices of periodic metallic via-holes connecting the periphery of the metallic patch with the bottom wall of the cavity. Two different lattices of via-holes are used to connect the bottom cavity wall with the metal layer M2 and then to the edges of the metallic patch. The miniature size of the via-holes and the very tiny gap between them ensure that these lattices of metallic via-holes provide a good electrical wall around the space between the metallic patch and the bottom cavity wall. A narrow metal frame on M2 around this enclosed space is also used to connect all the via-holes together to further improve the electrical wall and hence the capacitive property of the post. The capacitive post enhances the overall capacitance of the resonator by

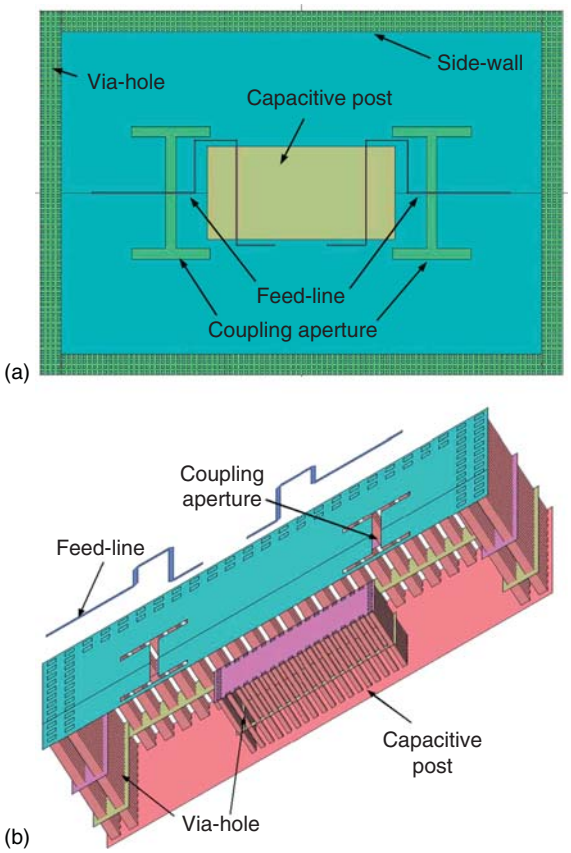


Figure 5.33. Sketch of the top-view (a) and three-dimensional (3D) view (b) of CMOS capacitively loaded cavity resonator (not to scale). The front wall in the 3D view is removed to show the inside. (After Miao and Nguyen [2]. Reprinted with permission of IEEE.)

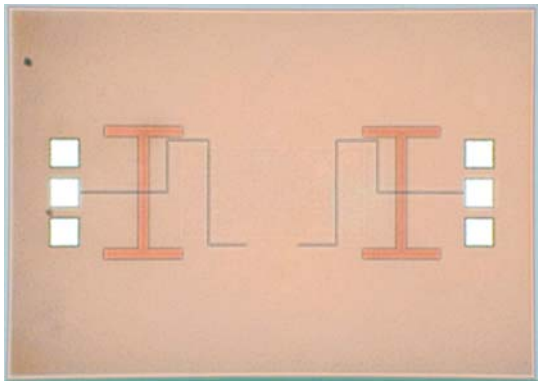


Figure 5.34. Photograph of the fabricated CMOS capacitively loaded cavity resonator showing the H-shaped coupling apertures, feed-lines, and RF pads. The cavity is underneath and thus not visible. (After Miao and Nguyen [2]. Reprinted with permission of IEEE.)

adding capacitive effect to the resonator, resulting in a reduction of the resonant frequency of the dominant TE_{101} mode of the CMOS cavity resonator.

Figure 5.34 shows a photograph of the fabricated CMOS cavity resonator. The complete cavity resonator, including two coupling apertures and two microstrip feed-lines, has lateral dimensions of 2 mm × 1.4 mm and a height of 6.28 μm. The size of the inserted rectangular metallic patch is 720 μm × 400 μm. RF pads connecting to the 50-Ω microstrip feed-lines, used for on-wafer measurements, are placed just above the cavity to save space.

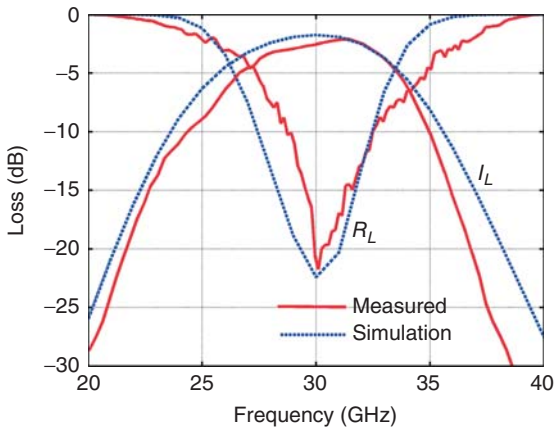


Figure 5.35. Measured and simulated return loss and insertion loss of the CMOS capacitively loaded cavity resonator. (After Miao and Nguyen [2]. Reprinted with permission of IEEE.)

Figure 5.35 shows the measured and simulated results of the developed CMOS cavity resonator. The measured resonance occurs at around 30 GHz with an insertion loss of 2 dB. The loss caused by the feed-lines is removed from the measured insertion loss. The measured and calculated unloaded quality factors are derived as 22.2 and 24.8 using the measured and calculated S-parameters, respectively. This unloaded Q is very low as compared to conventional discrete cavity resonators due to the fact that the height of the CMOS cavity resonator is only 6.28 μm, much smaller than that used in conventional cavity structures. This significantly smaller height results in reduced Q according to formula (5.168) for the cavity's Q. Larger heights are not possible in current CMOS processes due to a very small distance between the top-most and bottom-most metal layers. Nevertheless, the achieved Q is still higher than that of typical resonators made by CMOS lumped elements at 30 GHz and, particularly, CMOS cavity resonators should be useful for various millimeter-wave RFICs, where frequencies are so high that good lumped elements cannot be realized, or when waveguide circuit topologies, such as direct-coupled waveguide band-pass filters, need to be fabricated in CMOS processes where other high Q techniques cannot be used. Furthermore, it is expected that advances in modern RFIC fabrication technology will allow more metal layers and increase distance between the top- and bottom-most metal layers, which can be exploited to consequently improve the Q of CMOS cavity resonators and facilitate their applications for RFICs.

5.3.3.2 Circular Cavity Resonator. Figure 5.36 shows a conventional circular cavity resonator made from a circular waveguide and an equivalent structure in CMOS technology. Figure 5.37 shows several possible coupling mechanisms into a CMOS circular cavity resonator. As in CMOS rectangular cavity resonators, arrays of via-holes of very small dimensions placed very close to each other are used to realize good interconnections and walls which resemble solid metallic walls. Similar to rectangular cavity resonators, circular cavity resonators support both TE and TM modes.

TE Modes

We consider a circular waveguide cavity resonator with radius a and end-walls at $z = 0$ and d, as shown in Figure 5.38. We also assume the conductors and dielectric are perfect. We follow the same approach as used for rectangular cavity resonators, starting with the TE fields for circular waveguides, summing the electric fields corresponding to the forward and backward propagating waves, enforcing the boundary conditions of zero tangential electric field at the end-walls, and then deriving the total electric and magnetic fields.

The total electric and magnetic field components in a circular cavity resonator can be obtained from the summation of the individual forward and backward field components as

$$E_r(r,\phi,z) = \frac{jm\beta_{mn}Z_{\mathrm{TE}_{mn}}}{r(k^h_{c,mn})^2} J_m\left(\frac{x'_{mn}r}{a}\right)\sin(m\phi)(C^+e^{-j\beta_{mn}z} + C^-e^{j\beta_{mn}z})$$

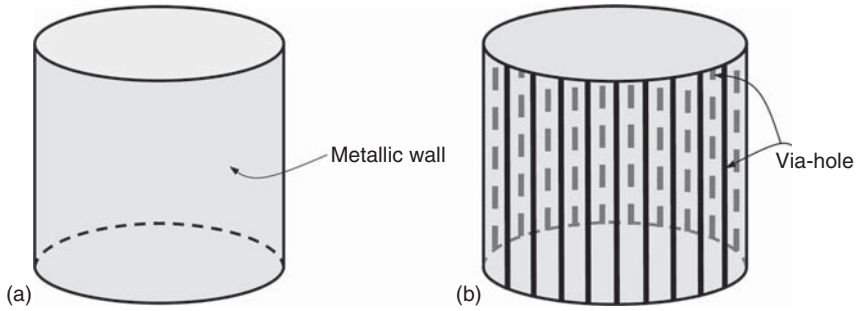


Figure 5.36. (a) Conventional circular cavity resonator using metallic walls. (b) Equivalent CMOS circular cavity resonator. Via-holes are used to simulate side-walls. Top and bottom walls may be realized by single or multiple metal strips in parallel depending on the wall size.

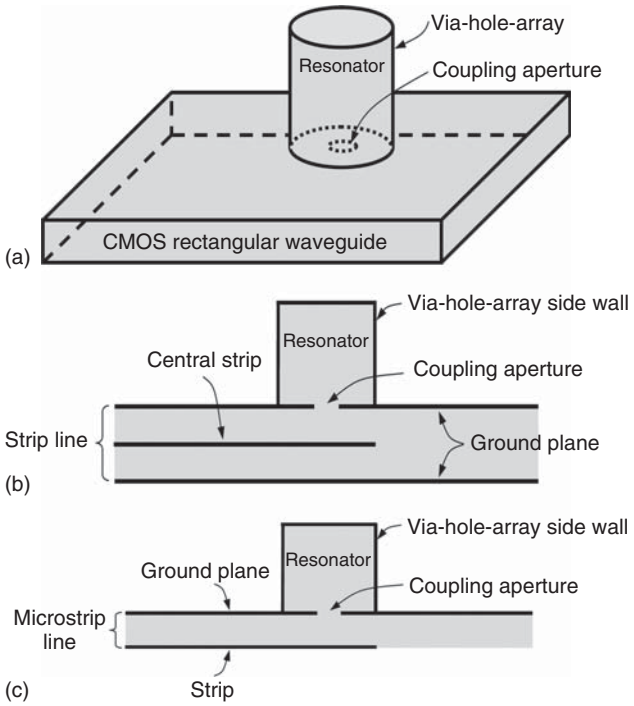


Figure 5.37. Coupling mechanisms to a CMOS circular cavity resonator from a CMOS rectangular waveguide (a), stripline (b), and microstrip line (c).

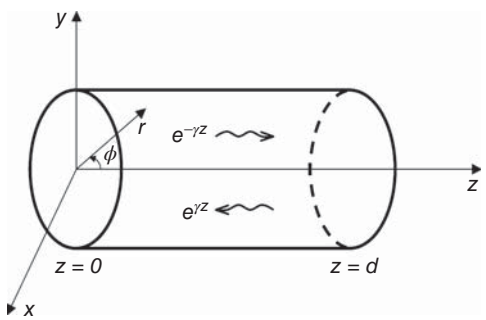


Figure 5.38. A circular cavity resonator.

$$
\begin{aligned}
E_\phi(r,\phi,z) &= \frac{j\beta_{mn}x'_{mn}Z_{\mathrm{TE}_{mn}}}{a(k^h_{c,mn})^2}J'_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)(C^+e^{-j\beta_{mn}z}+C^-e^{j\beta_{mn}z})\\
E_z(r,\phi,z) &= 0\\
H_r(r,\phi,z) &= -\frac{j\beta_{mn}}{k^h_{c,mn}}J'_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)(C^+e^{-j\beta_{mn}z}-C^-e^{j\beta_{mn}z})\\
H_\phi(r,\phi,z) &= \frac{jm\beta_{mn}}{r(k^h_{c,mn})^2}J_m\left(\frac{x'_{mn}r}{a}\right)\sin(m\phi)(C^+e^{-j\beta_{mn}z}+C^-e^{j\beta_{mn}z})\\
H_z(r,\phi,z) &= J_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)(C^+e^{-j\beta_{mn}z}+C^-e^{j\beta_{mn}z})
\end{aligned}
\tag{5.171}
$$

where $Z_{\mathrm{TE}_{mn}}$ is the characteristic wave impedance of TE modes; $m = 0, 1, 2, \ldots$; $n = 1, 2, 3, \ldots$; C^+ and C^- are arbitrary constants corresponding to forward and backward traveling waves, respectively; J_m is the Bessel function of the first kind of order m; $J'_m \equiv dJ_m(k^h_{c,mn}r)/dr$ is the derivative of J_m with respective to r; x'_{mn} is the root of $J'_m = 0$; $k^h_{c,mn} = x'_{mn}/a$; and β_{mn} is the phase constant given by

$$\beta_{mn} = \sqrt{k^2 - \left(\frac{x'_{mn}}{a}\right)^2} \tag{5.172}$$

where $k = \omega\sqrt{\varepsilon\mu}$, with ε and μ being the respective permittivity and permeability of the dielectric, is the wave number. Note that these fields already satisfy the boundary conditions along the resonator's circular wall. Now applying the remaining boundary conditions of $E_r = E_\phi = 0$ at $z = 0$ and d, we obtain

$$C^+ = -C^- \tag{5.173}$$

and

$$\beta_{mn} = \frac{p\pi}{d} \tag{5.174}$$

respectively, where $p = 1, 2, 3, \ldots$ Substituting (5.173) and (5.174) into (5.171), letting $C = -2jC^+$, and replacing $Z_{\mathrm{TE}_{mn}}$ with $k\eta/\beta_{mn}$ result in the following field equations for TE_{mnp} modes:

$$
\begin{aligned}
E_r(r,\phi,z) &= jC\frac{k\eta a^2 m}{r(x'_{mn})^2}J_m\left(\frac{x'_{mn}r}{a}\right)\sin(m\phi)\sin\left(\frac{p\pi}{d}z\right)\\
E_\phi(r,\phi,z) &= jC\frac{k\eta a}{x'_{mn}}J'_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)\sin\left(\frac{p\pi}{d}z\right)\\
E_z(r,\phi,z) &= 0\\
H_r(r,\phi,z) &= C\frac{\beta_{mn}a}{x'_{mn}}J'_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)\cos\left(\frac{p\pi}{d}z\right)\\
H_\phi(r,\phi,z) &= -C\frac{\beta_{mn}a^2 m}{r(x'_{mn})^2}J_m\left(\frac{x'_{mn}r}{a}\right)\sin(m\phi)\cos\left(\frac{p\pi}{d}z\right)\\
H_z(r,\phi,z) &= CJ_m\left(\frac{x'_{mn}r}{a}\right)\cos(m\phi)\sin\left(\frac{p\pi}{d}z\right)
\end{aligned}
\tag{5.175}
$$

where C is an arbitrary constant. As expected, there are no propagation terms in the field expressions indicating the existence of only standing waves inside the resonator.

The length of the cavity resonator is obtained from (5.174) as

$$d = \frac{p\pi}{\beta_{mn}} = \frac{p\lambda_{mn}}{2} \tag{5.176}$$

where λ_{mn} is the wavelength of the TE_{mn} wave propagating in the corresponding circular waveguide at the resonant frequency.

Equations (5.172) and (5.174) were derived from the boundary conditions along the circular wall and at the end-walls, respectively. Considering the fact that all the fields within the resonator must satisfy these boundary conditions, we can then write:

$$\left(\frac{p\pi}{d}\right)^2 = \omega^2\varepsilon\mu - \left(\frac{x'_{mn}}{a}\right)^2 \tag{5.177}$$

The resonant frequency, at which fields exist for a TE_{mn} mode, can thus be derived as

$$f^{\mathrm{TE}}_{mnp} = \frac{1}{2\pi\sqrt{\varepsilon\mu}}\sqrt{\left(\frac{x'_{mn}}{a}\right)^2 + \left(\frac{p\pi}{d}\right)^2} \tag{5.178}$$

As for rectangular cavity resonators, there exist possibly infinite TE_{mn} modes within a circular cavity resonator, with each mode having its own field distributions. Among these modes, TE_{111} has the lowest resonant frequency and is known as the dominant mode in circular cavity resonators, indeed corresponding to the dominant TE_{11} mode in circular waveguides.

Now consider conducting walls made of a (good) low loss conductor having finite conductivity σ and a lossy dielectric with permittivity $\varepsilon = \varepsilon' + j\varepsilon''$. The time-average energy stored inside a circular cavity resonator can be derived from (5.83) and (5.175) as

$$\begin{aligned} W &= \frac{\varepsilon'}{2}\int_0^a\int_0^{2\pi}\int_0^d (|E_r|^2 + |E_\phi|^2) r dz d\phi dr \\ &= C^2\frac{\varepsilon'\pi k^2\eta^2 a^2 d}{4(x'_{mn})^2}\left\{\int_0^a\left[J'^{\,2}_m\left(\frac{x'_{mn}r}{a}\right) + \left(\frac{ma}{x'_{mn}r}\right)^2 J_m^2\left(\frac{x'_{mn}r}{a}\right)\right] r dr\right\} \\ &= C^2\frac{\varepsilon'\pi k^2\eta^2 a^4 d}{8(x'_{mn})^2}\left[1 - \left(\frac{m}{x'_{mn}}\right)^2\right] J_m^2(x'_{mn}) \end{aligned} \tag{5.179}$$

making use of

$$\int_0^{x'_{mn}}\left[J'^{\,2}_m(x) + \frac{m^2}{x^2}J_m^2(x)\right] x dx = \frac{x'^{\,2}_{mn}}{2}\left(1 - \frac{m^2}{x'^{\,2}_{mn}}\right) J_m^2(x'_{mn}) \tag{5.180}$$

Applying the perturbation theory for small loss, as done in rectangular cavity resonators, we can write the power loss due to conductor as

$$\begin{aligned} P_c &\simeq \frac{R_s}{2}\left\{\int_0^d\int_0^{2\pi}\left[|H_\phi(r=a)|^2 + |H_z(r=a)|^2\right] a d\phi dz + 2\int_0^a\int_0^{2\pi}\left[|H_r(z=0)|^2 + |H_\phi(z=0)|^2\right] r dr d\phi\right\} \\ &\simeq C^2\pi\frac{R_s}{2}J_m^2\left(x'_{mn}\right)\left\{\frac{ad}{2}\left[1 + \left(\frac{\beta_{mn}ma}{x'_{mn}}\right)^2\right] + \left(\frac{\beta_{mn}}{x'_{mn}}\right)^2\left(1 - \frac{m^2}{x'^{\,2}_{mn}}\right)\right\} \end{aligned} \tag{5.181}$$

The quality factor due to conductor loss can then be derived from the energy W and power loss P_c as

$$Q_c \simeq \frac{c}{2\pi f_{mnp}\delta_s\sqrt{\varepsilon_r}} \frac{\left[1-\left(\frac{m}{x'_{mn}}\right)^2\right]\left[x'^2_{mn}+\left(\frac{a}{d}\pi p\right)^2\right]^{3/2}}{x'^2_{mn}+\frac{2a}{d}\left(\frac{a}{d}\pi p\right)^2+\left(1-\frac{2a}{d}\right)\left(\frac{a}{d}\frac{mp\pi}{x'_{mn}}\right)^2} \tag{5.182}$$

which is inversely proportional to the square-roof of the resonant frequency. For the dominant TE_{111} mode, occurring when $d/a \geq 2$, Q_c is obtained as

$$Q_c = \frac{0.27\lambda_{mnp}}{\delta_s} \tag{5.183}$$

where $\lambda_{mnp} = c/f_{mnp}\sqrt{\varepsilon_r}$.

The power loss due to dielectric loss can be derived from (5.85) and (5.175) as

$$P_d \simeq \frac{\omega_{mnp}\varepsilon'' k^2\eta^2 a^4}{8x'^2_{mn}} C^2\left[1-\left(\frac{m}{x'^2_{mn}}\right)^2\right] J_m^2(x'_{mn}) \tag{5.184}$$

which, along with the energy W, gives the same quality factor due to dielectric for rectangular cavity resonators as obtained in (5.169). The total quality factor due to both conductor and dielectric losses can then be obtained from the usual formula (5.170).

The conductor Q of circular cavity resonators, as examined from (5.182), for TE_{011} (TE_{012}) and TE_{111} modes reaches a maximum when the length of the resonator is approximately equal to its diameter and 2/3 of the diameter, respectively, a fact useful for the design of circular cavity resonators. It is particularly noted that, for given frequency and resonator's dimensions, different modes have different values for Q – for instance, the Q of the TE_{011} mode is about two to three times of that for the TE_{111} mode and the Q of the TE_{012} mode is higher than that of the TE_{011} mode. Regarding high Q consideration, a particular mode may thus need to be chosen in lieu of the usual dominant TE_{111} mode. Using a nondominant mode, however, may result in unwanted resonances caused by other modes that may be excited within the frequencies of interest. Therefore, if a nondominant mode such as TE_{011} is selected, proper values for a and d need to be chosen so that less modes can be excited within a particular frequency range. Furthermore, a proper coupling mechanism for the chosen mode must also be implemented to avoid possible excitations of other modes within the frequencies of interest.

TM Modes

Following the same procedure for TE modes, we can derive the field distributions and other parameters for circular cavity resonators. The fields are given as

$$\begin{aligned}
E_r(r,\phi,z) &= -jB\frac{\beta_{mn}a}{x_{mn}}J'_m\left(\frac{x_{mn}r}{a}\right)\cos(m\phi)\sin\left(\frac{p\pi}{d}z\right)\\
E_\phi(r,\phi,z) &= -jB\frac{\beta_{mn}a^2m}{rx^2_{mn}}J_m\left(\frac{x_{mn}r}{a}\right)\sin(m\phi)\sin\left(\frac{p\pi}{d}z\right)\\
E_z(r,\phi,z) &= jBJ_m\left(\frac{x_{mn}r}{a}\right)\cos(m\phi)\cos\left(\frac{p\pi}{d}z\right)\\
H_r(r,\phi,z) &= -B\frac{ka^2m}{r\eta x^2_{mn}}J_m\left(\frac{x_{mn}r}{a}\right)\sin(m\phi)\cos\left(\frac{p\pi}{d}z\right)
\end{aligned} \tag{5.185}$$

$$H_\phi(r,\phi,z) = B\frac{ka}{\mu x_{mn}} J'_m\left(\frac{x_{mn}r}{a}\right)\cos(m\phi)\sin\left(\frac{p\pi}{d}z\right)$$
$$H_z(r,\phi,z) = 0$$

where $m = 0, 1, 2, \ldots$; $n = 1, 2, 3, \ldots$; $p = 0, 1, 2, \ldots$; B is an arbitrary constant; and β_{mn} is the phase constant of the corresponding TM_{mn} mode given by

$$\beta_{mn} = \sqrt{k^2 - \left(\frac{x_{mn}}{a}\right)^2} \tag{5.186}$$

These field distributions are unique for each of the TM_{mnp} modes and exist only at the resonant frequency

$$f_{mnp}^{\text{TM}} = \frac{1}{2\pi\sqrt{\varepsilon\mu}}\sqrt{\left(\frac{x_{mn}}{a}\right)^2 + \left(\frac{p\pi}{d}\right)^2} \tag{5.187}$$

Note that the equations for the phase constant and resonant frequency for TM modes are the same as those for TM_{mnp} modes, given in (7.65) and (7.171), with x'_{mn} being replaced with x_{mn}.

The quality factor due to conductor loss for TM_{mnp} modes can be derived as

$$Q_c \simeq \frac{\eta\sqrt{x_{mn}^2 + \left(\frac{p\pi a}{d}\right)^2}}{2R_m(1 + 2a/d)} \tag{5.188}$$

while the dielectric quality factor is the same as that for the TE_{mnp} modes. For the dominant TM_{010}, occurring when $d/a < 2$, Q_c can be obtained from (5.188) with $x_{01} = 2.405$ as

$$Q_c = \frac{1.202\eta}{R_m(1 + a/d)} \tag{5.189}$$

It is noted that, while the Q of a rectangular cavity resonator operating in the TE and TM modes is a function of the resonator's individual dimensions a, b, and d, the Q of a circular cavity resonator for the TE and TM modes depends on the resonator's dimension ratio a/d. At first, we might think that the problem of limited physical dimensions in CMOS/BiCMOS structures, which make it very difficult to realize high Q rectangular cavity resonators mentioned earlier, can be overcome for high Q circular cavity resonators since their Q only depends on the ratio between the diameter and length. However, as the resonant frequency still depends on a and d individually as seen from (5.178) and (5.187), it is still very difficult to design high Q circular cavity resonators in CMOS/BiCMOS structures at low frequencies even in the millimeter-wave region. It is possible, however, to realize high Q RFIC circular cavity resonators at extremely high frequencies in the terahertz range.

Examining the cutoff frequencies for the TE and TM modes given in (5.178) and (5.187), respectively, reveals that, for $d/a \geq 2$, the TE_{111} (corresponding to $x'_{11} = 1.841$) is the dominant mode, while, for $d/a < 2$, the TM_{010} (corresponding to $x_{01} = 2.405$) is dominant. This is especially interesting and important for CMOS applications since, due to limited numbers of metal layers and distances between them in CMOS processes, d is much more limited than a. As a results, $d/a < 2$ is easier to obtain than $d/a \geq 2$, signifying that it is better to design a circular cavity resonator TM_{010} in CMOS/BiCMOS structures for the TM_{010} mode operation than for the TE_{111} mode.

As an example, we consider a circular cavity resonator having $d = 6.4\,\mu\text{m}$ and $a = 1\,\text{mm}$, copper wall with $\sigma = 4\pi \times 10^7\,\text{S/m}$, and SiO_2 dielectric with $\varepsilon_r = 4$ and $\tan\delta = 0.0002$, which can be realized in a submicron CMOS/BiCMOS structure. The Q of this resonator for the dominant TM_{010} mode at 57-GHz resonant frequency can be obtained from (7.162), (7.163), and (7.180) as $Q \simeq 65$ which is much higher than any lumped-element CMOS/BiCMOS resonators at 57 GHz.

5.4 RESONATOR'S SLOPE PARAMETERS

We have seen in previous sections that the characteristics of series and parallel resonators – whether lumped elements, distributed elements, or a combination of them – can be found in terms of their input impedance or admittance. "Resonator's slope parameters" or, specifically, "reactance" and "susceptance" slope parameters, related to the input impedance and admittance, have also been used as an alternative for characterizing the resonance of resonators. They are defined for series and parallel resonators as [4]

$$X_s = \frac{\omega_r}{2}\frac{dX}{d\omega}|_{\omega=\omega_r} \quad (\Omega) \tag{5.190}$$

and

$$B_s = \frac{\omega_r}{2}\frac{dB}{d\omega}|_{\omega=\omega_r} \quad (\text{mhos}) \tag{5.191}$$

where X and B are the reactance and susceptance of the input impedance and admittance of the series and parallel resonators, respectively. These parameters were also used in deriving the design equations for various microwave band-pass and band-stop filters. The reactance and susceptance slope parameters for series and parallel RLC resonators can be derived from (5.190) and (5.191) as

$$X_s = \omega_r L = \frac{1}{\omega_r C} \tag{5.192}$$

and

$$B_s = \omega_r C = \frac{1}{\omega_r L} \tag{5.193}$$

The quality factors for series and parallel resonators can be obtained from (5.42) and (5.15), making use of (5.192) and (5.194), as

$$Q = \frac{X_s}{R} \tag{5.194}$$

and

$$Q = \frac{B_s}{G} \tag{5.195}$$

where $G = Q/R$ is the parallel resonator's conductance.

5.5 TRANSFORMATION OF RESONATORS

Versatility in RF circuit design is always a desirable feature and, for RF circuits employing resonators, this may require transformation from one type of resonators into another type. The ability of transforming one resonator type into another provides more flexibility in circuit design and may help improve circuit performance and size as well as realize certain circuit functions. For instance, band-pass filters require alternative use of series and parallel resonators, and realizing both these types of resonators using only one kind of resonators (either series or parallel resonator) would facilitate the filter design.

5.5.1 Impedance and Admittance Inverters

Conversion of resonators from series to parallel or vice versa can be accomplished by using "impedance" or "admittance inverters." Impedance and admittance inverters can also be used to transform impedance or admittance and convert inductance into capacitance or vice versa.

5.5.1.1 Ideal Inverters. Recall from the transmission-line theory that the input impedance of a lossless quarter-wavelength transmission line, having characteristic impedance K, terminated with impedance Z_b, is given as

$$Z_a = \frac{K^2}{Z_b} \tag{5.196}$$

which shows that a series or shunt resonator connecting to a quarter-wavelength line at one end is seen as a shunt or series resonator, respectively, at the other end of the transmission line. A quarter-wavelength transmission line can thus act as an impedance or admittance inverter within a narrow frequency range around the quarter-wavelength frequency, at which the impedance or admittance inversion is exact. It is noted that, over a narrow bandwidth around that frequency, the length of the transmission line, or inverter, varies from longer to shorter than the quarter-wavelength. These length deviations are, however, typically absorbed into the resonators adjacent to the inverter.

An ideal impedance inverter with impedance parameter K is defined as a network that functions as a quarter-wavelength transmission line having a constant characteristic impedance K at all frequencies. Similarly, an ideal admittance inverter with admittance parameter J operates as a quarter-wavelength transmission line with a characteristic admittance J at all frequencies. K and J are referred to as the image impedance and admittance of the inverter, respectively. The ideal admittance inverter converts admittance Y_b connected at one end of the inverter into another admittance of

$$Y_a = \frac{J^2}{Y_b} \tag{5.197}$$

The chain matrix of the ideal quarter-wavelength inverter with $K = Z_o$ can be derived as

$$[T] = \begin{bmatrix} 0 & \pm jK \\ \pm\frac{j}{K} & 0 \end{bmatrix} \tag{5.198}$$

which forms the basis for formulating other inverters that can be used in RF circuits. As can be seen, the ideal inverter has an image phase shift or transmission phase of an odd multiple of $\pm 90°$.

An impedance inverter can also act as an admittance inverter or vice versa. The choice of either impedance or admittance is basically for the convenience of design. Also, as can be seen from (5.196) and (5.197), a series or parallel resonator with particular R, L, C can be realized from a parallel or series resonator connecting to an impedance or admittance inverter. Furthermore, an RF circuit consisting of both series and parallel resonators would have identical electrical performance as one using either series or parallel resonators in conjunction with proper impedance or admittance inverters, respectively. Figure 5.39 illustrates these ideal inverters.

We can also see, from Figure 5.39, that an inductor or capacitor connected at one end of an inverter is electrically transformed into a capacitor or an inductor at the other end, respectively. This result may be useful for impedance transformation and matching and can be used in designing low and high pass filters using only one kind of lumped elements – either inductor or capacitor.

5.5.1.2 Practical Inverters. The foregoing analysis establishes that a quarter-wavelength transmission line forms the basic (ideal) inverter, from which other inverters can be based on. Any RF network can be considered to be either an impedance or an admittance inverter if it was designed to electrically behave as a quarter-wavelength transmission line; that is, the image phase or phase of the transmission coefficient is an odd multiple of $\pm 90°$ and the image impedance or admittance is real in the operating frequency band. Using this concept, various RF networks can be configured and designed as inverters.

Figure 5.40 shows five impedance inverters that can be used with series resonators [4].

The inverters in Figure 5.40(a) and (b) contain negative inductance and capacitance, respectively, and are suitable for use with series resonators that can absorb those negative elements to form circuits having only positive components – for instance, use of the inverter in Figure 5.40(a) at one end of a series resonator having

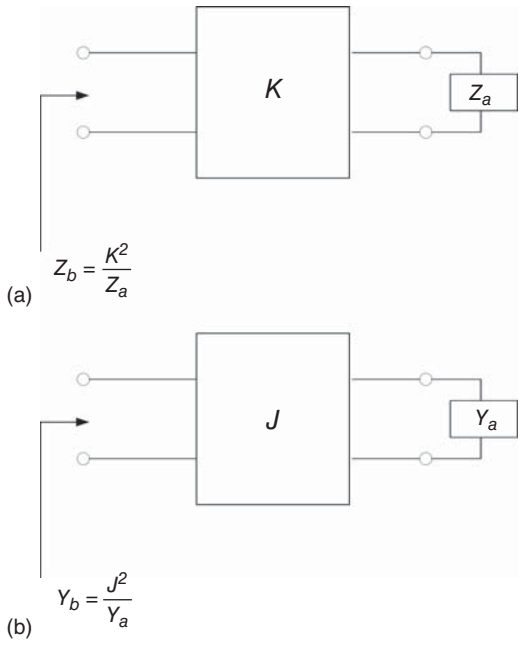


Figure 5.39. Ideal impedance (a) and admittance (b) inverters.

inductance greater than L. The inverters shown in Figure 5.40(c) and (d) have transmission lines of negative and positive electrical lengths, respectively. These inverters are useful for distributed (transmission-line) series resonators, where the negative- or positive-length line can be subtracted from or added to adjacent transmission lines having the same characteristic impedance of the distributed resonators. The inverter shown in Figure 5.40(e) is useful for use with T-networks such as those representing equivalent-circuit models of transmission-line elements, for example, a step discontinuity in transmission lines or a via-hole (neglecting its resistive elements). It should be noted that the transmission phase of these inverters includes the phase of the transmission line of electrical length ϕ and the phase of the shunt and series elements.[4]

The impedance parameter K, image or transmission phase, and other parameters of an inverter can be derived by equating the chain matrix of the inverter to that of the ideal inverter consisting of a quarter-wavelength transmission line of characteristic impedance K given in (5.198). The parameters for the five impedance inverters shown in Figure 5.40 are also included in the same figure.

Figure 5.41 shows five admittance inverters that can be used with parallel resonators [4]. They are the duals of the inverters in Figure 5.40.

The inverters in Figure 5.41(a) and (b) contain negative inductance and capacitance, respectively, and are suitable for use with parallel resonators that can absorb those negative elements to form circuits with only positive components – for instance, using the inverter in Figure 5.41(b) at one end of a parallel resonator having capacitance greater than C. The inverters shown in Figure 5.41(c) and (d) have transmission lines of positive and negative electrical lengths, respectively. These inverters are useful for distributed parallel resonators, where the positive- or negative-length line can be added to or subtracted from adjacent transmission lines with the same characteristic impedance as that of the distributed resonators. The inverter shown in Figure 5.41(e) is useful for use with π-networks such as those representing equivalent-circuit models of transmission-line elements, for example, a gap in transmission lines. The transmission phase of these inverters includes the phase of the transmission line of electrical length ϕ and the phase of the series and shunt elements.

The admittance inverter parameter J, image or transmission phase, and other parameters can be derived by equating the chain matrix of the inverter to that of the ideal inverter consisting of a quarter-wavelength

[4] Any electronic elements including lumped elements possess an electrical length, which causes phase delay or advance in circuits.

transmission line of characteristic admittance $J = 1/K$ given in (5.198). The parameters for the five admittance inverters shown in Figure 5.41 are included in the same figure.

The dependence of the image impedance or admittance and transmission phase of inverters on frequency is important for design involving inverters. These parameters need to be evaluated carefully over interested frequency ranges for selecting proper inverters. There are two main design requirements for inverters. One is realizablility and another is small variation of the image impedance or admittance and transmission phase over the operating frequency range, which may be difficult over a wide bandwidth.

It is noted that the inverters shown in Figures 5.40 and 5.41 have a broader bandwidth than the quarter-wavelength inverter. Inverters 3, 4, and 5, as shown in Figures 5.40(c)–(e) and 5.41(c)–(e), are also

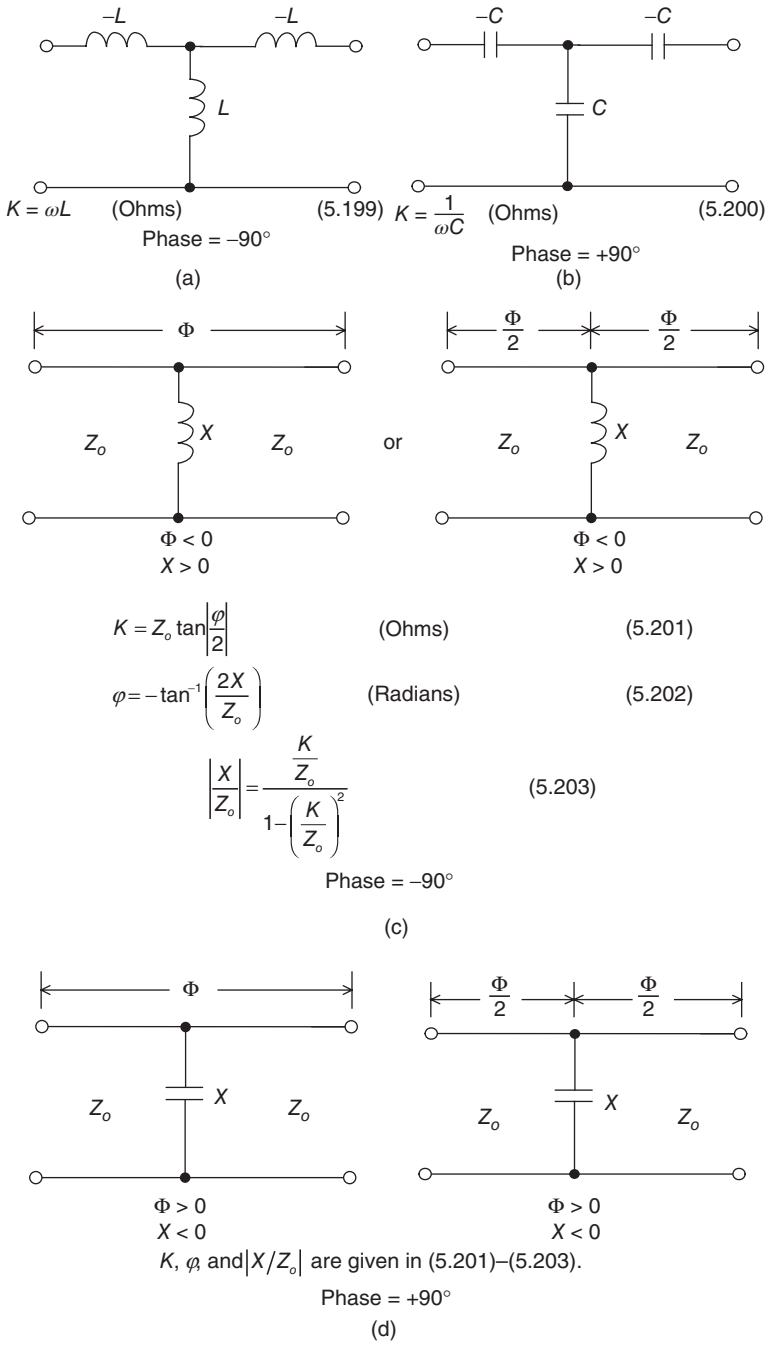
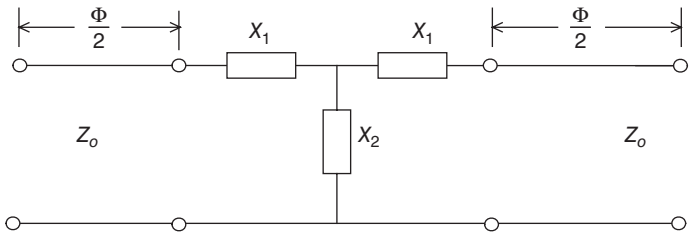


Figure 5.40. (a–e) Impedance inverters for series resonators.



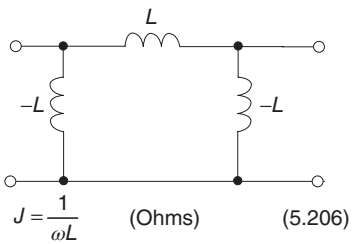
(e)

$$K = Z_o \left| \tan\left(\frac{\varphi}{2} + \tan^{-1} \frac{X_1}{Z_o} \right) \right| \quad \text{(Ohms)} \quad (5.204)$$

$$\varphi = -\tan^{-1}\left(\frac{2X_2}{Z_o} + \frac{X_1}{Z_o} \right) - \tan^{-1} \frac{X_1}{Z_o} \quad \text{(Radians)} \quad (5.205)$$

Figure 5.40. (*Continued*)

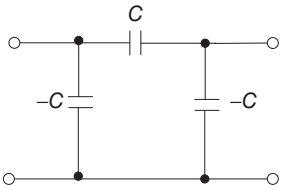
useful for evaluating characteristics of possible inverters produced by intentionally or unintentionally made discontinuities in transmission lines, such as gaps or steps in transmission lines. These inverters, if properly and accurately evaluated, can be exploited to realize or improve the performance of certain circuits. It is interestingly noted that, although typical transmission-line discontinuities are unwanted disruptions in circuits which degrade circuit performance, properly configured and designed discontinuities may help improve circuit performance or even realize certain circuit functions. This view adds another versatile dimension into circuit design and could be useful in facilitating the design.



$$J = \frac{1}{\omega L} \quad \text{(Ohms)} \quad (5.206)$$

Phase = +90°

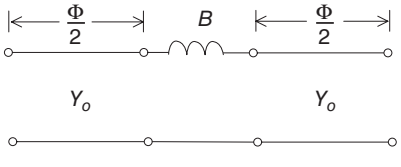
(a)



$$J = \omega C \quad \text{(Ohms)} \quad (5.207)$$

Phase = –90°

(b)



Φ > 0, B < 0

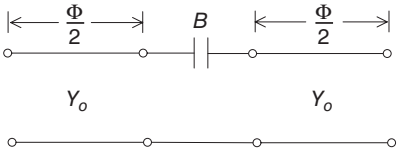
$$J = Y_o \tan \left| \frac{\phi}{2} \right| \quad \text{(Ohms)} \quad (5.208)$$

$$\phi = -\tan^{-1}\left(\frac{2B}{Y_o} \right) \quad \text{(Radians)} \quad (5.209)$$

$$\left| \frac{B}{Y_o} \right| = \frac{\frac{J}{Y_o}}{1 - \left(\frac{J}{Y_o} \right)^2} \quad (5.210)$$

Phase = –90°

(c)



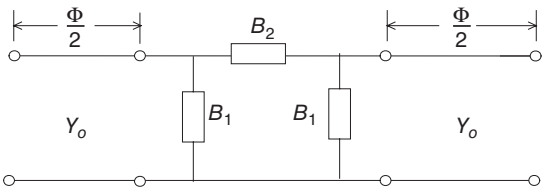
Φ < 0, B > 0

K, ϕ, and $|XZ_o|$ are given in (5.208)-(5.210).

Phase = +90°

(d)

Figure 5.41. (a–e) Admittance inverters for parallel resonators.



$$J = Y_o \left| \tan\left(\frac{\phi}{2} + \tan^{-1} \frac{B_1}{Y_o} \right) \right| \qquad \text{(Ohms)} \qquad (5.211)$$

$$\phi = -\tan^{-1}\left(\frac{2B_2}{Y_o} + \frac{B_1}{Y_o} \right) - \tan^{-1} \frac{B_1}{Y_o} \qquad \text{(Radians)} \qquad (5.212)$$

(e)

Figure 5.41. (*Continued*)

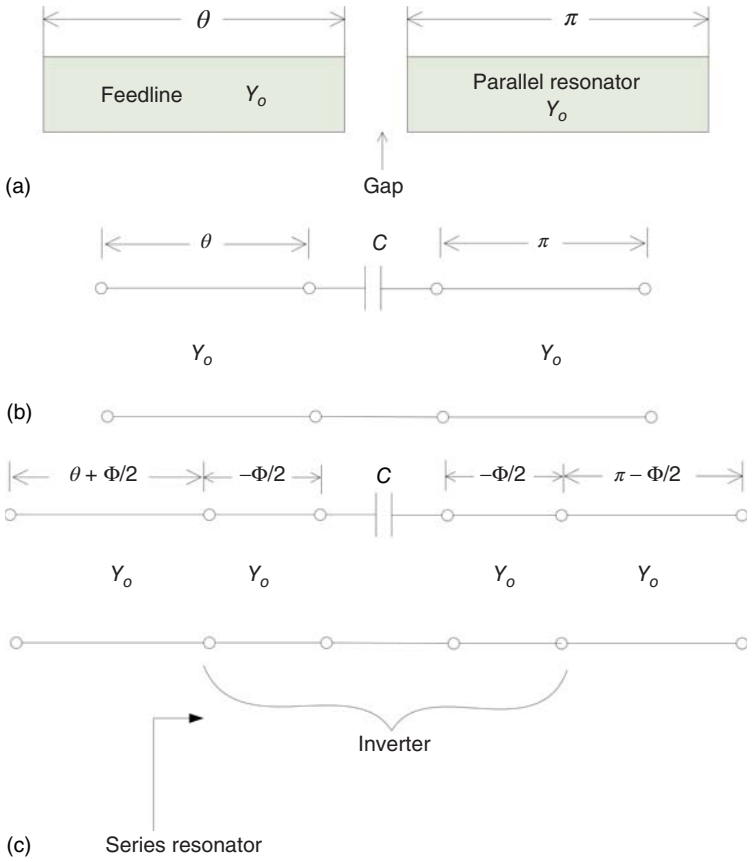


Figure 5.42. (a–c) An open-circuited half-wavelength resonator coupled to external circuits using a gap. The gap is approximately represented by a series capacitance.

5.5.2 Examples of Resonator Transformation

We now recognize that a parallel resonator can be transformed into a series resonator and vice versa using impedance or admittance inverters. This concept is very useful in RF circuit design. The following examples serve to demonstrate this concept.

It was proved earlier that an open-circuited half-wavelength resonator behaves as a parallel resonator. Now consider an open-circuited half-wavelength resonator connected to a feed-line (or external elements) via a gap, as shown in Figure 5.42(a), which is typically used for coupling between external circuits and

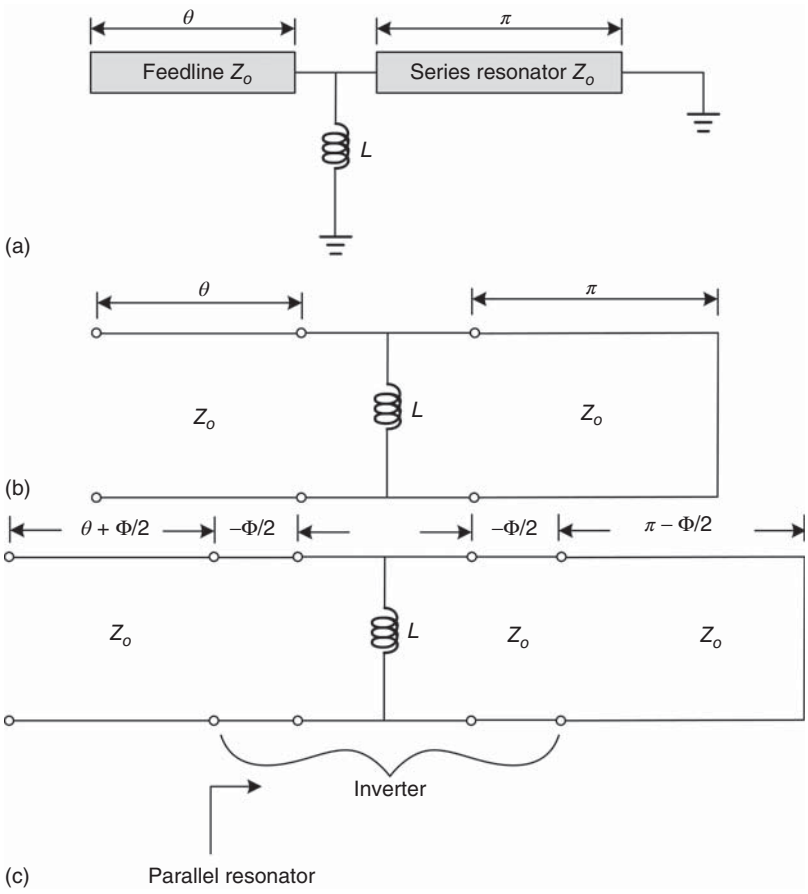


Figure 5.43. A short-circuited half-wavelength resonator connected with a shunt inductance.

transmission-line resonators. Assume the gap is sufficiently narrow such that the gap's fringing capacitance is negligibly small as compared to the gap's coupling capacitance. Figure 5.42(b) shows the equivalent circuit with C representing the series coupling capacitance. The transmission line at each end of the gap can be redrawn to include a negative length of $-\phi/2$, where ϕ is positive, as shown in Figure 5.42(c). The capacitance and two adjacent negative-length transmission lines function as an admittance inverter, referring to Figure 5.41(d), to convert the open-circuited half-wavelength parallel resonator into a series resonator.

As another example to demonstrate conversion of another resonator type, we consider a short-circuited half-wavelength resonator connected with a shunt inductance and the equivalent circuits as shown in Figure 5.43. The short-circuited half-wavelength resonator itself behaves as a series resonator. However, the impedance inverter formed by the shunt inductance and two adjacent negative-length lines convert the series resonator into a parallel resonator.

REFERENCES

1. R. E. Collin, *Foundations for Microwave Engineering*, McGraw-Hill, Inc., New York, New York, 1966, pp. 313–314.
2. R. K. Hoffmann, *Handbook of Microwave Integrated Circuits*, Artech House, Inc., pp. 321–322, Norwood, Massachusetts, 1987.
3. M. Miao and C. Nguyen, "A Novel Multilayer Aperture-Coupled Cavity Resonator for Millimeter-Wave CMOS RFICs," *IEEE Trans. on Microwave Theory and Technique*, Vol. MTT-55, No. 5, pp. 783–787, May 2007.
4. G. Matthaei, L. Young, and E. M. T. Jones, *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, pp. 144–145, 214–217, 430–438, Artech House, Inc., Norwood, Massachusetts, 1980.

PROBLEMS

5.1 Derive Eq. 5.6.

5.2 Prove that the resonator shown in Figure 5.2(a) can be modeled as a parallel resonator whose elements are given as

$$R' = \frac{L}{RC}$$
$$L' = \frac{R}{\omega_r^2 LC}$$
$$C' = \frac{LC}{R}$$

where $\omega_r = \sqrt{\frac{1}{LC} - \left(\frac{R}{L}\right)^2}$ is the resonant frequency.

5.3 Prove that the resonator shown in Figure 5.2(b) behave as a parallel resonator R', L', and C' over a narrow frequency band around the resonant frequency. Assume the transmission line is lossless. Derive the equivalent R', L', and C'.

5.4 Consider a parallel RLC resonator with resonant frequency ω_r and its practical equivalent-circuit representation as shown in Figure P5.1, in which L and C are replaced with their corresponding equivalent circuits. Prove that these two circuits are equivalent over a narrow bandwidth around ω_r and derive conditions for this equivalence to occur [e.g., elements of Figure P5.1(a) in terms of those in Figure P5.1(b)].

5.5 Consider a parallel RLC resonator, whose element values at resonance are R_o, L_o, and C_o. Derive

$$\frac{R_o}{Q} = -\frac{2}{\omega_r^2}\frac{d\omega}{dC}$$

where Q is the unloaded quality factor at the resonant frequency ω_r. This equation indicates that the ratio between R_o and Q can be determined by measuring or calculating the rate of change of frequency with small changes in capacitance, from which R_o can also determined from Q. The derivative of frequency with respect to capacitance, $d\omega/dC$, can be obtained as the slope of the capacitance-frequency curve at the point C_o.

5.6 Derive Eqs. (5.115)–(5.119) for half-wavelength open-circuited resonators.

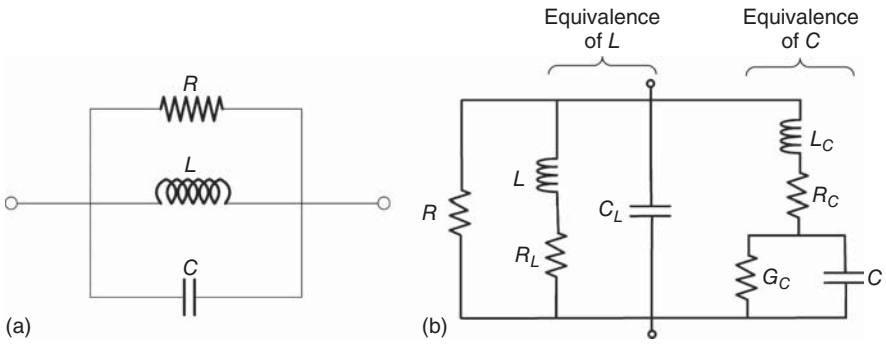


Figure P5.1.

5.7 Let Δt be the time duration during which the voltage amplitude or, in general, the time-domain response of the resonator's voltage, reduces to $1/e$ or 36.8% of its initial value. Determine the number of Δt, after which the energy stored within resonators decays to 1% of its initial value.

5.8 Derive the resonant frequency of series and parallel (RLC) resonators in terms of the (unloaded) Q and the resonant frequency of corresponding lossless resonators. Comment on the resultant resonant frequency.

5.9 Derive Eqs. (5.113)–(5.124) for quarter-wavelength open-circuited resonators.

5.10 Consider a low loss short-circuited $\lambda/8$ transmission line. Determine whether or not it can behave as a parallel or series resonant circuit within a narrow frequency range around the frequency at which the transmission-line length is $\lambda/8$. If it cannot function as a resonator, derive possible condition, such as loading condition at one end, to make it a resonator. Derive the equivalent R, L, and C of the corresponding resonant circuit if any.

5.11 Repeat Problem 5.10 for a low loss open-circuited $\lambda/8$ transmission line.

5.12 Consider a lossy open-circuited half-wavelength resonator with characteristic impedance Z_o connected with a generator (external circuit) having an internal impedance equal to Z_o. Derive the following relation at the resonance:

$$\frac{Q}{2} \ll Q_L < Q$$

where Q and Q_L are the unloaded and loaded quality factors of the resonator, respectively.

5.13 Derive Eq. (5.30).

5.14 Determine whether or not the following transmission-line network is a resonator. Provide your rationale.

a) A lossless transmission line terminated with a real impedance equal to the transmission line's characteristic impedance.

b) A half-wavelength lossless transmission line terminated with a real impedance different from the characteristic impedance.

5.15 Consider a low loss short-circuited transmission line having characteristic impedance Z_o terminated with a load resistor $R_L \neq Z_o$ at the remaining end.

a) Prove that this transmission-line circuit behaves as a resonator and derive the length of the transmission line for the resonance to occur. Note that there may be two resonance conditions corresponding to two different sets of length. Does the (external) load affect the resonant frequency and/or the Q of the short-circuited transmission line?

b) For a given length, what happens to the resonant frequency f_r when the short is replaced with an inductor or a capacitor. Provide your rationale. If f_r changes, then is it reduced or increased when an inductor or capacitor is used in place of the short? Is it possible to use Smith chart to prove that the resonant frequency is increased or reduced?

5.16 Repeat Problem 5.15 with the short replaced with an open.

5.17 Consider a short-circuited quarter-wavelength resonator at 10 GHz. The transmission line is microstrip with $Z_o = 50\,\Omega$, SiO_2 dielectric having $\varepsilon_r = 3.9$, $h = 10\,\mu$m and $\tan\delta = 0.000015$, and copper having $\sigma = 5.8 \times 10^7$ mho/m and thickness of 1 μm. The microstrip's bottom conducting ground plane is assumed to be infinitely large.

a) Plot the magnitude of the input impedance $|Z_{\text{in}}| = |Z_o \tanh(\alpha + j\beta)\ell|$ as a function of frequency and determine the unloaded Q from the resonant frequency and 3-dB bandwidth using Eq. (5.19).

b) Calculate Q from $Q = \pi/4\alpha\ell = \beta/2\alpha$ and compare result to that obtained in Part (a).

5.18 Repeat Problem 5.17 with Si as the dielectric having $\varepsilon_r = 11.7$, $h = 250$ μm and $\tan\delta = 0.005$.

5.19 Consider microstrip resonators. The dielectric is SiO_2 having $\varepsilon_r = 3.9$, $h = 10\,\mu m$ and $\tan\delta = 0.000015$, the conductor is copper with $\sigma = 5.8 \times 10^7$ mho/m and thickness of $1\,\mu m$, and $Z_o = 50\,\Omega$. The microstrip's bottom conducting ground plane is assumed to be infinitely large.

a) Calculate the radiation quality factor Q_r and the (total) unloaded Q for half-wavelength and quarter-wavelength open- and short-circuited microstrip resonators at 35 GHz. We assume the open-circuited half- and quarter-wavelength resonators have the same Q_r and the Q_r of the short-circuited half- and quarter-wavelength resonators is also equal.

b) Assume the half-wavelength open- and quarter-wavelength short-circuited resonators are loaded with a load resistor of $10\,k\Omega$, determine the external quality factor Q_e and loaded quality factor Q_L.

c) Assume the half-wavelength short- and quarter-wavelength open-circuited resonators are loaded with a load resistor of $1\,\Omega$, determine the external quality factor Q_e and loaded quality factor Q_L. Compare and discuss the results to those obtained in Part (b).

5.20 Repeat Problem 5.19 with SiO_2 replaced by Si having $\varepsilon_r = 11.7$, $h = 250\ \mu m$ and $\tan\delta = 0.005$.

5.21 A good approximation often used for narrow bandwidth $\Delta\omega$ around a center or resonant frequency ω_o of RF circuits is

$$\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \simeq \frac{2\Delta\omega}{\omega_o}$$

where $\Delta\omega = \omega - \omega_o$ with $\Delta\omega \ll \omega_o$. Derive this equation.

5.22 A resonator is constructed from a 50-Ω lossless microstrip line open at one end and terminated with an inductor at the other end, as shown in Figure P5.2. The length of the microstrip line is $\lambda/5$ at 10 GHz, where λ is the corresponding wavelength of the microstrip line. The microstrip line's substrate has $\varepsilon_r = 2.2$ and thickness $h = 0.031$ in.

a) Determine L so that the lowest-order resonance occurs at 10 GHz.

b) Place a 5000-Ω resistor in parallel with the inductor and calculate the corresponding quality factor.

5.23 Repeat Problem 5.10 for a low loss open-circuited $\lambda/8$ transmission line.

5.24 A series RLC resonant circuit is connected to a lossless transmission line that is a quarter-wavelength long at frequency f_o.

a) Show that the combined network of the series RLC and transmission line behaves as a parallel resonator at f_o. Derive the elements of the equivalent parallel resonator in terms of the elements of the series RLC and the transmission line.

b) Assume the transmission line's characteristic impedance is $Z_o = 50\,\Omega$, $R = 2\,\Omega$, $L = 40$ nH, and $C = 30$ pF. Determine the 3-dB bandwidth of the new resonator that consists of the transmission line and the series RLC.

5.25 Consider a resonator, consisting of a varactor, coupling network 1 and microstrip line, connected to an oscillator via coupling network 2, as shown in Figure P5.3.

a) Derive the following equation:

$$Q_{oV} = \frac{Q_T(1 + K_1)(1 + K_2)Q_{oM}}{Q_{oM} - Q_T(1 + K_2)}$$

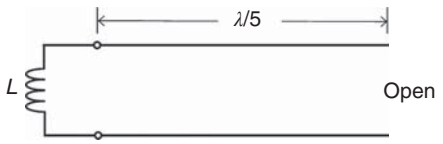


Figure P5.2.

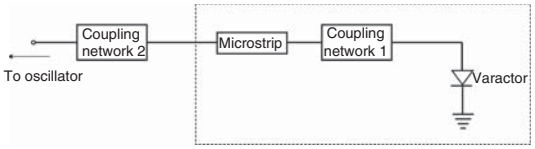


Figure P5.3.

where Q_{oV} is the unloaded quality factor of the varactor, Q_T is the quality factor of the combined circuit of resonator and coupling network 2, Q_M is the unloaded quality factor of the microstrip line, and K_1 and K_2 are the coupling coefficients of the coupling networks 1 and 2, respectively.

b) Determine the required value for Q_{oV} at 35 GHz, assuming $Q_T = 3$, $Q_M = 60$, $K_1 = 0$ dB (direct coupling), and $K_2 = -10$ dB.

5.26 Derive Eq. (5.159) for the fields of the TE modes for rectangular cavity resonators.

5.27 Derive the expression for Q of TE_{10p} mode in a rectangular cavity resonator of dimensions a, b, and d. It is assumed that the resonator's conducting walls are lossy and its dielectric is perfect with a relative dielectric constant of ε_r.

5.28 Derive Eq. (5.185) for the fields of the TM modes for circular cavity resonators.

5.29 Derive Eq. (5.187) for the resonant frequency of the TM modes for circular cavity resonator.

5.30 Derive Eqs. (5.188) and (5.189) for the conductor quality factors of the TM_{mnp} and TM_{010} modes, respectively.

5.31 Find the resonant frequency and Q of a rectangular cavity resonator having $a = b = d = 2$ cm for the TE_{101} mode. The resonator is made of copper with air as the dielectric.

5.32 Find the resonant frequency and Q for the TE_{111} mode of a circular cavity resonator, made of copper, having radius and length of 1 cm. The dielectric of the resonator has $\varepsilon_r = 2.1$ and $\tan\delta = 0.0004$.

5.33 Determine the dimensions of an air-filled circular cavity resonator operating in the TE_{011} mode so that it has a maximum Q at 10 GHz.

5.34 Derive the expression for Q of TE_{011} mode in a rectangular cavity resonator of dimensions a, b, and d. The resonator's conducting walls are assumed to be perfect and its dielectric has $\hat{\varepsilon} = \varepsilon' - j\varepsilon''$.

5.35 A rectangular cavity resonator is designed for a millimeter-wave electron device for operation in the TE_{101} mode at $f = 300$ GHz. Assume the conductor is copper with $\sigma = 5.8 \times 10^7$ mho/m, the enclosed dielectric is a low loss material with $\varepsilon_r = 3.9$, and the resonator's length and width are equal ($a = d$). Moreover, the height of the resonator must be kept small ($b < 0.1$ mm) to minimize the electron transit time. Since the height is small, use of CMOS technology is possible.

a) Calculate the dimensions of the resonator.

b) Calculate the Q and 3-dB bandwidth of the resonator.

5.36 Consider a CMOS structure as shown in Figure P5.4. Assume the width of the metal in each layer can be any value. The distance between metal layers, however, is limited as shown in the figure. Design a circular cavity resonator operating in the TM_{010} mode that can be realized in that CMOS process to have highest Q at 50 GHz and 100 GHz.

5.37 Repeat Problem 5.36 for 400 GHz. Based on the results of Problem 5.36 and this problem, what do you suggest of the operating frequency ranges of circular cavity resonators if we want to realize them using CMOS processes.

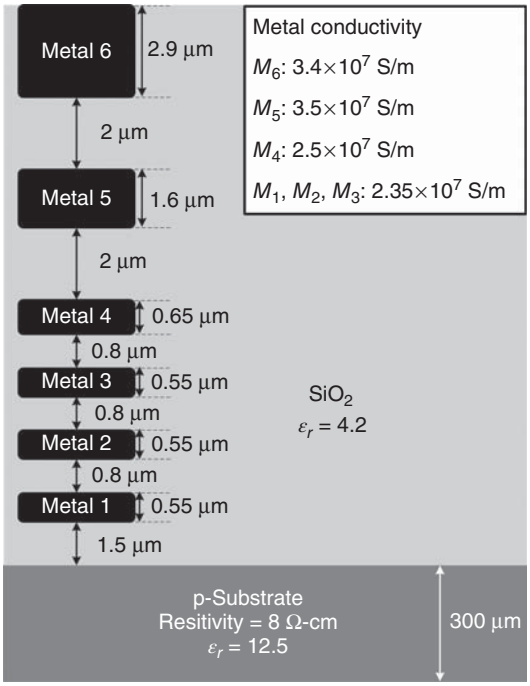


Figure P5.4.

5.38 Consider the CMOS profile as shown in Figure P5.4 and a circular cavity resonator having $d = 6.4$ μm and $d/a = 3/2$ in that structure. Assume the loss tangent of the SiO_2 dielectric is $\tan\delta = 0.0002$. Determine the resonant frequency for the TM_{010} mode and calculate the corresponding quality factor. You can assume all the conductors have the same conductivity as metal 4. Can this resonator be realized in a CMOS structure? If no, give a rationale. If yes, give a design example including dimensions for d and a, along with metal layers to be used, as well as dimensions of and gaps between the via-holes used for the wall for a 0.18-μm CMOS or any available submicron CMOS process.

5.39 Repeat Problem 5.38 for $d/a = 5 \times 10^{-3}$. Compare and comment on the results between these two problems. If both resonators can be realized, which one is better for CMOS structures with respect to physical realization and electrical performance (operating frequency and Q)? What about when d/a approaches smaller values than 5×10^{-3}?

5.40 Repeat Problem 5.38 for $d/a = 2$ for the TE_{111} mode. Is this kind of resonators applicable for the normal RF range up to 300 GHz or only possible in the terahertz range?

5.41 Derive Eqs. (5.201)–(5.203) for the inverters in Figure 5.40(c) and (d).

5.42 Derive Eqs. (5.204)–(5.205) for the inverter in Figure 5.40(e).

5.43 Derive Eqs. (5.208)–(5.210) for the inverters in Figure 5.41(c) and (d).

5.44 Derive Eqs. (5.211)–(5.212) for the inverter in Figure 5.41(e).

5.45 Consider a network consisting of two parallel complex admittances Y_1 and Y_2. Prove that resonance occurs when $Y_1 = Y_2^*$ where * indicates the conjugate.

5.46 Prove that the input impedance of a low loss short-circuited $\lambda/8$ transmission line in a narrow frequency range around f, where f is the frequency at which the transmission-line length is $\lambda/8$, is given as

$$Z_{\text{in}} \simeq Z_o \frac{\alpha\ell + j}{1 + j\alpha\ell} \simeq jZ_o$$

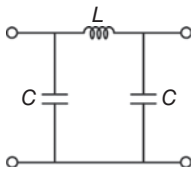


Figure P5.5.

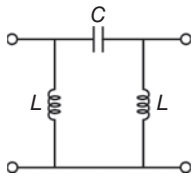


Figure P5.6.

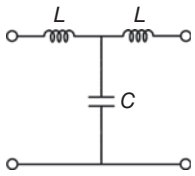


Figure P5.7.

where Z_o, α, and ℓ are the characteristic impedance, attenuation constant and length of the transmission line.

5.47 For Problem 5.10, can another low loss transmission line be added to the short-circuited $\lambda/8$ transmission line to make it a resonator in the vicinity of the frequency at which the transmission-line length is $\lambda/8$? Determine the length of this transmission line and R, L, and C of the resultant resonator in terms of the transmission-line parameters.

5.48 Consider an admittance inverter as shown in Figure P5.5. Derive the image admittance J and phase of the inverter.

5.49 Consider an admittance inverter as shown in Figure P5.6. Derive the image admittance J and phase of the inverter.

5.50 Consider an impedance inverter as shown in Figure P5.7. Derive the image impedance K and phase of the inverter.

CHAPTER 6

IMPEDANCE MATCHING

Impedance matching involves transformation of an impedance to another impedance to achieve certain desired results such as an optimum noise figure or minimum reflection. Impedance matching is inevitable in radio frequency integrated circuits (RFICs). Virtually all RFICs require impedance matching; therefore, the design of impedance-matching networks is important in RFIC design. This chapter is focused on some fundamental design techniques for impedance-matching networks.

6.1 BASIC IMPEDANCE MATCHING

Impedance matching does not necessarily involve only minimization of reflection or voltage standing wave ratio (VSWR) in RFICs. As an example, we consider a general radio frequency (RF) amplifier as shown in Figure 6.1. For low noise design, as discussed in Section 11.2 of Chapter 11, a certain impedance Z_S needs to be presented at the input terminal of the transistor in order to achieve a particular noise figure. The input impedance-matching network is therefore designed to transfer the source impedance Z_o to Z_S. The output impedance-matching network, on the other hand, is typically designed to transfer the load impedance Z_o to Z_L, corresponding to $\Gamma_L = \Gamma_{\text{out}}^*$, for maximum gain. At the same time, these impedance-matching networks also minimize the reflection at the input and output ports of the amplifier. With respect to minimization of reflection, which is the typical objective of impedance-matching network design, ideally it is also desired to have out-of-band rejection, thus imposing a filter function upon matching networks as well.

One of the most important tools in the design of impedance-matching networks is the Smith chart invented by Smith in the 1930s [1]. The characteristics of the Smith chart are given as follows.

6.1.1 Smith Chart

We consider a lossless transmission line terminated with a load as shown in Figure 6.2 and define the normalized load impedance as

$$z \equiv \frac{Z_L}{Z_o} = r + jx \tag{6.1}$$

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

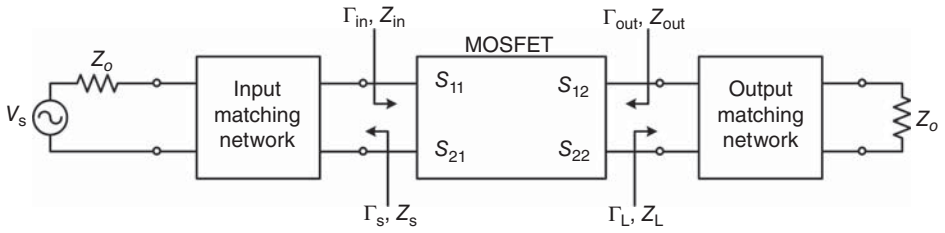


Figure 6.1. General block diagram of RF amplifiers.

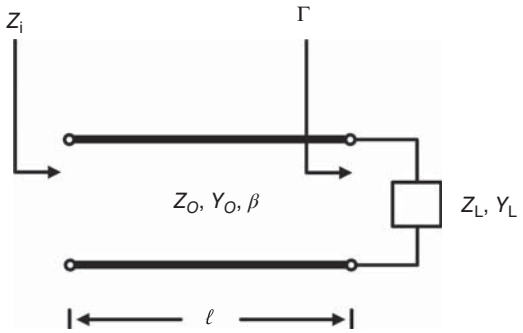


Figure 6.2. Transmission line with characteristic impedance Z_o (or characteristic admittance Y_o) and phase constant β terminated with a load of impedance Z_L (or admittance Y_L).

where r and x represent the normalized resistance and reactance. The reflection coefficient Γ at the load is given as

$$\Gamma = \Gamma_r + j\Gamma_i = \frac{Z_L - Z_o}{Z_L + Z_o} = \frac{z-1}{z+1} \tag{6.2}$$

where Γ_r and Γ_i are the real and imaginary parts of Γ. We can obtain from (6.2):

$$z = \frac{1+\Gamma}{1-\Gamma} \tag{6.3}$$

Substituting $z = r + jx$ from (6.1) and $\Gamma = \Gamma_r + j\Gamma_i$ from (6.2) into (6.3), and equating the corresponding real and imaginary parts leads to

$$r = \frac{1-\Gamma_r^2-\Gamma_i^2}{(1-\Gamma_r)^2+\Gamma_i^2} \tag{6.4}$$

$$x = \frac{2\Gamma_i}{(1-\Gamma_r)^2+\Gamma_i^2} \tag{6.5}$$

Equations (6.4) and (6.5) can be rearranged as

$$\left(\Gamma_r - \frac{r}{1+r}\right)^2 + \Gamma_i^2 = \left(\frac{1}{1+r}\right)^2 \tag{6.6}$$

$$(\Gamma_r - 1)^2 + \left(\Gamma_i - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2 \tag{6.7}$$

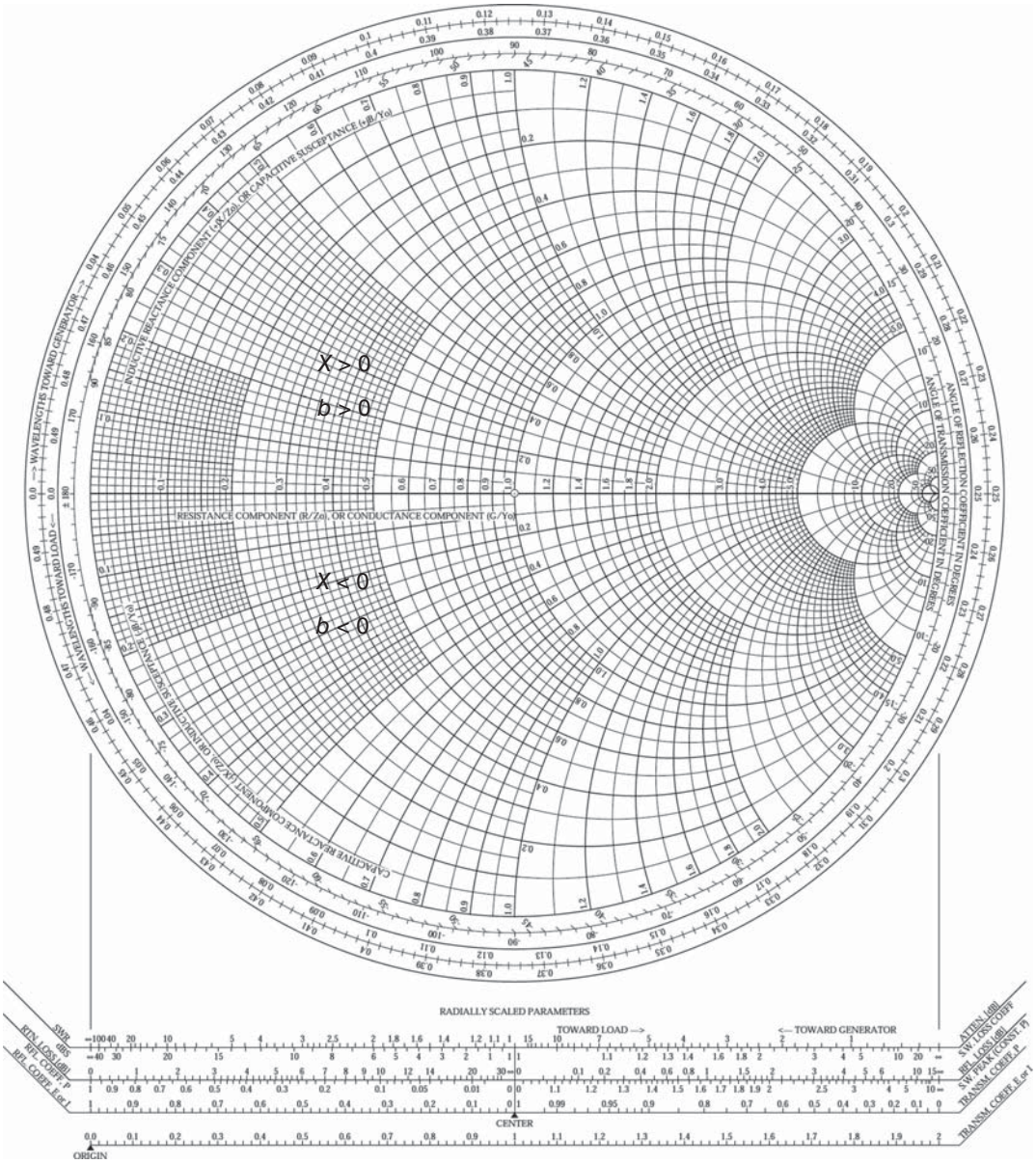


Figure 6.3. Smith chart. Note that both x and b are positive in the upper half and negative in the lower half of the Smith chart when it is used as the Z-chart and Y-chart, respectively.

It is apparent that each of these equations describes a family of circles. Equation (6.6) represents a set of circles having center at $\Gamma_r = r/(1+r)$, $\Gamma_i = 0$ and radius of $1/(1+r)$, with r as an independent variable, in the (Γ_r, Γ_i) plane. These circles are referred to as the constant-resistance (or constant-r) circles. Equation (6.7) represents another set of circles centered at $\Gamma_r = 1$, $\Gamma_i = 1/x$ with radius of $1/x$, with x as an independent variable, in the (Γ_r, Γ_i) plane. These circles are known as the constant-reactance (or constant-x) circles.

Smith chart is formed by plotting the constant-r and constant-x circles on the same (Γ_r, Γ_i) plane as shown in Figure 6.3. The intersection of a constant-r and a constant-x circle gives the impedance $z = r + jx$. This characteristic is a significant fundamental of the Smith chart from which many calculations can be performed and various design techniques can be evolved. The Smith chart can also be used for admittance based on the normalized load admittance

$$y \equiv \frac{Y_L}{Y_o} = \frac{1}{z} = g + jb \tag{6.8}$$

where g and b correspond to the constant-conductance (or constant-g) and constant-susceptance (or constant-b) circles on the Smith chart, respectively. These circles form the admittance version of the Smith chart which is the counterpart of the Smith chart based on the constant-r and -x circles.

The input impedance of a lossless transmission can be derived by making use of (4.187) and (4.190) of Chapter 4 as

$$z_i = \frac{1 + \Gamma e^{-j2\beta\ell}}{1 - \Gamma e^{-j2\beta\ell}} \tag{6.9}$$

As can be seen from (6.3) and (6.9), the normalized load impedance z is indeed equal to the normalized input impedance z_i when $\ell = 0$ as expected. This further implies that the input impedance can be obtained from the load impedance by rotating this impedance an angle of $2\beta\ell$ in the clockwise (CW) direction on the Smith chart. This direction is marked as "toward generator" on the Smith chart. On the other hand, rotating the input impedance in the counter-clockwise (CCW) direction moves toward the load impedance, and correspondingly the direction is written as "toward load" on the Smith chart. These additional features allow transmission-line calculations to be performed using the Smith chart.

Important characteristics of the Smith chart are summarized as follows.

1. Impedance and admittance on the Smith chart are normalized with respect to the characteristic impedance and admittance, respectively.
2. Values of the normalized resistance r or conductance g are read on the central horizontal line.
3. Values of the normalized reactance x or susceptance b are read on the corresponding constant-x or constant-b circles.
4. Reactance x and susceptance b are positive and negative in the upper and lower half of the Smith chart, respectively, while r is always positive inside the Smith chart.
5. Location for the perfect match, corresponding to zero reflection or unity VSWR, is at the center of the Smith chart ($r = 1, x = 0$) or ($g = 1, b = 0$).
6. Reflection coefficient (Γ) and VSWR can be determined directly from the Smith chart.
7. Traveling along a lossless transmission line is equivalent to moving on a circle having center at the Smith chart's center and radius equal to the magnitude of Γ, which is known as the constant-reflection (constant-Γ) or constant-VSWR circle. Along this circle, only the phase of Γ changes.
8. Admittance can be obtained by rotating 180° from the impedance point along a constant-Γ circle.
9. Moving away from the load or the input of a transmission line means moving toward generator or toward load as indicated on the Smith chart, respectively.
10. Traversing a transmission line over an electrical length $\theta = \beta\ell$ is equivalent to moving over 2θ on the Smith chart.
11. Inside the Smith chart, $|\Gamma| < 1$ and on the outer circle, $|\Gamma| = 1$.
12. The left-most and right-most points on the Smith chart correspond to the short (or open) and the open (or short) for impedance (or admittance), respectively.

Impedance-matching networks normally consist of both series and shunt elements and hence their analysis and design are more conveniently handled by a mixture of impedance and admittance. To facilitate the design of impedance-matching networks, both the impedance and admittance forms of the Smith chart should be used simultaneously. This leads to the formation of the Smith impedance–admittance chart, or simply Smith Z–Y chart, as shown in Figure 6.4, where the Smith chart for impedance and Smith chart for admittance, obtained by rotating the impedance chart 180°, are superimposed. The Smith Z–Y chart is typically employed for the design of impedance-matching networks. It is noted that the normalized impedance z and admittance y are read directly on the Z and Y chart, respectively, with $x > 0$, $b < 0$ and $x < 0$, $b > 0$ in the upper and lower half, respectively, and $r > 0$ and $g > 0$ at any point within the Smith Z–Y chart.

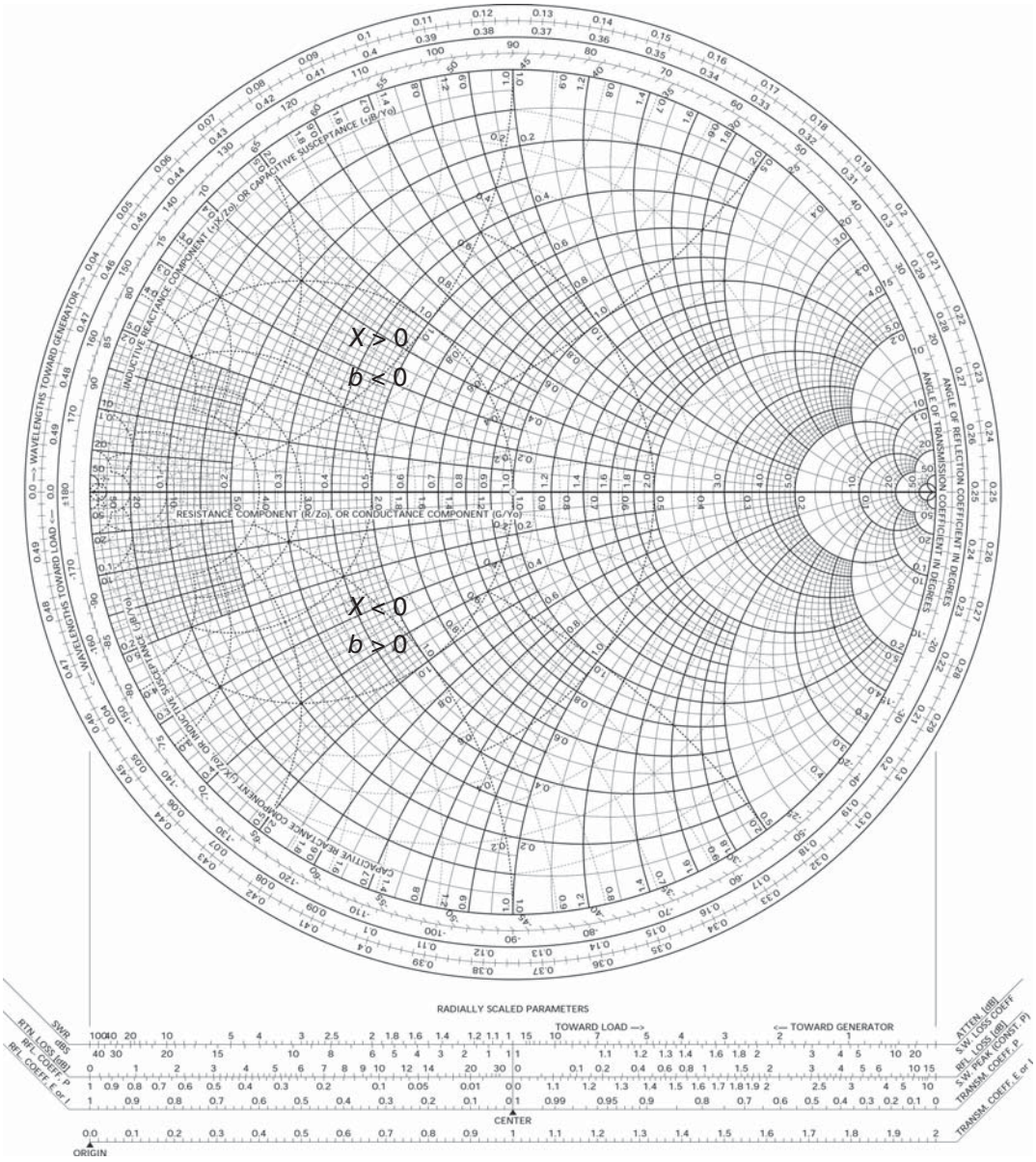


Figure 6.4. Smith Z–Y chart consisting of Z-chart for impedance Z (darker lines) and Y-chart for admittance Y (lighter lines). Note that $x > 0$ ($b < 0$) and $x < 0$ ($b > 0$) in the upper and lower half of the Z-chart (Y-chart), respectively.

6.2 DESIGN OF IMPEDANCE-MATCHING NETWORKS

Impedance-matching networks can consist of only lumped elements (inductors, capacitors, resistors) or transmission lines, or a combination of lumped elements and transmission lines. Here we consider only inductors and capacitors for the lumped elements, although resistors can also be used, especially for some wideband impedance matching. The lumped-element and transmission-line matching networks can be derived independently or from each other. Using lumped elements for impedance-matching networks is typically preferred than transmission lines for RFICs due to significantly smaller size for lumped elements. However, at frequencies in the high end of the RF spectrum, transmission-line impedance-matching networks may need to be used because lumped elements do not behave truly as lumped elements due to electromagnetic (EM) effects and/or the size of transmission lines become comparable to that of lumped elements while having higher quality factor.

6.2.1 Impedance-Matching Network Topologies

Figure 6.5 shows eight possible topologies for impedance-matching networks considering only two lumped elements. These impedance-matching networks represent the simplest practical lumped-element matching networks.

In Section 4.8.5, we show that an inductor and capacitor are equivalent to an open- or short-circuited lossless line. Therefore, the lumped-element impedance-matching networks in Figure 6.5 can be transferred to equivalent impedance-matching networks using open- and/or short-circuited transmission lines. Figure 6.6 shows two of them corresponding to the lumped-element impedance-matching networks in Figure 6.5(a) and (c), where (series/shunt) short- and open-circuited transmission lines are used for (series/shunt) inductor and capacitor, respectively. It is noted that series transmission lines are very difficult, and even impossible for certain transmission-line configurations, to be realized in RFIC. To avoid the series-connection problem, Kuroda identities [2], as described later, can be employed to transform series into shunt transmission lines.

6.2.2 Impedance Transformation through Series and Shunt Inductor and Capacitor

To facilitate the design of lumped-element impedance-matching networks such as those shown in Figure 6.5 and, in turn, transmission-line impedance-matching networks like those in Figure 6.6, using the Smith $X-Y$ chart, we examine the effects of transforming an arbitrary impedance through a series or shunt inductor and capacitor on the Smith $Z-Y$ chart. Without loss of generality, we analyze the effects on a specific impedance (or admittance) when it is transformed through a specific inductance or capacitance as follows.

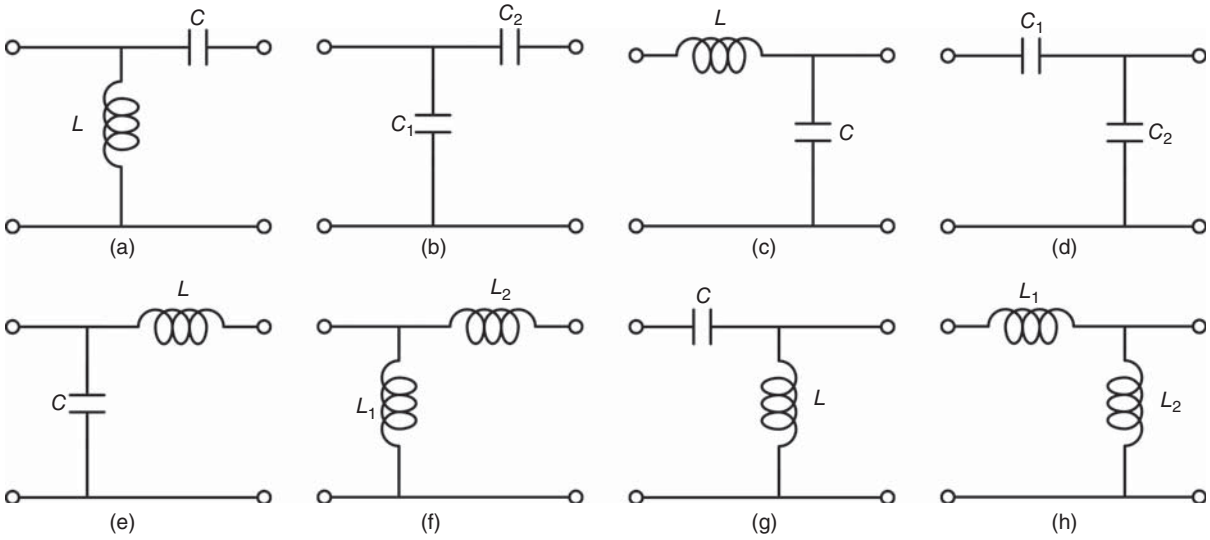


Figure 6.5. (a–h) Possible impedance-matching networks using two lumped elements.

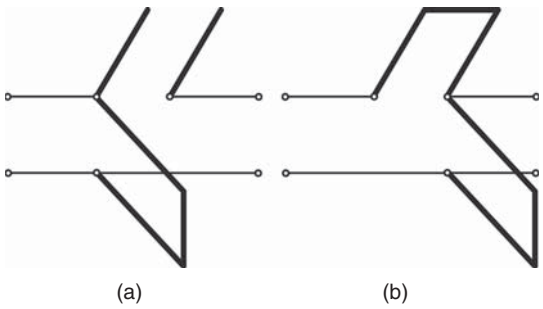


Figure 6.6. Transmission-line impedance-matching networks corresponding to Figure 6.5(a) and (c).

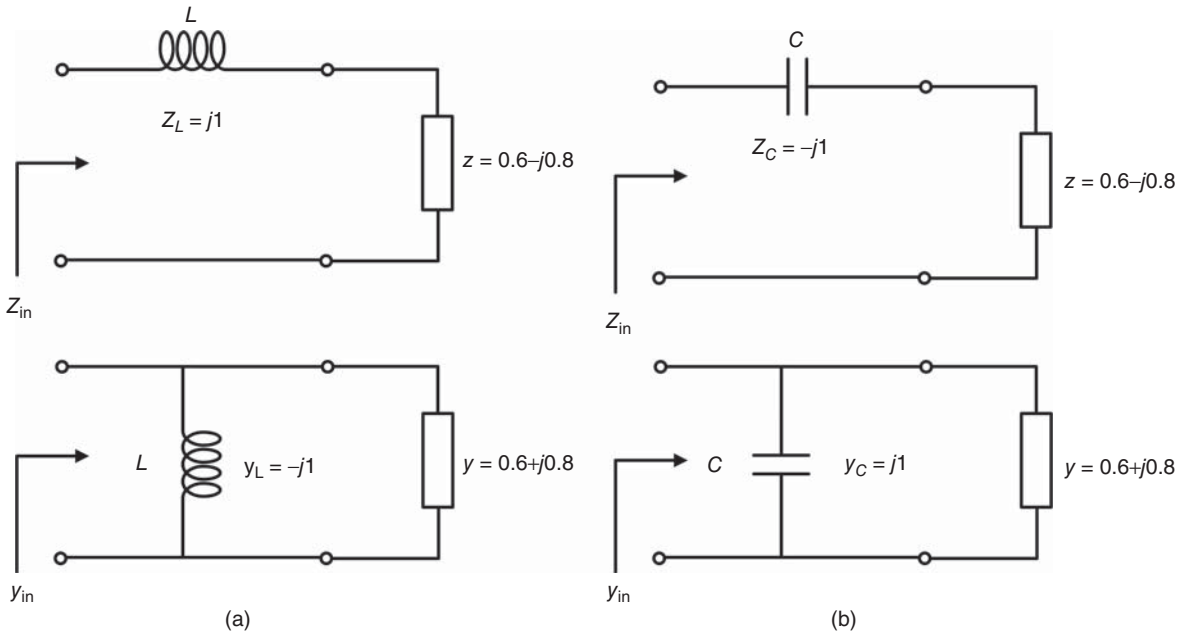


Figure 6.7. (a–d) Impedance in series or shunt with inductor or capacitor. $z = 0.6 - j0.8$ and $y = 1/z = 0.6 + j0.8$ are the normalized impedance and admittance, respectively. $z_L = j1$ and $y_L = 1/z_L = -j1$ are the normalized impedance and admittance of the inductor, respectively. $z_C = -j1$ and $y_C = 1/z_C = j1$ are the normalized impedance and admittance of the capacitor, respectively.

Effect of Series Inductor

We consider an impedance connected in series with an inductor as shown in Figure 6.7(a). The normalized input impedance representing the transformed impedance is $z_{\text{in}} = z_L + z = 0.6 + j0.2$. The impedance z and its transformed impedance z_{in} are located at points A and B on a Smith Z–Y chart, respectively, as shown in Figure 6.8. Note that these points are located according to the Z-chart portion of the Smith Z–Y chart. We can then see that a series inductor transfers an impedance (A) to another impedance (B) on the same constant-r circle in the CW direction. The constant-r circle is referenced to the Z-chart portion.

Effect of Series Capacitor

We now consider Figure 6.7(b) which shows the same impedance z, as shown in Figure 6.7(a), in series with a capacitor. The transformed impedance is $z_{\text{in}} = z_C + z = 0.6 - j1.8$ and is located at point C in Figure 6.8. It is apparent that a series capacitor transforms an impedance (A) to another impedance (C) on the same constant-r circle in the CCW direction.

Effect of Shunt Inductor

Figure 6.7(c) shows an admittance y in parallel with an inductor. The transformed admittance is $y_{\text{in}} = y_L + y = 0.6 - j0.2$ and is located at point D in Figure 6.8. Note that D is located according to the Y-chart portion of the Smith Z–Y chart. The result indicates that a shunt inductor moves an admittance (A) to another admittance (D) located on the same constant-g circle in the CCW direction. The constant-g circle is referenced to the Y-chart portion.

Effect of Shunt Capacitor

Figure 6.7(d) shows the same admittance y in parallel with a capacitor. The transformed admittance is $y_{\text{in}} = y_C + y = 0.6 + j1.8$ and is located at point E in Figure 6.8. We can see that a shunt capacitor causes an admittance (A) to move to another admittance (E) on the same constant-g circle in the CW direction.

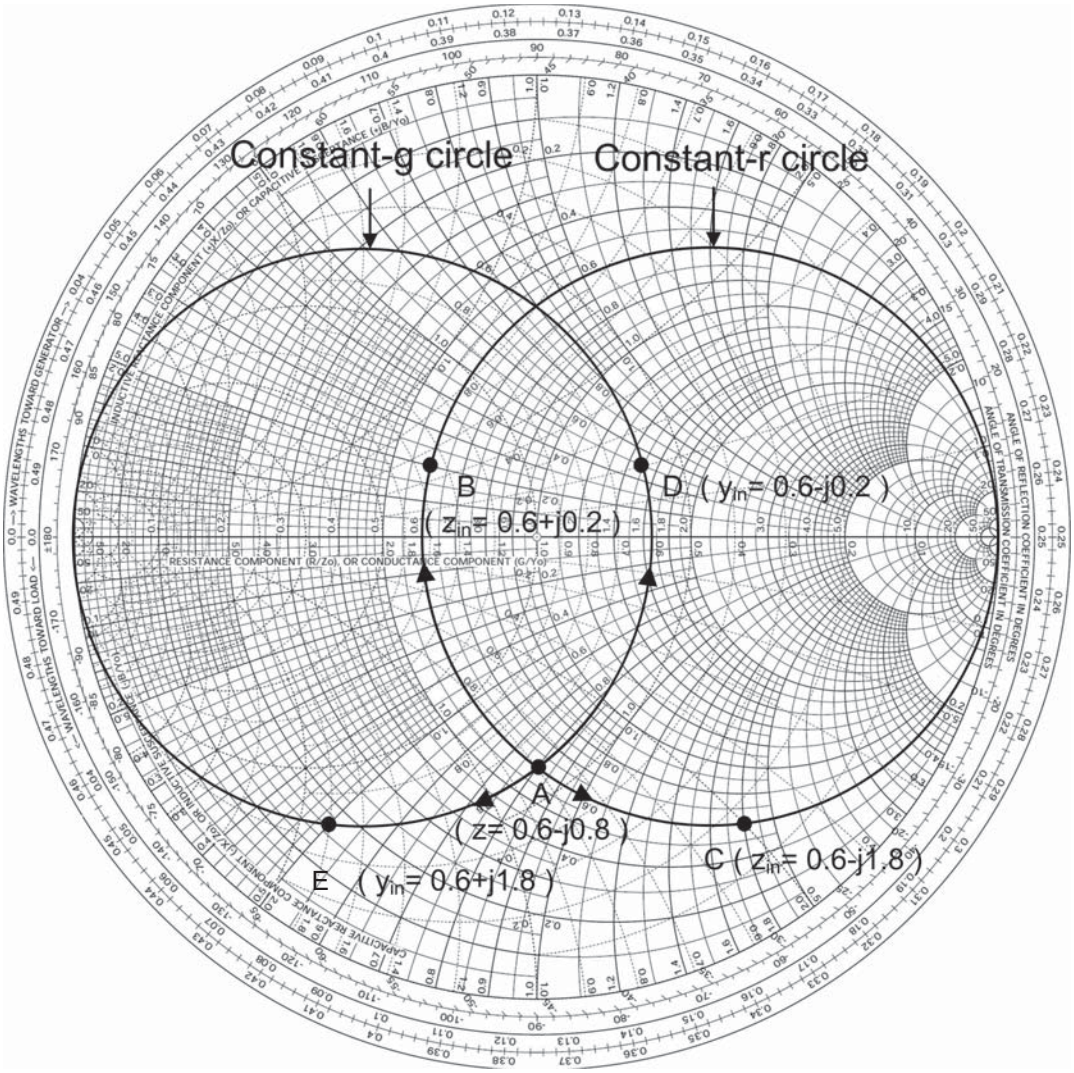


Figure 6.8. Impedance transformation through inductor or capacitor on a Smith $Z-Y$ chart.

TABLE 6.1. Summary of Transformation Effects of Inductors and Capacitors on Impedance or Admittance

Element	Transformation direction	Transformation circle
Series inductor	CW	Constant-r (Z-chart)
Series capacitor	CCW	Constant-r (Z-chart)
Shunt inductor	CCW	Constant-g (Y-chart)
Shunt capacitor	CW	Constant-g (Y-chart)

CW, clockwise; CCW, counter-clockwise.

Table 6.1 and Figure 6.9 summarize the transformation effects of inductors and capacitors on impedance (or admittance). It is noted that Table 6.1 or Figure 6.9 can also be used to identify whether a series or shunt inductor or capacitor is needed when an impedance (or admittance) is moved in a particular direction on a constant-r or -g circle. They can be used to design virtually any lumped-element impedance-matching network.

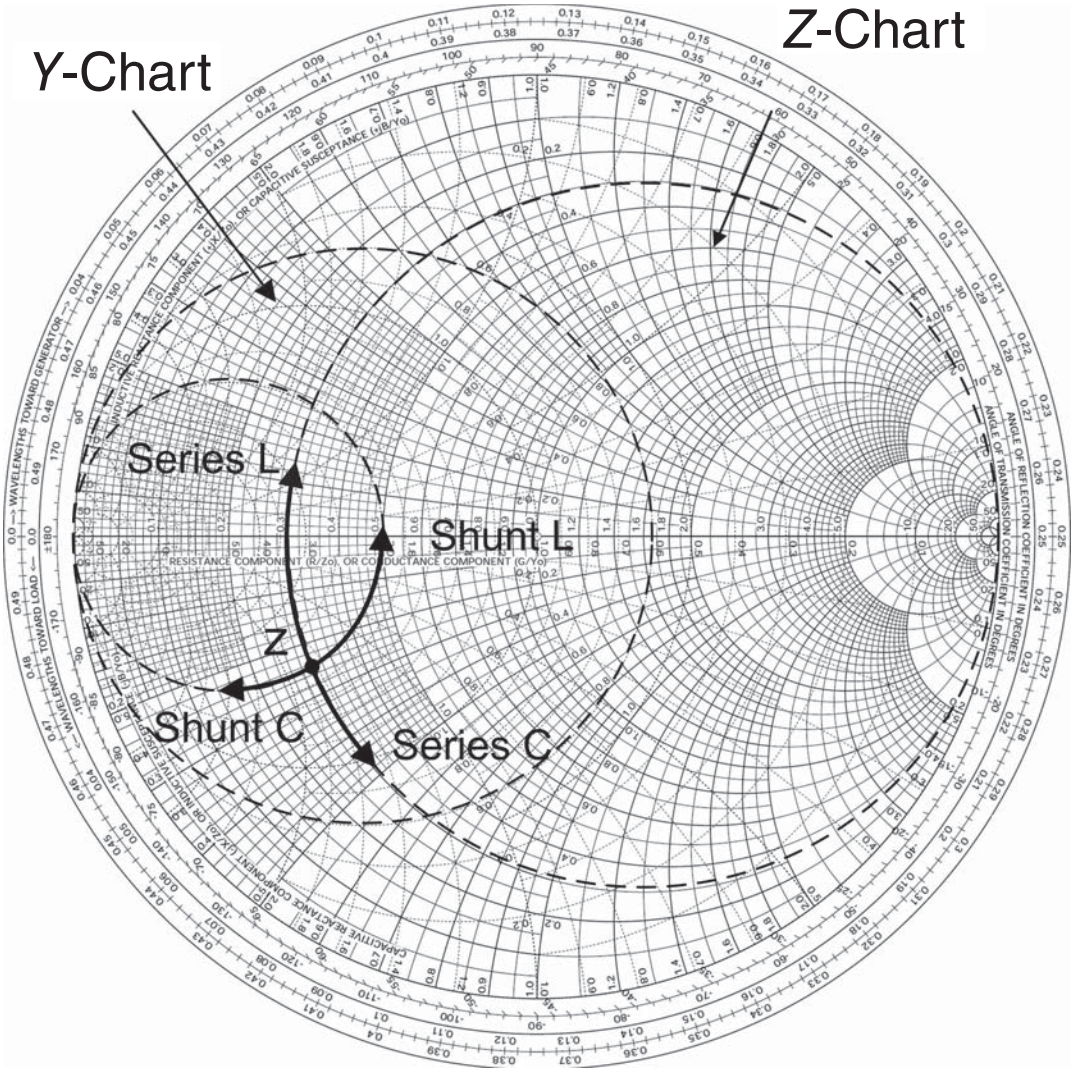


Figure 6.9. Transformation effects of inductors and capacitors on impedance or admittance on the Smith $Z-Y$ chart.

6.2.3 Examples of Impedance-Matching Network Design

As examples to illustrate a procedure for the design of impedance-matching networks, we describe the design of two impedance-matching networks. In the first example, we design a lumped-element impedance-matching network to transfer or match a complex impedance given as $Z_1 = 10 + j10\ \Omega$ to a real impedance of $Z_2 = 50\ \Omega$ at 10 GHz. The design steps are as follows.

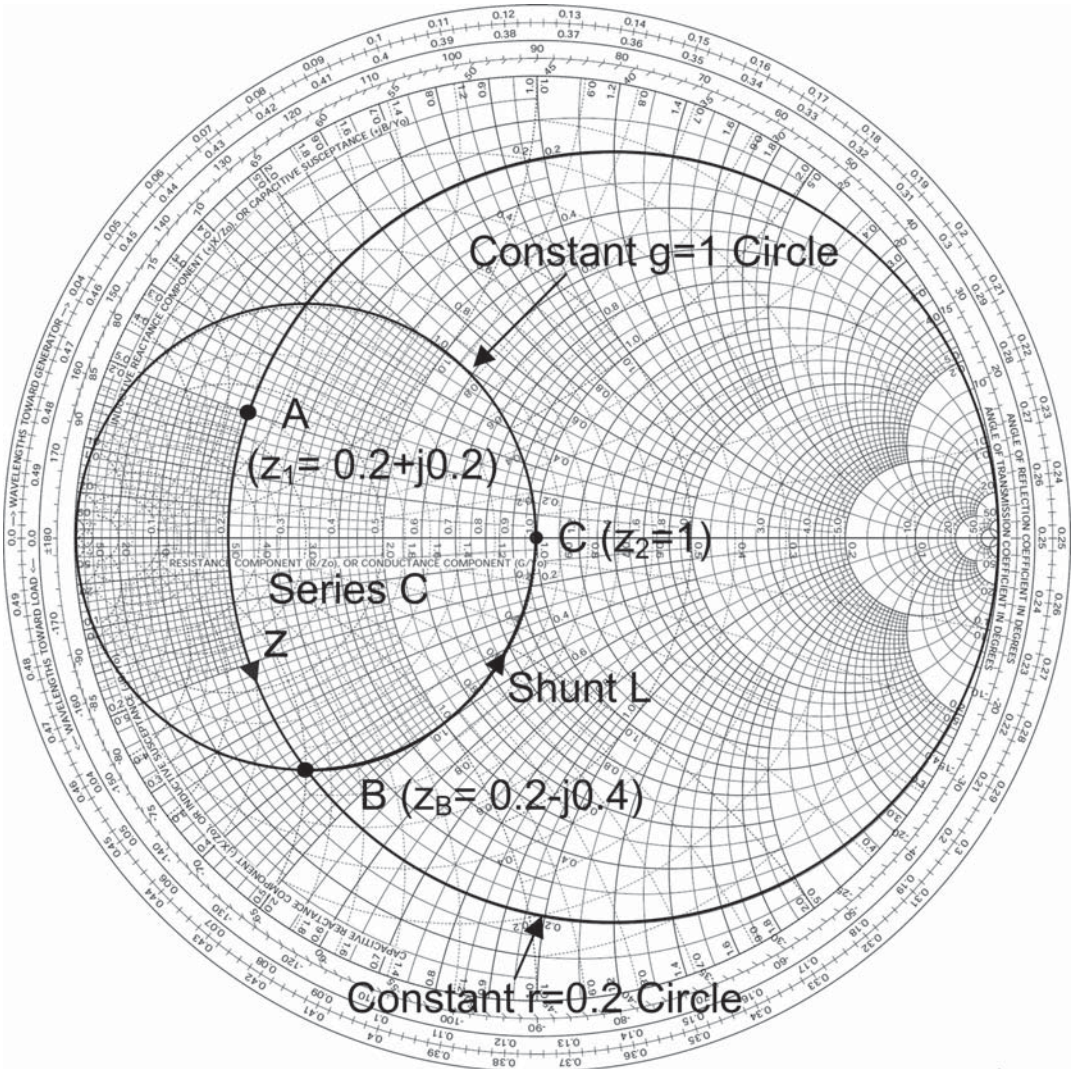
1. Locate the normalized impedance $z_1 = (10 + j10)/50 = 0.2 + j0.2$ and $z_2 = 50/50 = 1$ (with respect to 50 Ω) at points A and C on a Smith $Z-Y$ chart, respectively, as shown in Figure 6.10(a).
2. Move from point A along the $r = 0.2$ circle (on the Z-chart) to point B corresponding to the normalized impedance $z_B = 0.2 - j0.4$ (read from the Z-chart), or normalized admittance $y_B = 1 + j2$ (read from the Y-chart), and then along the $g = 1$ circle (on the Y-chart) to reach the desired normalized impedance z_2 at point C as described below.
   a. Move from A to B along the $r = 2$ circle: This movement is in the CCW direction which, as indicated by Table 6.1, corresponds to a series capacitor C as shown in Figure 6.10(b). Figure 6.10(b) shows that the normalized impedance of the capacitor C is obtained as $z_C = x_B - x_1 = -j0.4 - j0.2 = -j0.6$

which, after de-normalized (by 50 Ω), becomes $Z_C = -j0.6 \times 50 = -j30$ Ω. The capacitance C can then be determined as $C = 1/\omega Z_C = 1/2\pi \times 10 \times 10^9 \times 30 = 0.53$ pF.

b. Move from B to C along the $g = 1$ circle: This movement is in the CCW direction which, as indicated by Table 6.1, corresponds to a shunt inductor L as shown in Figure 6.10(c). Figure 6.10(c) shows that the normalized admittance of the inductor L is obtained as $y_L = b_2 - b_B = -j0.2$ which, after de-normalized (by 0.02 ℧), becomes $Y_L = -j2 \times 0.02 = -j0.04$ ℧. The inductance L can then be calculated as $L = 1/\omega Y_L = 1/0.04 \times 2\pi \times 10 \times 10^9 = 0.398$ nH.

Figure 6.10(c) shows the designed impedance-matching network that consists of a series capacitor $C = 0.53$ pF and a shunt inductor $L = 0.398$ nH. It is noted, as can be seen in the Smith chart, that there are various paths leading A to C, which correspond to different impedance-matching networks consisting of two or more lumped elements connected in series and shunt.

The same design procedure can also be used for the design of a transmission-line impedance-matching network equivalent to the lumped-element impedance-matching network in Figure 6.10(c), in which series and shunt open-circuited (or short-circuited) transmission lines are used for the series capacitor and shunt



(a)

Figure 6.10. (a–e) Design of impedance-matching network matching complex to real impedance.

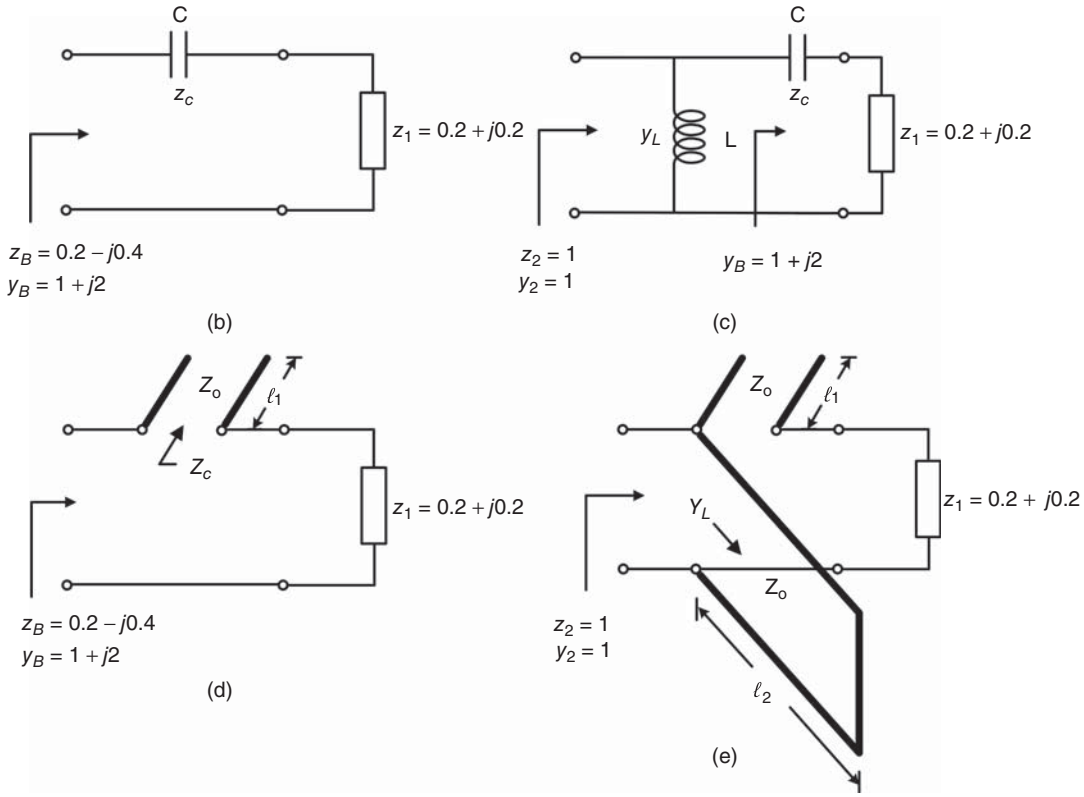


Figure 6.10. (*Continued*)

inductor, respectively. For illustration purpose, we use a series open-circuited and shunt short-circuited transmission line for the series capacitor C and shunt inductor L, respectively. We also assume that the transmission lines are 82-Ω microstrip lines on SiO_2 dielectric having thickness of 15 μm, relative dielectric constant of 3.9, and loss tangent of 0.0002. The conductors are copper having thickness of 1 μm and conductivity of 5.8×10^7 ℧/m. The effective dielectric constant (ε_{eff}) of the considered 82-Ω microstrip line at 10 GHz is calculated as 2.4505.

Figure 6.10(d) shows the resultant circuit as the normalized impedance z_1 at A is moved to B. The impedance looking into the open-circuited transmission line is $Z_C = -j30\ \Omega = -jZ_o \cot \beta \ell_1$, where Z_o and $\beta = 2\pi f \sqrt{\varepsilon_{\text{eff}}}/c$ are the characteristic impedance and phase constant of the transmission line, respectively. For the considered 82-Ω microstrip line at 10 GHz, we obtain $\ell_1 = 3.721$ mm.

Figure 6.10(e) shows the resultant impedance-matching circuit after the impedance z_1 is moved from A to B and then to C. The admittance looking into the shunt short-circuited transmission line is $Y_L = -j0.04\ ℧ = -j/Z_o \tan \beta \ell_2$, from which, we can calculate $\ell_2 = 0.903$ mm for the considered 82-Ω microstrip line at 10 GHz.

We notice from Figure 6.10(a) that we can also use an open- or short-circuited transmission line in series with z_1 to transform z_1 at A to a real impedance (located on the central horizontal line), and then transfer that real impedance to the desired impedance z_2 at C using a quarter-wavelength transmission line.

As can be seen from the Smith chart shown in Figure 6.10(a), we can also move from C corresponding to z_2 to B and then to A corresponding to z_1 to obtain an impedance-matching network matching z_2 to z_1. This impedance-matching network, however, is different from the one obtained in Figure 6.10(c). Going from C to B requires a shunt capacitor (corresponding to the CW movement along the *g*-circle), instead of a shunt inductor, and a series inductor (corresponding to CW movement along the *r*-circle) instead of a series capacitor. We can of course obtain exactly the same impedance-matching circuit as shown in Figure 6.10(c) if we move along the CCW direction from C to another location and then to A.

In the second example, a matching network is designed to match a real impedance $Z_1 = 50\ \Omega$ to a complex impedance $Z_2 = 50 - j50\ \Omega$ at 10 GHz. The design steps are similar to the previous example as described below.

1. Locate the normalized impedance $z_1 = 1$ and $z_2 = 1 - j1$ with respect to 50 Ω at A and C on a Smith Z–Y chart, respectively, as shown in Figure 6.11(a).
2. Move from A along the $r = 1$ circle (on the Z-chart) to B, corresponding to the normalized impedance $z_B = 1 + j1$ or normalized admittance $y_B = 0.5 - j0.5$, and then along the $g = 0.5$ circle (on the Y-chart) to reach the desired normalized impedance z_2 at C. It is noted that there are other impedance-matching networks corresponding to different paths from A to C.
   a. Move from A to B along the $r = 1$ circle: This movement is in the CW direction which, according to Table 6.1, corresponds to a series inductor L as shown in Figure 6.11(b). The normalized impedance of the inductor L is obtained as $z_L = x_B - x_1 = j1$, from which $Z_L = j50\ \Omega$. The inductance can then be determined as $L = 50/\omega = 50/2\pi \times 10 \times 10^9 = 0.796$ nH.
   b. Move from B to C along the $g = 0.5$ circle: This movement is in the CW direction which, according to Table 6.1, corresponds to a shunt capacitor C as shown in Figure 6.11(c). The normalized admittance of the capacitor C is obtained as $y_C = b_2 - b_B = j1$, from which we obtain $Y_C = j0.02$ ℧. The capacitance C can then be determined as $C = 0.02/\omega = 1/2\pi \times 10 \times 10^9 = 0.318$ pF.

The designed matching network consists of a series inductor $L = 0.796$ nH and a shunt capacitor $C = 0.318$ pF as shown in Figure 6.11(c). Following the same procedure as in the first example, we can also design a matching network consisting of transmission lines.

We now can see that, theoretically, virtually any lumped-element or transmission-line impedance-matching network can be designed to match one impedance to another impedance. We also see that these networks have the high pass filter configuration, such as that in Figure 6.10(c), or the low pass filter topology like that Figure 6.11(c). With further expansion, we can also include the band-pass topology for impedance-matching networks. In practice, however, not all of the designed matching networks can be realized since the obtained elements such as inductors or capacitors may not be realizable. This, in turn, limits the number of possible impedance-matching networks. Furthermore, the choice of a particular impedance-matching network depends on the number of elements used, in which less elements are typically preferred due to possible smaller loss and size, the quality factor of one element versus another (such as inductor versus capacitor), and the desired frequency response like low pass, high pass, or band-pass response. It is noted that, although only two elements are used for the impedance-matching networks in the foregoing examples, the same procedure can be easily extended for impedance-matching networks having more than two elements.

6.2.4 Transmission-Line Impedance-Matching Networks

Transmission-line matching networks are inevitable in high RF ranges where lumped elements are not realizable and/or have comparable size with lower quality factor to transmission lines. In general, transmission-line matching networks consist of series and shunt open- and short-circuited transmission lines, and cascaded transmission lines. Different kinds of transmission lines (e.g., microstrip line or coplanar waveguide) or a combination of various transmission lines can be (or may need to be) employed to realize a particular high or low characteristic impedance, a particular transmission line configuration (e.g., series connection or short-circuited), a particular interface or connection with other circuit elements or other circuits.

6.2.4.1 Quarter-Wave Transformer. Quarter-wave transformer is a transmission line that has a physical length of $\lambda/4$, where λ is the wavelength, or an electrical length of 90°, used to match between two real impedances. We consider a lossless quarter-wave transformer connected to a real impedance at either end as shown in Figure 6.12. The input impedance of the quarter-wave transformer is obtained as

$$Z_{\text{in}} = Z_o \frac{R_L + jZ_o \tan \beta\ell}{Z_o + jR_L \tan \beta\ell} = \frac{Z_o^2}{R_L} \tag{6.10}$$

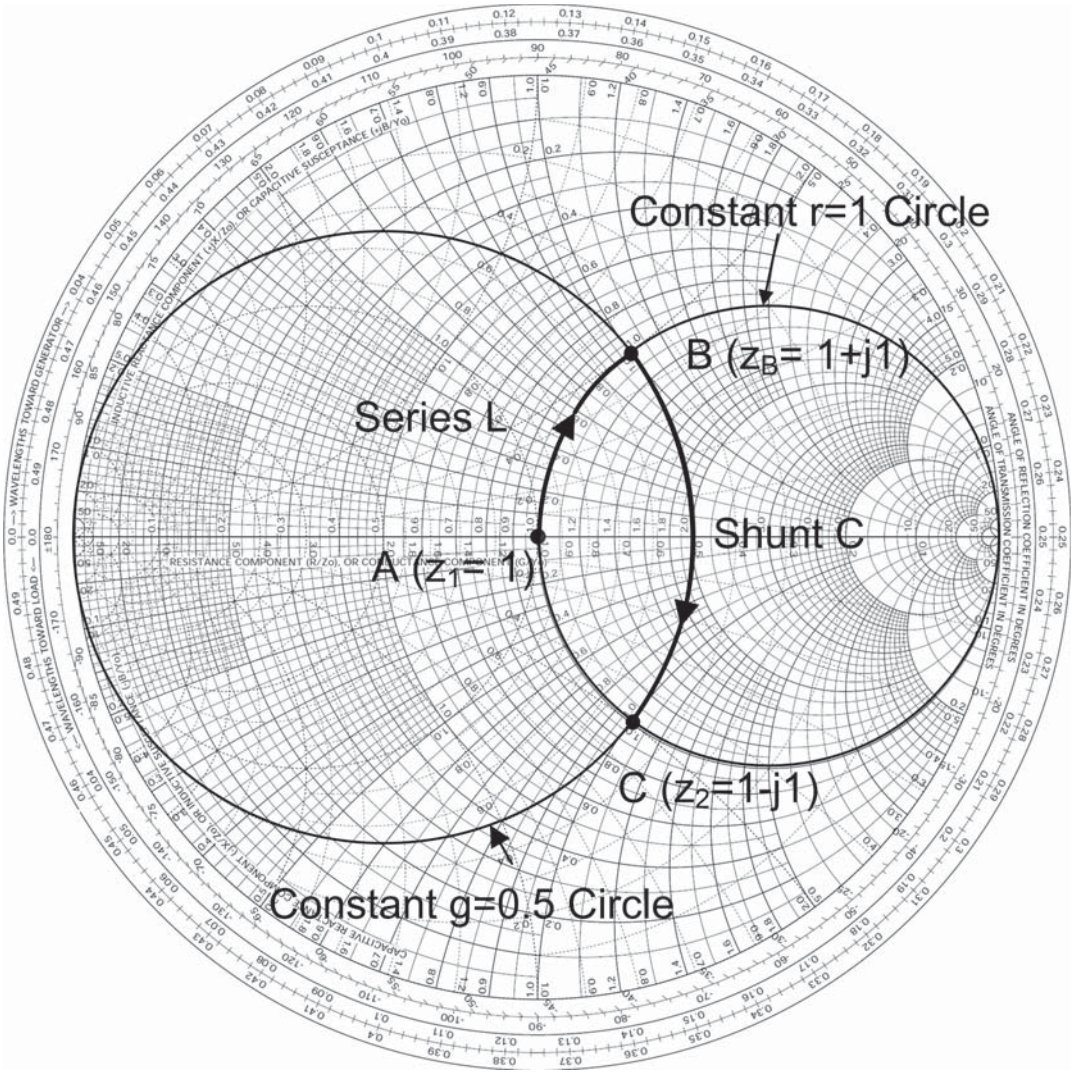


Figure 6.11. (a–c) Design of impedance-matching network matching real to complex impedance.

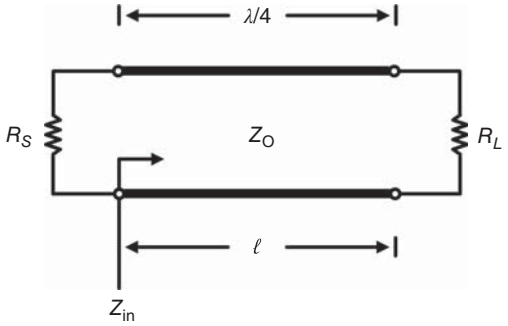


Figure 6.12. Quarter-wave transformer.

making use of $\beta\ell = \pi/2$. Equation (6.10) leads to, after enforcing the perfect impedance-match condition $Z_{\text{in}} = R_s$:

$$Z_o = \sqrt{R_L R_s} \tag{6.11}$$

which specifies the characteristic impedance of a quarter-wave transformer that provides impedance-matching between two real impedances.

If one of the impedances is not real, for instance, the load impedance Z_L, then we can transform Z_L to a real impedance R_L by inserting a transmission line immediately in front of Z_L before using a quarter-wave transformer to match R_L to R_s as shown in Figure 6.13(a). Assume the inserted transmission line is lossless, moving along the transmission line is equivalent to moving on a constant-Γ circle on a Smith chart. Therefore, the complex impedance Z_L can be transformed to a real impedance R_L by moving toward generator (CW direction) along the constant-Γ circle at Z_L on a Smith chart, as shown in Figure 6.13(b). There exist two real values for R_L which result in two possible values for the transmission-line length (d_1 and d_2) that can be determined directly from the Smith chart as noted in Figure 6.13(b). The characteristic impedance Z_o of the quarter-wave transformer can then be determined using (6.11). Knowing Z_o, R_s, d_1 (or d_2), and Z_L enable the characteristic impedance Z_o' of the inserted transmission line to be determined.

6.2.4.2 Single-Stub Impedance-Matching Networks. Figure 6.14 shows four transmission-line impedance-matching networks, namely single-stub impedance-matching networks, which consist of a shunt or series open- or short-circuited transmission line (stub) and a cascaded transmission line. The matching networks based on series stubs are difficult to realize unless the series stubs are transformed to shunt stubs using techniques such as Kuroda identities. These transmission-line matching networks are very basic matching networks and the ones employing shunt elements are well-known to RF engineers. These matching networks as well their designs serve as the basic that can be extended to develop and design other topologies for transmission-line matching networks.

We illustrate the design formulation by considering a single-stub network using a short-circuited stub as shown in Figure 6.15, that matches a complex load impedance Z_L to a real impedance Z_o representing the characteristic impedance of the transmission line connected to the input of the matching network or the impedance presented at that input port. All impedances and admittances are normalized with respect to the characteristic impedance Z_o and characteristic admittance $Y_o = 1/Z_o$, respectively. We define y_M as the normalized admittance looking into the matching network, y_S as the normalized admittance looking into the short-circuited stub, y_T as the normalized admittance looking into the cascaded transmission line, and y_L and Z_L as the respective normalized load admittance and impedance.

We note that, for perfect match to occur between Y_L and Y_o, the admittance looking into the matching network must be equal to Y_o, or

$$y_M = y_T + y_S = 1 \tag{6.12}$$

Equation (6.12) sets the criterion from which the lengths of the transmission lines in the matching network can be determined for given characteristic impedances. This criterion can be rewritten into the following

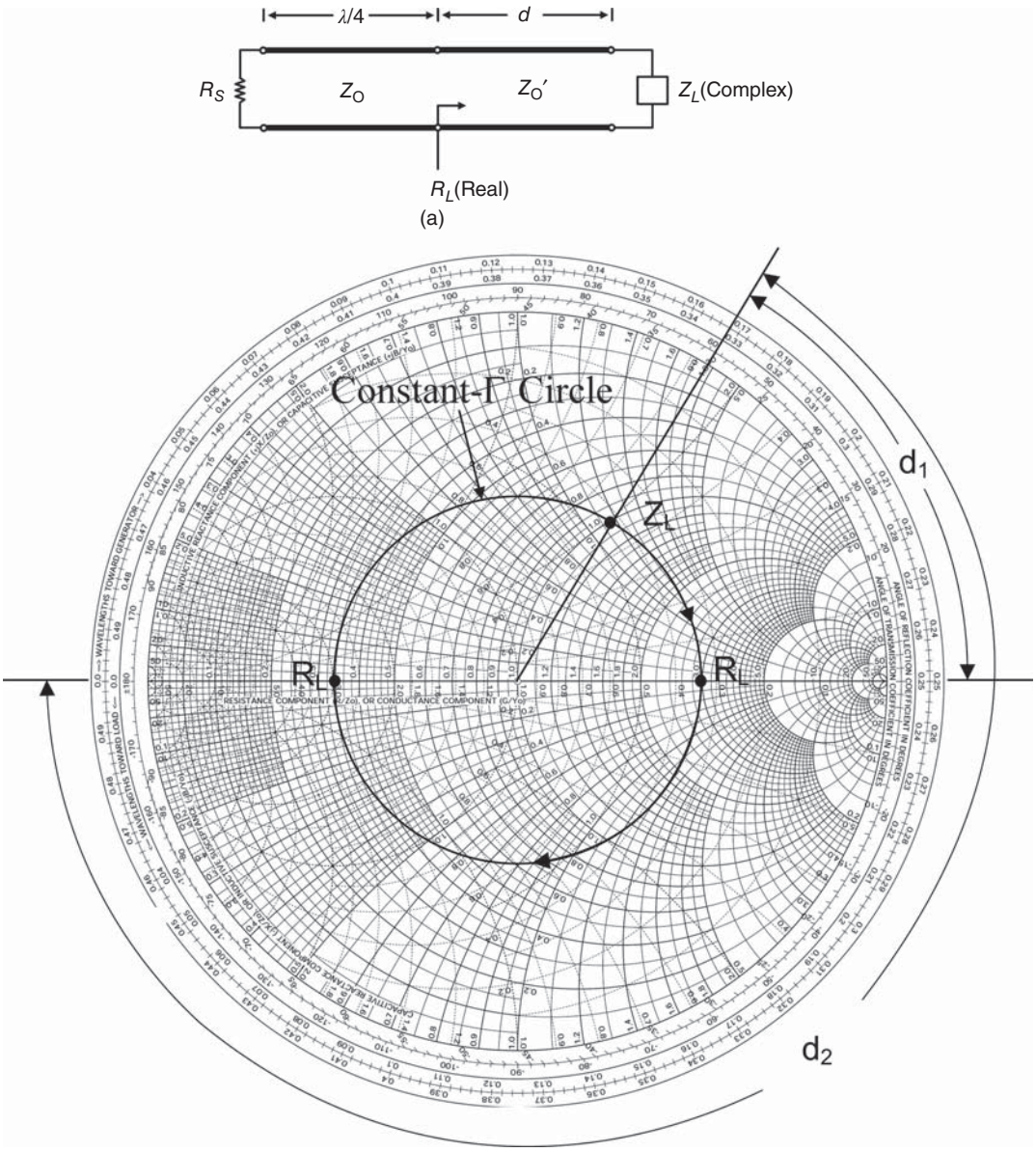


Figure 6.13. (a, b) Matching between a complex and real impedances using a quarter-wave transformer and transmission line.

two conditions upon using the normalized input admittance of the short-circuited stub $y_S = -jY_o' \cot \beta\ell / Y_o \equiv -jb_S$:

$$y_T = 1 + jb_S \tag{6.13}$$

$$y_S = -jb_S \tag{6.14}$$

Equation (6.13) sets the criterion for determining the length d; it shows that d is determined such that the real part of y_T is equal to 1. Equation (6.14) shows that the length ℓ is determined such that$y_S = -jb_S$. It is noted that the short-circuited stub in Figure 6.15 can be replaced with an open-circuited stub whose normalized admittance is $y_S = jY_o' \tan \beta\ell / Y_o \equiv -jb_S$. The same design formulation applies to single-stub matching networks employing open-circuited stubs and can also be extended for multi-stub impedance-matching networks. Figure 6.16 shows some possible double-stub impedance-matching networks.

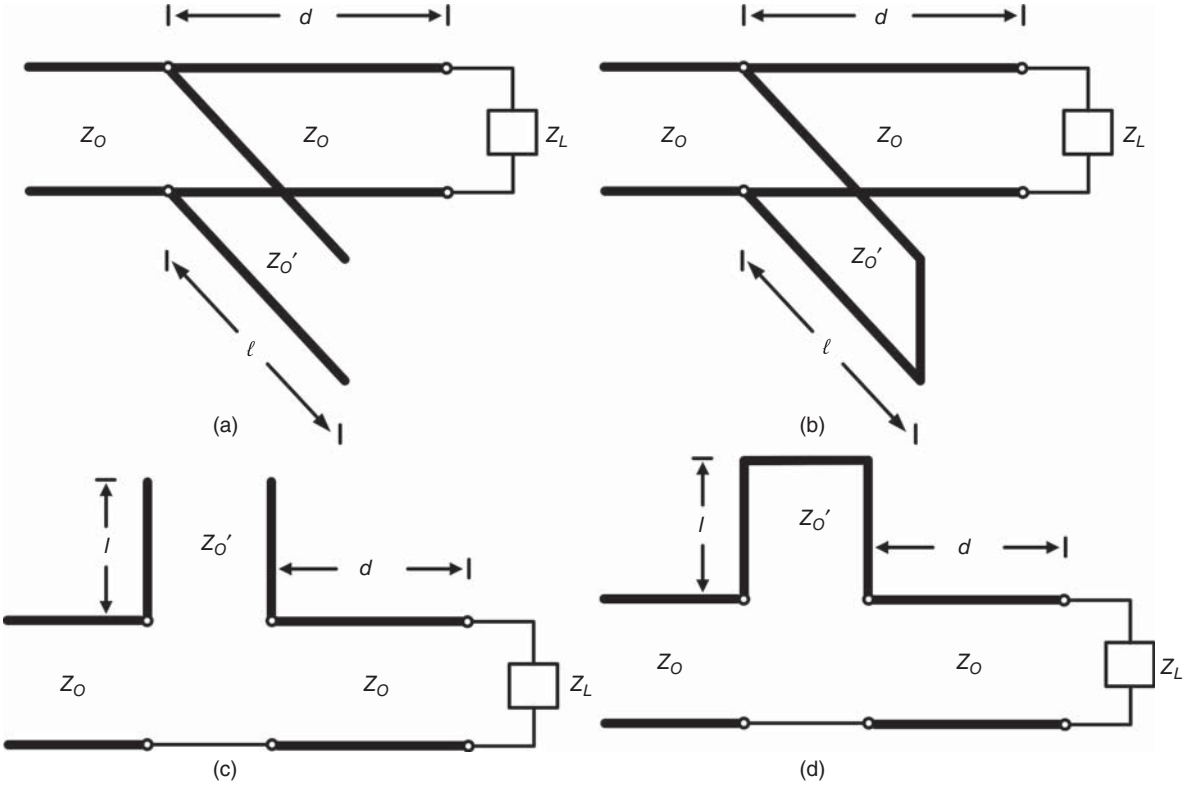


Figure 6.14. (a–d) Single-stub impedance-matching networks.

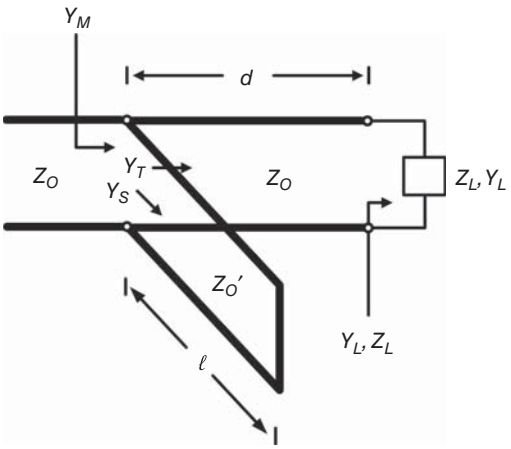


Figure 6.15. Single short-circuited stub impedance-matching network.

To illustrate the design procedure, we consider again the single-stub network as shown in Figure 6.15 and assume $Z_L = 100 + j100\ \Omega$, $Z_o = 50\ \Omega$, and $Z_o' = 100\ \Omega$. The following steps show the design procedure:

1. Locate the normalized load impedance $z_L = (100 + j100)/50 = 2 + j2$ at point A on a Smith chart as shown in Figure 6.17(a).
2. Locate the normalized load admittance y_L at point B which is opposite from Z_L on the constant-Γ circle through the center of the Smith chart.

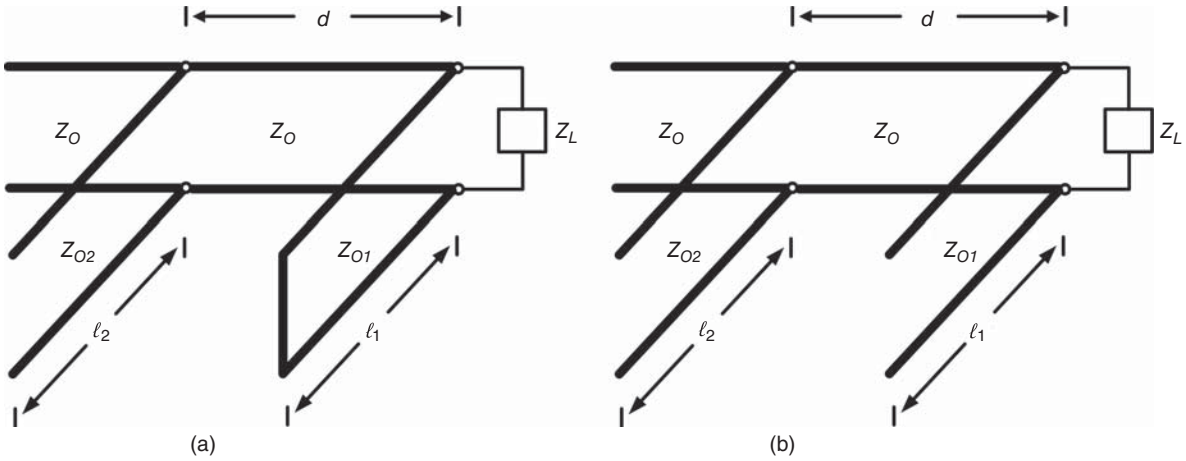


Figure 6.16. (a, b) Double-stub impedance-matching networks.

3. Draw the constant-Γ circle passing through y_L. This circle intersects the constant $g = 1$ circle at two points corresponding to normalized admittances: $y_{T1} = 1 + j1.6 = 1 + jb_{S1}$ and $y_{T2} = 1 - j1.6 = 1 + jb_{S2}$ at C and D, respectively.
4. There are two different solutions for the length d of the cascaded transmission line according to different movement from y_L to y_{T1} or y_{T2} as follows.
   a. Move from y_L to y_{T1} toward generator along the constant-Γ circle: The locations of y_L at B and y_{T1} at C correspond to 0.458λ and 0.180λ on the circle marked "wavelength toward generator," respectively, which give $d_1 = 0.50\lambda - 0.458\lambda + 0.180\lambda = 0.222\lambda$.
   b. Move from y_L to y_{T2} toward generator along the constant-Γ circle: The location of y_{T2} at D corresponds to 0.32λ on the circle marked "wavelength toward generator" from which we can obtain $d_2 = 0.50\lambda - 0.458\lambda + 0.32\lambda = 0.362\lambda$.
5. Since the characteristic impedance of the short-circuited stub (Z'_o) is different from the characteristic impedance (Z_o) used for normalization, we need to de-normalize and then normalize with respect to $Y_o = 1/Z_o$ and $Y'_o = 1/Z'_o$, respectively.
   a. Use the first value for d (d_1): $y_{S1} = -jb_{S1}Y_o/Y'_o = -j3.2$ utilizing jb_{S1} from step 3. Find the length ℓ of the short-circuited stub to produce $y_{S1} = -j3.2$ by moving toward generator from the short location (infinite admittance) corresponding to 0.25λ to $-j3.2$ at point E corresponding to 0.298λ on the circle marked "wavelength toward generator." This results in $\ell_1 = 0.298\lambda - 0.25\lambda = 0.048\lambda$. Note that if $Z'_o = Z_o$ then we can use $y_{S1} = -jb_{S1} = j1.6$ directly from step 3 and obtain $\ell_1 = 0.339\lambda - 0.25\lambda = 0.089\lambda$.
   b. Use the second value for d (d_2): $y_{S2} = -jb_{S2}Y_o/Y'_o = j3.2$ utilizing jb_{S2} from step 3. Find the length ℓ of the short-circuited stub to produce $y_S = j3.2$ by moving toward generator from the short location to $j3.2$ at point F corresponding to 0.202λ on the circle marked "wavelength toward generator." This results in $\ell_2 = 0.5\lambda - 0.202\lambda = 0.298\lambda$.
6. We choose $d_1 = 0.222\lambda$ and $\ell_1 = 0.089\lambda$ since they are shorter than d_2 and ℓ_2, respectively, and hence have less loss. Figure 6.17(b) shows the designed single-stub impedance-matching network. The physical lengths and dimensions of the transmission lines can be determined for specific transmission lines (e.g., microstrip line) at a certain frequency.

The foregoing design methods revolve around a single frequency and hence results in narrow-band impedance-matching networks. It is relatively straight forward to design an impedance-matching network, using either lumped elements or transmission lines, to match two real impedances over a wide bandwidth (pass-band) while rejecting the signals out-of-band (stop-band) by following the design for high pass, low

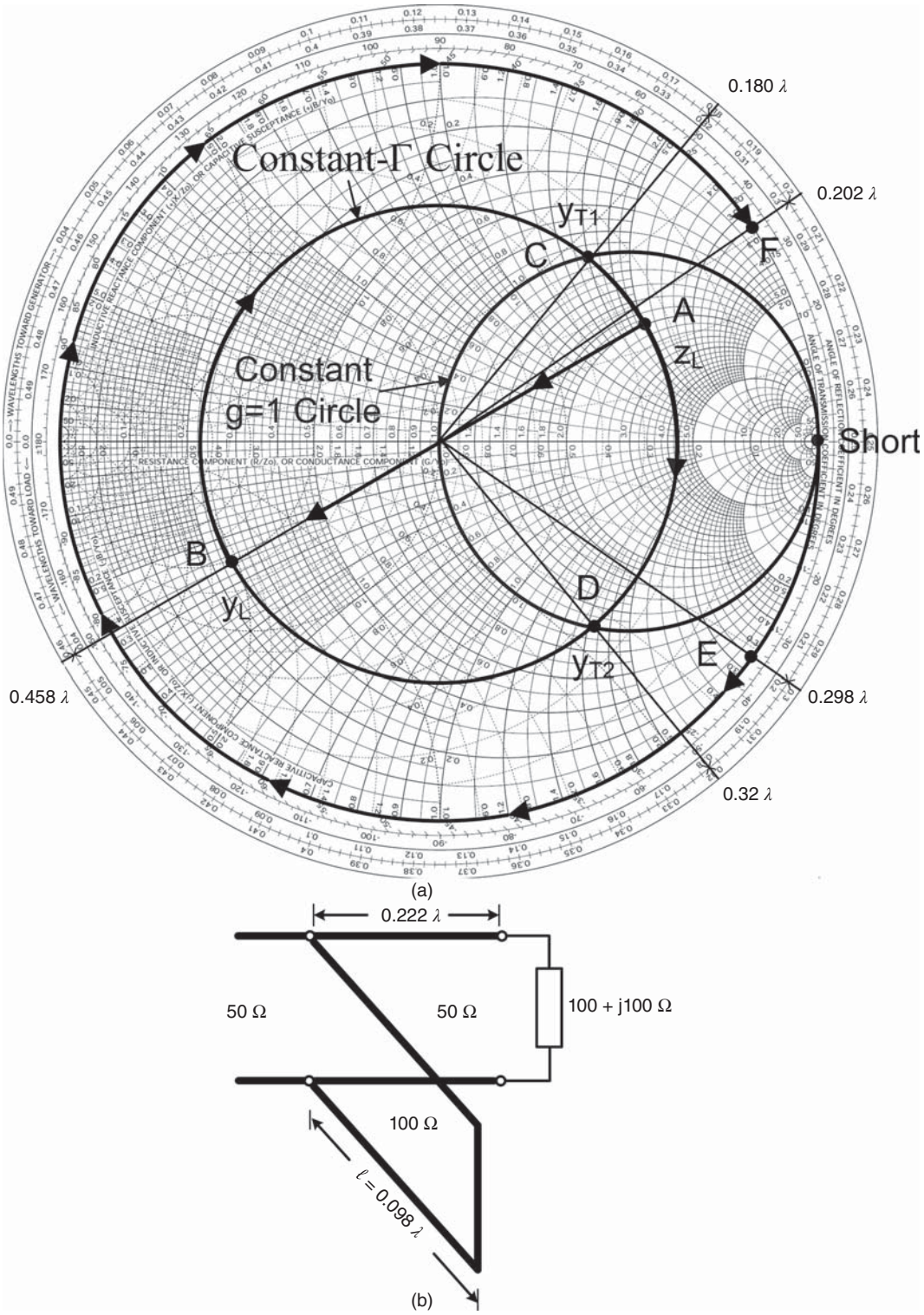


Figure 6.17. (a, b) Design of single-stub impedance-matching network.

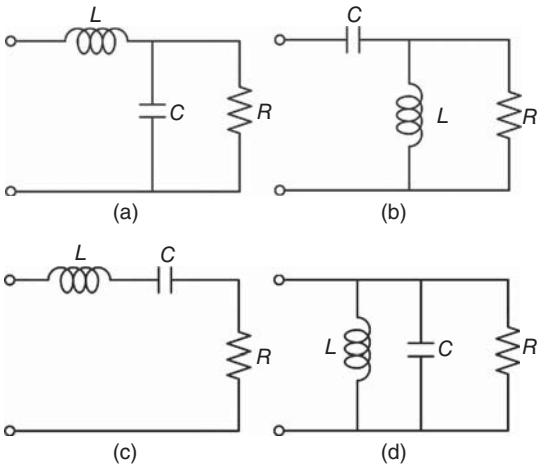


Figure 6.18. Representation of complex impedances for broadband matching: (a) low pass, (b) high pass, and (c), (d) band-pass.

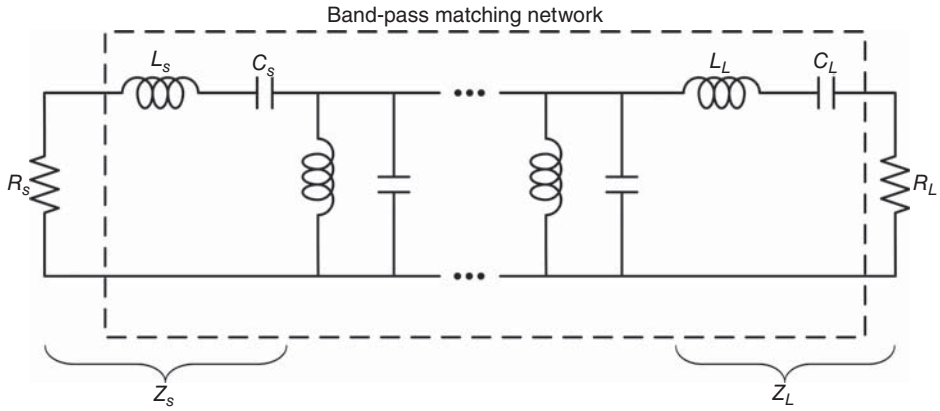


Figure 6.19. Absorption of reactive parts (L_s, C_S) and (L_L, C_L) of respective complex impedances Z_S and Z_L into broadband matching network.

pass, or band-pass filter. However, when one or both impedances are complex, which vary as a function of frequency, the design of impedance-matching networks is more involved. An attractive way to overcome the frequency-dependence of the matching impedances, and hence producing broadband impedance matching, is to incorporate the (frequency-dependent) reactive parts of the matching impedances as part of the matching network, or equivalently absorbing them into the matching network. This concept can be employed to design broadband impedance-matching networks such as those for wideband amplifiers. The basic idea is to separate the complex impedance into a real part and a reactive part that represents either a low pass, high pass, or band-pass element, as shown in Figure 6.18, depending on desired frequency response. The impedance-matching network can then be designed to match between the real parts of the complex impedances across a desired bandwidth based on some topologies such as low pass, high pass, or band-pass filter. Once the design is completed, the reactive parts of the complex impedances are absorbed into the impedance-matching network as part of the matching network. Figure 6.19 illustrates the incorporation of the reactive parts in a band-pass impedance-matching network.

6.3 KURODA IDENTITIES

We mention briefly in Section 6.2.1 that it is very difficult, if not impossible, to connect transmission lines in series. This difficulty is due to the fact that transmission lines have four terminals (two for each end or port

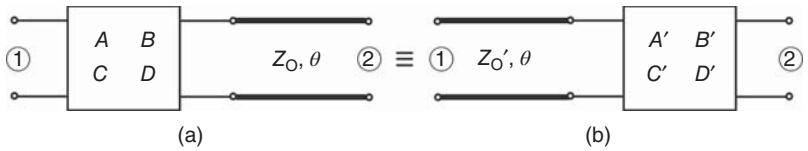


Figure 6.20. (a, b) Principle of Kuroda identities. Transmission lines have the same electrical length.

of transmission lines), which differ from lumped elements that have only two terminals. Kuroda identities [2] overcome this problem by converting a transmission-line network having series open- and/or short-circuited transmission lines into an equivalent one having only shunt transmission lines.

Figure 6.20 shows the principle of the Kuroda identities based on the equivalence between two two-port networks. One network consists of an element, represented by a chain matrix with parameters A, B, C, D, in cascade with a lossless transmission line, having characteristic impedance Z_o and electrical length θ, as shown in Figure 6.20(a). Another network comprises a lossless transmission line, with characteristic impedance Z_o' and electrical length θ, in cascade with an element, represented by a chain matrix whose parameters are A', B', C', D', as shown in Figure 6.20(b).

The chain matrices of the two-port networks in Figures 6.20(a) and (b) can be derived, based on Section 7.6, as

$$\frac{S}{\sqrt{1-S^2}}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} \frac{1}{S} & Z_o \\ \frac{1}{Z_o} & \frac{1}{S} \end{bmatrix} = \frac{S}{\sqrt{1-S^2}}\begin{bmatrix} \frac{A}{S}+\frac{B}{Z_o} & AZ_o+\frac{B}{S} \\ \frac{C}{S}+\frac{D}{Z_{o1}} & CZ_o+\frac{D}{S} \end{bmatrix} \tag{6.15}$$

and

$$\frac{S}{\sqrt{1-S^2}}\begin{bmatrix} \frac{1}{S} & Z_o' \\ \frac{1}{Z_o'} & \frac{1}{S} \end{bmatrix}\begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \frac{S}{\sqrt{1-S^2}}\begin{bmatrix} \frac{A'}{S}+C'Z_o' & \frac{B'}{S}+D'Z_o' \\ \frac{A'}{Z_o'}+\frac{C'}{S} & \frac{B'}{Z_o'}+\frac{D'}{S} \end{bmatrix} \tag{6.16}$$

respectively, where $S = j\tan\theta$. When the networks in Figure 6.20 are equivalent, their chain matrices, and correspondingly their respective A, B, C, and D parameters, are equal:

$$\frac{A}{S}+\frac{B}{Z_o} = \frac{A'}{S}+C'Z_o' \tag{6.17}$$

$$\frac{C}{S}+\frac{D}{Z_o} = \frac{A'}{Z_o'}+\frac{C'}{S} \tag{6.18}$$

$$AZ_o+\frac{B}{S} = \frac{B'}{S}+D'Z_o' \tag{6.19}$$

$$CZ_o+\frac{D}{S} = \frac{B'}{Z_o'}+\frac{D'}{S} \tag{6.20}$$

Solving (6.17) and (6.18), we get

$$A' = \frac{A+\left(\frac{S}{Z_o}\right)B-(SZ_o')C-\left(S^2\frac{Z_o'}{Z_o}\right)D}{1-S^2} \tag{6.21}$$

$$C' = \frac{C+\left(\frac{S}{Z_o}\right)D-\left(\frac{S}{Z_o'}\right)A-\left(\frac{S^2}{Z_oZ_o'}\right)B}{1-S^2} \tag{6.22}$$

while solving (6.19) and (6.20) gives

$$B' = \frac{B + (SZ_o)A - (SZ'_o)D - (S^2 Z_o Z'_o)C}{1 - S^2} \tag{6.23}$$

$$D' = \frac{D + (SZ_o)C - \left(\frac{S}{Z'_o}\right)B - \left(S^2 \frac{Z_o}{Z'_o}\right)A}{1 - S^2} \tag{6.24}$$

The impedance parameters $Z_{11,1}$ and $Z_{11,2}$ at port 1 of the two-port networks in Figures 6.20(a) and 6.20(b), respectively, can be derived from their chain matrix parameters according to the conversion formulas given in Table 7.1 as

$$Z_{11,1} = \frac{\frac{A}{S} + \frac{B}{Z_o}}{\frac{C}{S} + \frac{D}{Z_o}} \tag{6.25}$$

$$Z_{11,2} = \frac{\frac{A'}{S} + C'Z'_o}{\frac{A'}{Z'_o} + \frac{C'}{S}} \tag{6.26}$$

For $S = 1$, we can obtain from (6.26):

$$Z_{11,2}(S=1) = \frac{A'(S=1) + C'(S=1)Z'_o}{\frac{A'(S=1)}{Z'_o} + C'(S=1)} = Z'_o \tag{6.27}$$

For the equivalent two-port networks, $Z_{11,1} = Z_{11,2}$ which, upon letting $S = 1$ and utilizing (6.27), leads to

$$Z'_o = \frac{A(S=1)Z_o + B(S=1)}{C(S=1)Z_o + D(S=1)} \tag{6.28}$$

Equations (6.21)–(6.24) and (6.28) indicate that the two-port network shown in Figure 6.20(b) can be determined completely from its equivalent network shown in Figure 6.20(a). Similarly, we can also derive equations for the chain matrix parameters A, B, C, D and the characteristic impedance Z_o of the transmission line in the two-port network shown in Figure 6.20(a) in terms of the corresponding parameters of the two-port network shown in Figure 6.20(b). This, in turn, allows us to find the two-port network in Figure 6.20(a) from that in Figure 6.20(b).

Making use of (6.21)–(6.24) and (6.28), various equivalent two-port networks can be derived. Some of which are shown in Figure 6.21. It is noted that the electrical lengths of all the transmission lines in each pair of the equivalent networks are the same. Figure 6.20(e), (f) and 6.20(g), (h) are obtained directly from Figure 6.20(a), (b) and 6.20 (c), (d), respectively, by making use of the equivalence in Figure 6.20 which shows that the network in Figure 6.20(b) is essentially the same as the network in Figure 6.20(a) when ports 1 and 2 are interchanged.

The Kuroda identities, through Figure 6.21, demonstrate that we can convert a series short-circuited (or similarly, a series open-circuited) transmission line to a shunt transmission line, hence avoiding the complication of connecting a transmission line in series in RFICs. This is achieved by inserting a transmission line (assumed lossless), having characteristic impedance equal to the source or load impedance (assumed real) and length equal to the length of the series short-circuit (or open-circuited) transmission line, between the

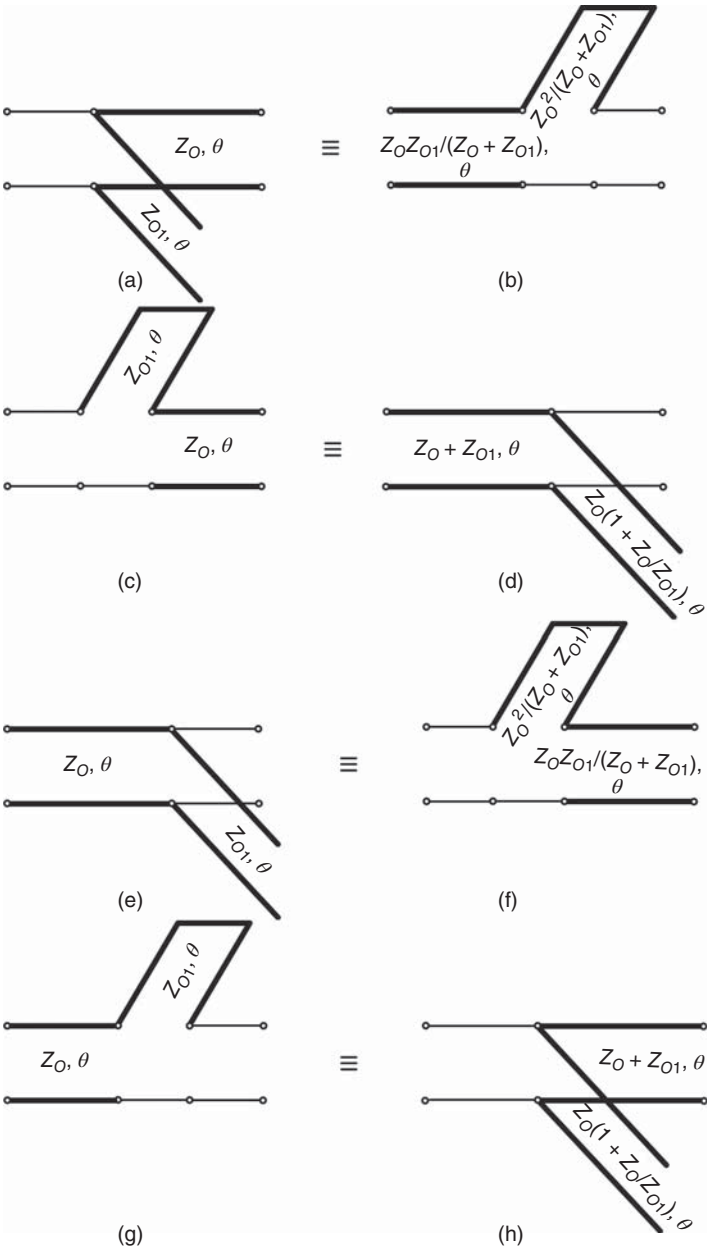


Figure 6.21. (a–h) Equivalent networks based on Kuroda identities.

source or load, respectively, and the series short-circuit (or open-circuited) transmission line. The insertion of such transmission line at the source or load does not affect the performance of the network since it does not change the impedance level of the source or load. As an example, we consider a lossless transmission-line network as shown in Figure 6.22(a). At the source, we insert a lossless transmission line having characteristic impedance equal to the source real impedance Z_S and electrical length θ_1. At the load, we insert another lossless transmission line having characteristic impedance equal to the load real impedance Z_L and electrical length θ_3. The new two-port network is shown in Figure 6.22(b). Now by applying the Kuroda identities to the network in Figure 6.22(b), or specifically the equivalences in Figure 6.21(g), (h), (c), and (d), we can derive the final equivalent two-port network as shown in Figure 6.22(c) consisting of only cascaded and shunt transmission lines that can be easily realized.

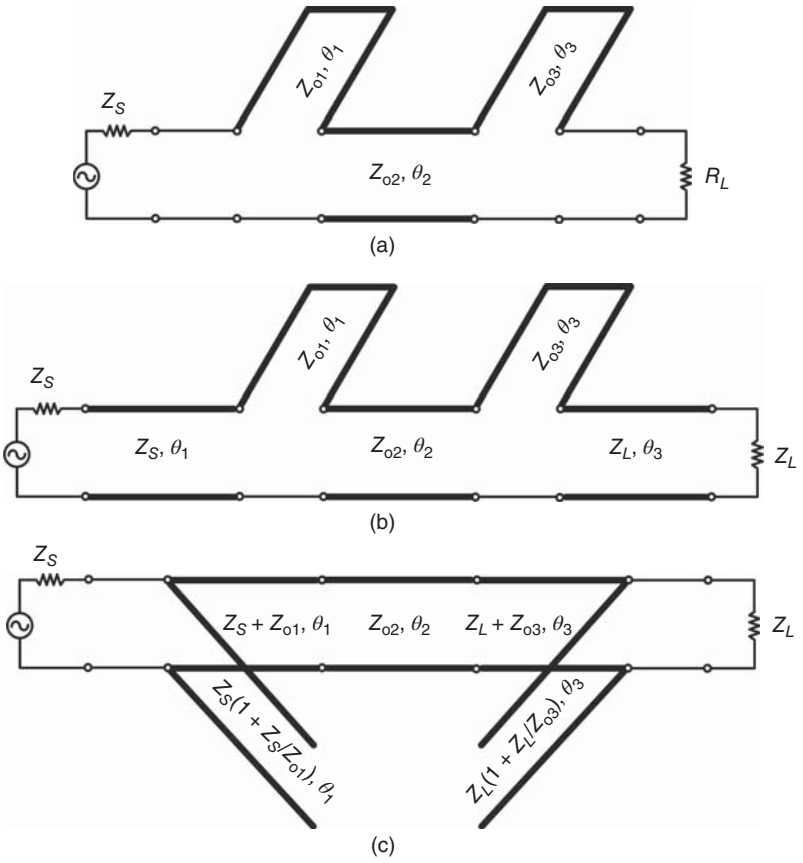


Figure 6.22. Application of Kuroda identities: (a) original transmission-line network, (b) insertion of transmission lines, and (c) equivalent transmission-line network.

REFERENCES

1. P. H. Smith, "Transmission-Line Calculator," *Electronics*, Vol. 12, pp. 29–31, Jan. 1939.
2. G. C. Temes and T. W. Lapatra, *Introduction to Circuit Synthesis and Design*, McGraw-Hill, New York, New York, pp. 342–348, 1977.

PROBLEMS

6.1 Design the two-element impedance-matching networks shown in Figure P6.1 to match the load impedance $Z_L = 100 + j40\ \Omega$ to 50 Ω at 20 GHz. Verify by calculating the input impedance of the matching networks using your own analysis or a commercially available computer program.

6.2 Consider an amplifier as shown in Figure P6.2. Design the output impedance-matching network to enable maximum power transfer between the transistor and the 50-Ω load at 15 GHz. Verify by calculating the impedance looking into the output matching network.

6.3 Two types of LC networks are shown in Figure P6.3.

a) Select the network that can match the load $Y_L = (8 - j12)10^{-3}$ ℧ to 50-Ω. Provide the rationale of your choice.

b) Design the impedance-matching network in Part (a) at 10 GHz. Verify by calculating the input impedance of the matching network.

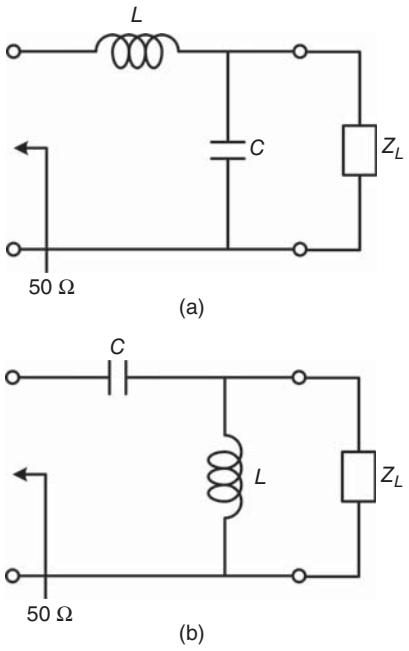


Figure P6.1.

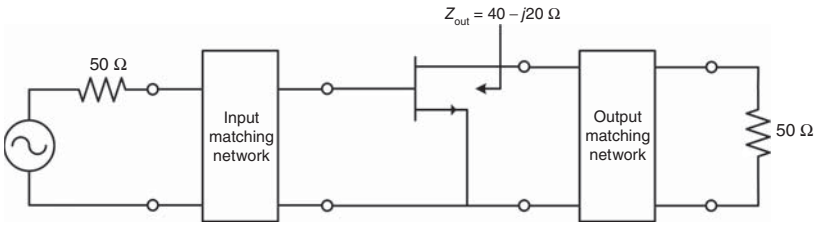


Figure P6.2.

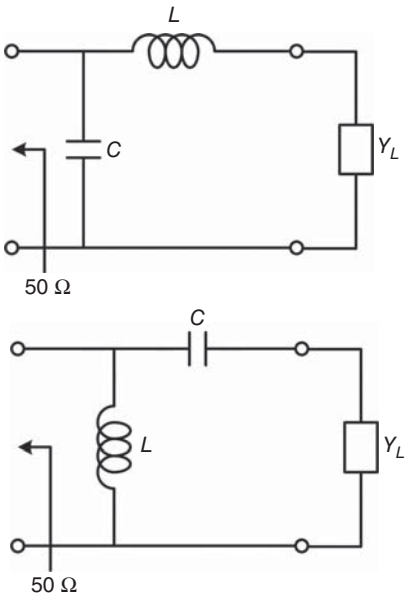


Figure P6.3.

6.4 Design a lumped-element impedance-matching network to match $Z_1 = 70 + j30\ \Omega$ to $Z_2 = 15 - j20\ \Omega$ at 30 GHz. Verify by calculating the input impedance of the matching network.

6.5 Design a lumped-element impedance-matching network to match $Z_1 = 150 - j75\ \Omega$ to $Z_2 = 80 - j35\ \Omega$. at 20 GHz. Verify by calculating the input impedance of the matching network.

6.6 Design a lumped-element impedance-matching network to match $70\ \Omega$ to $10 - j20\ \Omega$ at 20 GHz. Verify by calculating the input impedance of the matching network.

6.7 We consider the design problem in Section 6.2.2 that requires impedance-matching from $Z_1 = 10 + j10\ \Omega$ to $Z_2 = 50\ \Omega$ at 10 GHz. Design another impedance-matching network using an open- or short-circuited transmission line in series with z_1 and a quarter-wavelength transmission line. Assume the transmission lines are microstrip lines on SiO_2 dielectric having thickness of $15\ \mu m$ and relative dielectric constant of 3.9. SiO_2 and the conductors are assumed to be perfect and the conductor thickness is negligible. Verify by calculating the input impedance of the matching network.

6.8 We consider again the design problem in Section 6.2.2 that requires impedance-matching between $Z_1 = 10 + j10\ \Omega$ and $Z_2 = 50\ \Omega$ at 10 GHz. Referring to Figure 6.10(a), design another impedance-matching network to match Z_2 to Z_1 by moving from C to B and then to A. Verify by calculating the input impedance of the matching network. Compare this impedance-matching network to that in Figure 6.10(c).

6.9 We consider again the design problem in Section 6.2.2 that requires impedance-matching from $Z_1 = 10 + j10\ \Omega$ to $Z_2 = 50\ \Omega$ at 10 GHz. Instead of moving from A to B and C to obtain the impedance-matching network in Figure 6.10(c), design another impedance-matching network to match Z_1 to Z_2 by moving along the CCW direction from C to another location and then to A. Verify by calculating the input impedance of the matching network. Compare the resultant impedance-matching network to that in Figure 6.10(c).

6.10 Design a transmission-line impedance matching network to match $Z_1 = 50\ \Omega$ to $Z_2 = 50 - j50\ \Omega$ at 10 GHz using series short-circuited and shunt open-circuited microstrip lines having characteristic impedances of 100 and 30 Ω, respectively. Verify by calculating the input impedance of the matching network. The microstrip line structure is the same as that in Problem 6.7.

6.11 Design a transmission-line network as shown in Figure 6.13(a) to match $R_s = 50\ \Omega$ to $Z_L = 100 + j70\ \Omega$ at 35 GHz. Determine two possible values for d and the corresponding characteristic impedance Z'_o using microstrip lines employed in Problem 6.7. Verify by calculating the input impedance of the matching network.

6.12 Design a transmission-line network to transform $50\ \Omega$ to $30 + j10\ \Omega$ at 60 GHz. You can choose any topology for the network and values for the characteristic impedances of the transmission lines. The transmission lines are microstrip lines on SiO_2 dielectric having thickness of $15\ \mu m$ and relative dielectric constant of 3.9. We assume that SiO_2 is lossless and the conductors are perfect with zero metallization thickness. Verify by calculating the input impedance of the matching network.

6.13 We wish to match $Z_L = 155 - j70\ \Omega$ to $R_s = 50\ \Omega$ at 40 GHz using a transmission line immediately in front of Z_L followed by a quarter-wave transformer as shown in Figure 6.13(a). Determine Z_o, Z'_o, and d. Assume the transmission lines are microstrip lines given in Problem 6.7. Verify by calculating the input impedance of the matching network.

6.14 Design a single-stub impedance-matching network having the topology as shown in Figure 6.15 that matches $Z_L = 200 + j100\ \Omega$ to a transmission line having $Z_o = 100\ \Omega$ at 60 GHz. Assume $Z'_o = 50\ \Omega$ and the transmission lines are microstrip lines given in Problem 6.7. Verify by calculating the input impedance of the matching network.

6.15 Repeat Problem 6.14 using an open-circuited stub.

6.16 Repeat Problem 6.14 using $Z_o' = 100\ \Omega$.

6.17 Repeat Problem 6.15 using $Z_o' = 100\ \Omega$.

6.18 Design the impedance-matching network shown in Figure P6.4 to transfer a 50-Ω load impedance to an impedance of $20 - j30\ \Omega$ at 60 GHz. Assume $Z_{o2} = 100\ \Omega$ and the transmission lines are microstrip lines given in Problem 6.7. Verify by calculating the input impedance of the matching network.

6.19 Design the impedance-matching network shown in Figure P6.5 to match the load $Z_L = 100 + j100\ \Omega$ to a 50-Ω transmission line at 35 GHz. Assume the transmission lines are microstrip lines given in Problem 6.7. Verify by calculating the input impedance of the matching network.

6.20 Derive the two-port network in Figure 6.21(b) from that in Figure 6.21(a).

6.21 Derive the two-port network in Figure 6.21(d) from that in Figure 6.21(c).

6.22 Derive the two-port network in Figure P6.6(b) from that in Figure P6.6(a). The transformer is an ideal transformer which enables impedance transformation for the network.

6.23 In Problem 7.29, we derive an equivalent network of a two-port network consisting of a parallel-coupled transmission line. This equivalent network consists of a transmission line of characteristic impedance $Z_{oe}/2$ and electrical length θ_e in series with a short-circuited transmission line of characteristic

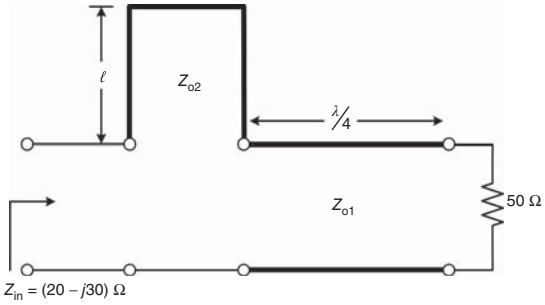


Figure P6.4.

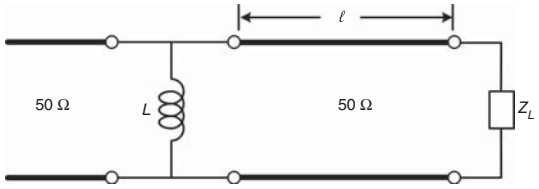


Figure P6.5.

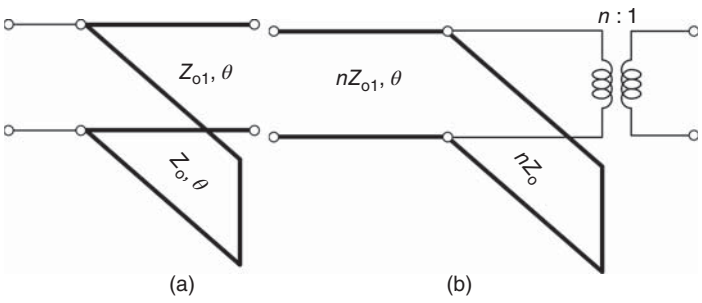


Figure P6.6.

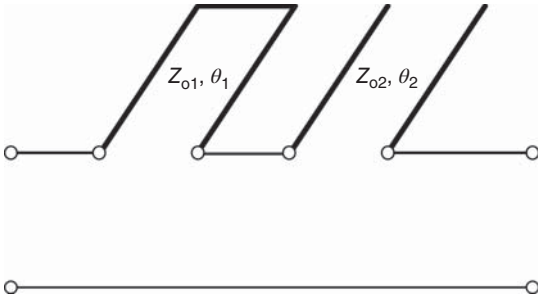


Figure P6.7.

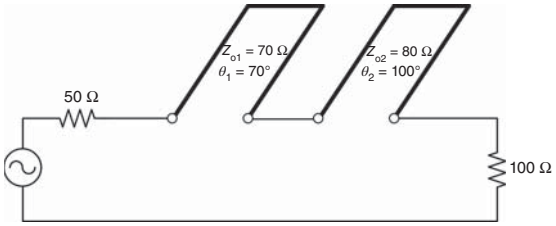


Figure P6.8.

impedance $Z_{oo}/2$ and electrical length θ_o. Due to the series transmission line, this equivalent network is very difficult to realize in practice. Assume the parallel-coupled transmission line is embedded in homogeneous medium (e.g., strip line) so that it has equal even- and odd-mode electrical length, derive an equivalent circuit of the two-port network using the Kuroda identities so that it can be easily realized.

6.24 Use the Kuroda identities to derive an equivalent network of the two-port transmission-line network consisting of series short- and open-circuited transmission lines as shown in Figure P6.7.

6.25 Use the Kuroda identities to transform the two-port transmission-line network consisting of two series short-circuited transmission lines as shown in Figure P6.8 to another two-port transmission-line network.

CHAPTER 7

SCATTERING PARAMETERS

Analysis, simulation, and measurement of radio frequency (RF) circuits, components and active devices are typically performed using the scattering (S) parameters. While the impedance and admittance matrix parameters are the basic and well-known parameters used for the design and analysis of general electrical circuits, their use is rather limited for RF circuit design due to the fact that open and short circuits, needed for the determination of these parameters, is very difficult, if not impossible, to achieve at high frequencies, particularly across wide frequency ranges, thereby hindering the use of the impedance and admittance matrices for RF circuit measurement. On the other hand, the S-parameters only require the use of matched terminations, which are easy to achieve even at high frequencies and across wide bandwidths. Therefore, the S-parameters are always used for the design of RF circuits and hence are perhaps the most important parameters in RF circuits. Knowledge of the S-parameters is absolutely essential for RF engineers. In this chapter, we will present the formulation and characteristics of the S-parameters as well as important parameters related to them. In order to demonstrate the usefulness of the S-parameters, we also briefly discuss the impedance and admittance parameters well-known to electrical engineers.

7.1 MULTIPORT NETWORKS

RF circuits, components, and systems, both passive and active, can have a single port or multiple ports. So in general, we can consider RF circuits, components, and systems are multiport networks. Figure 7.1 shows a general N-port network that can represent not only a multiport RF circuit, component, or system, but also multiport circuit elements, such as a junction between different transmission lines as shown in Figure 7.2, encountered in RF circuits, components, or systems. One particular remark needs to be mentioned here. In N-port circuits, there are N physical ports. From electromagnetic (EM) point of view, there are possibilities of multiple modes in a circuit even when there is only one RF signal incident to the circuit. Assuming an N-port circuit has only one mode, this circuit is a truly an N-port network. However, if a circuit with N physical ports has M modes, then it is represented electrically as a multiport network having $M \times N$ ports. In this MN-port network, there is a signal corresponding to the mth mode ($m = 1, 2, \ldots, M$) entering and leaving each (physical) port n ($n = 1, 2, \ldots, N$). Figure 7.3 illustrates an equivalent network of a three-port network having two modes.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

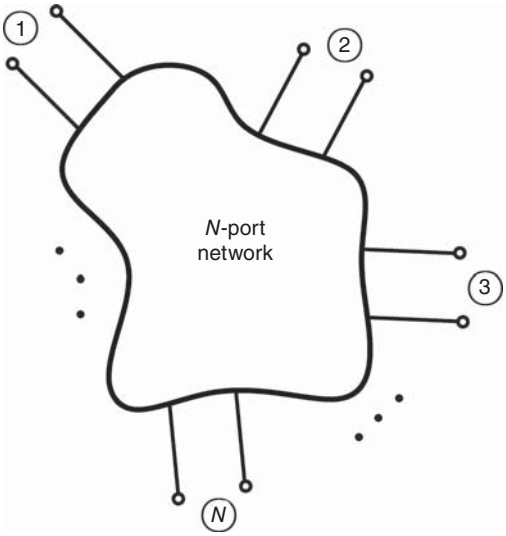


Figure 7.1. *N*-port network.

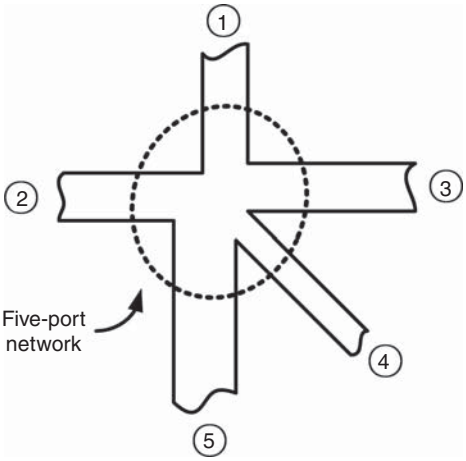


Figure 7.2. Five-port network formed by junction between transmissions.

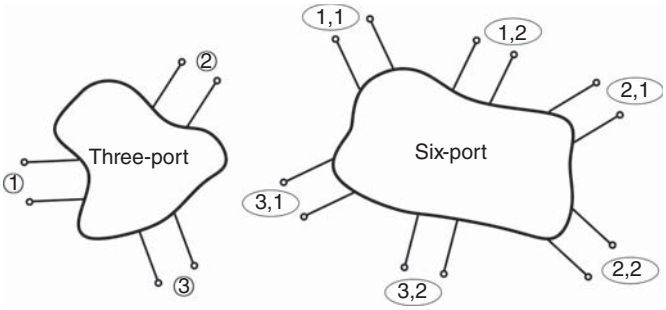


Figure 7.3. Three-port network (a) and its equivalent six-port network (b) assuming two different modes.

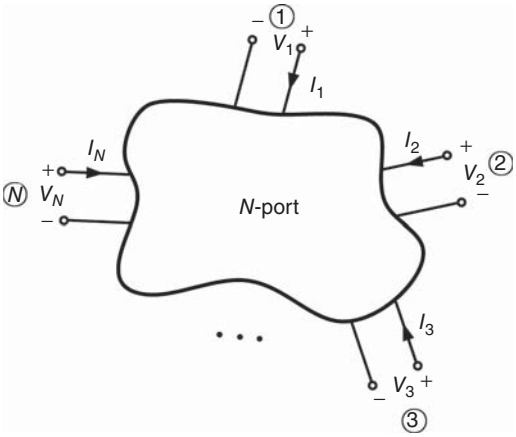


Figure 7.4. N-port network with voltage and current at each port.

An N-port network can be represented by an impedance matrix $[Z]$, an admittance matrix $[Y]$ as we learn from basic electrical circuit theory, and scattering matrix $[S]$. In the following, we will first review briefly the impedance and admittance matrices and point out the reason why these matrices are not very useful for RF circuit design. The formulation of the S-matrix useful for RF circuits is then given.

7.2 IMPEDANCE MATRIX

An N-port network, as shown in Figure 7.4, can be represented by an impedance matrix $[Z]$ defined by the following impedance-matrix equation:

$$\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1N} \\ Z_{21} & Z_{22} & \cdots & Z_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ Z_{N1} & Z_{N2} & \cdots & Z_{NN} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix} \tag{7.1}$$

where V_i and $I_i (i = 1, 2, \ldots, N)$ are the voltage and current at port i, respectively. The impedance parameters can be derived from (7.1) as

$$\begin{aligned} Z_{ii} &= \left.\frac{V_i}{I_i}\right|_{I_k\,(k\neq i)=0} \\ Z_{ij} &= \left.\frac{V_i}{I_j}\right|_{I_k\,(k\neq i,j)=0} \end{aligned} \tag{7.2}$$

For reciprocal networks, $Z_{ij} = Z_{ji}$, while for symmetrical networks, $Z_{ii} = Z_{jj}$, and for lossless network, Z_{ij} is purely imaginary.

The impedance matrix is useful for the analysis of networks connected in series. However, as (7.2) implies, an open circuit is required for determining its parameters. While this is not a problem in theoretical analysis, it is very difficult, if not impossible, in practice to achieve an open circuit at high frequencies, especially over a wide frequency range. As an example, we consider a transmission line open at one end. There is always a fringing capacitance C (to a ground) at the open end, which produces impedance $Z = 1/j\omega C$. While Z approaches infinity at DC, which represents a perfect open, it is finite at RF and varies across a bandwidth, thus making it impossible to achieve a truly open circuit. Therefore, it is very difficult to measure accurately

the impedance parameters of RF circuits, and hence the impedance matrix is not used for measurement of RF circuits. Another problem with open circuit is that active devices such as transistors may not be stable when its terminals are open-circuited, preventing the use of impedance matrix for both theoretical analysis and practical measurement of these devices.

7.3 ADMITTANCE MATRIX

The admittance matrix $[Y]$ of an N-port network as shown in Figure 7.4 is described in the following admittance-matrix equation:

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1N} \\ Y_{21} & Y_{22} & \cdots & Y_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{N1} & Y_{N2} & \cdots & Y_{NN} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix} \tag{7.3}$$

where V_i and $I_i (i = 1, 2, \ldots, N)$ are again the voltage and current at port i, respectively. The admittance parameters can be derived from (7.3) as

$$\begin{aligned} Y_{ii} &= \left.\frac{I_i}{V_i}\right|_{V_k(k\neq i)=0} \\ Y_{ij} &= \left.\frac{I_i}{V_j}\right|_{V_k(k\neq i,j)=0} \end{aligned} \tag{7.4}$$

For reciprocal networks, $Y_{ij} = Y_{ji}$, while for symmetrical networks, $Y_{ii} = Y_{jj}$, and for lossless network, Y_{ij} is purely imaginary. The admittance matrix is related to the impedance matrix by $[Y] = [Z]^{-1}$.

The admittance matrix is useful for the analysis of networks connected in shunt. However, as (7.4) suggests, a short circuit is required for the determination of the admittance parameters. While this is not a problem in theoretical analysis, it is very difficult, if not impossible, in practice to achieve a short circuit at high frequencies, especially over a wide frequency range. As an example, we consider a transmission line short-circuited at one end. It is not possible to connect the transmission-line end directly to a ground to create a perfect short at that end. An interconnect such as a via-hole is typically used for connecting to a ground. This interconnect, in the simplest case, behaves as an inductor with inductance L, resulting in impedance $Z = j\omega L$. While Z approaches zero at DC, thereby representing a perfect short, it is different from zero at RF and varies across a bandwidth, thus making it impossible to achieve a perfect short circuit. As a result, it is very difficult to measure accurately the admittance parameters of RF circuits; consequently, the admittance matrix is not useful for RF circuit measurement. Another problem with open circuit is that active devices may not be stable when its terminals are open-circuited, preventing the use of admittance matrix for both theoretical analysis and practical measurement of these devices.

7.4 IMPEDANCE AND ADMITTANCE MATRIX IN RF CIRCUIT ANALYSIS

As mentioned earlier, although the impedance and admittance matrices are not employed for RF circuit measurement, they are useful for the analysis and design of certain RF circuits. In these particular circuits, the impedance and admittance parameters are often determined directly from the impedance and admittance matrix equations, or indirectly via conversions from other parameters such as the S-parameters obtained through calculations or measurements. In the following, we discuss two important cases that employ T- and π-network to represent any two-port RF circuits utilizing the impedance and admittance matrices. These formulations serve as the basis for extension to other equivalent circuits different from the T- and π-network for RF circuits consisting of two or more ports.

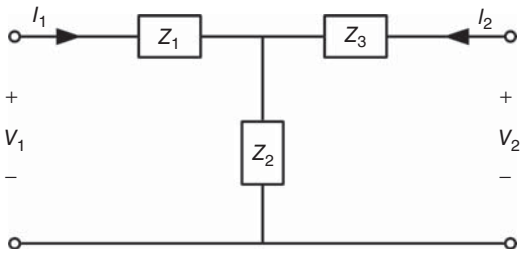


Figure 7.5. T-network.

7.4.1 T-Network Representation of Two-Port RF Circuits

We consider a particular T-network as shown in Figure 7.5 and a general two-port network. The voltages at ports 1 and 2 of the T-network are obtained as

$$\begin{aligned} V_1 &= (Z_1 + Z_2)I_1 + Z_2 I_2 \\ V_2 &= Z_2 I_1 + (Z_2 + Z_3)I_2 \end{aligned} \tag{7.5}$$

while those for a general two-port network are given using the impedance parameters by

$$\begin{aligned} V_1 &= Z_{11} I_1 + Z_{12} I_2 \\ V_2 &= Z_{21} I_1 + Z_{22} I_2 \end{aligned} \tag{7.6}$$

Comparing (7.5) with (7.6) gives the impedance parameters of the two-port network as

$$\begin{aligned} Z_{11} &= Z_1 + Z_2 \\ Z_{12} &= Z_{21} = Z_2 \\ Z_{22} &= Z_2 + Z_3 \end{aligned} \tag{7.7}$$

By solving (7.7), we can obtain the T-network elements in terms of the impedance parameters of the general two-port network as

$$\begin{aligned} Z_1 &= Z_{11} - Z_{12} \\ Z_2 &= Z_{12} \\ Z_3 &= Z_{22} - Z_{12} \end{aligned} \tag{7.8}$$

Equation (7.8) shows that any two-port network represented by an impedance matrix $[Z]$ can be modeled as a T-network as shown in Figure 7.6.

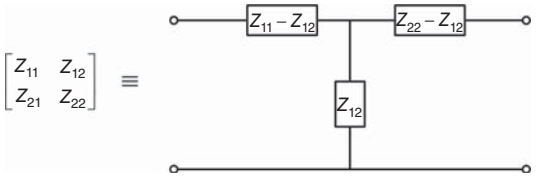


Figure 7.6. Representation of a two-port network characterized by an impedance matrix $[Z]$ by a T-network.

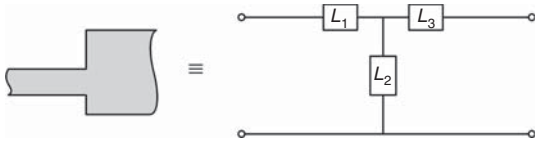


Figure 7.7. Transmission-line step and its equivalent circuit.

Equation (7.7) can be used to determine the impedance parameters of any T-network. For instance, we can obtain the impedance parameters of a step on transmission line, or a discontinuity between two transmission lines, modeled by an equivalent circuit as shown in Figure 7.7 as

$$\begin{aligned} Z_{11} &= j\omega(L_1 + L_2) \\ Z_{12} &= Z_{21} = j\omega L_2 \\ Z_{22} &= j\omega(L_2 + L_3) \end{aligned} \tag{7.9}$$

Now we consider a lossless transmission line having characteristic impedance Z_o and length ℓ, and assume that it is electrically equivalent to a T-circuit as shown in Figure 7.8. We proceed with the modeling or synthesis of the transmission line by terminating the transmission line and its equivalent circuit with a load impedance Z_L as shown in Figure 7.9. The impedances looking into the input ports of Figure 7.9(a) and (b) are obtained as

$$Z_{\text{in}} = Z_o \frac{Z_L + jZ_o \tan \beta\ell}{Z_o + jZ_L \tan \beta\ell} \tag{7.10}$$

and

$$Z_{\text{in},e} = \frac{Z_L + Z_1 \frac{Z_1+2Z_2}{Z_1+Z_2}}{1 + \frac{Z_L}{Z_1+Z_2}} \tag{7.11}$$

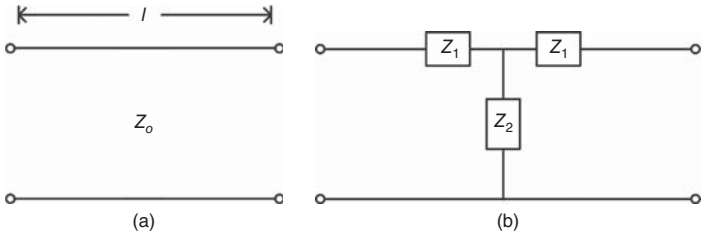


Figure 7.8. Lossless transmission line (a) and its assumed equivalent circuit (b).

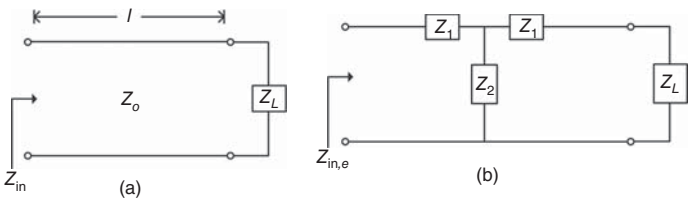


Figure 7.9. Lossless transmission line (a) and its assumed equivalent circuit (b) terminated with a load impedance.

respectively. Enforcing the equivalence between these networks, we get $Z_{\text{in}} = Z_{\text{in},e}$ which is satisfied if we let:

$$Z_1 \frac{Z_1 + 2Z_2}{Z_1 + Z_2} = jZ_o \tan \beta\ell$$
$$\frac{1}{Z_1 + Z_2} = j\frac{\tan \beta\ell}{Z_o} \tag{7.12}$$

which leads to

$$Z_1 = jZ_o \tan \frac{\beta\ell}{2}$$
$$Z_2 = -jZ_o \csc \beta\ell \tag{7.13}$$

Substituting (7.13) into (7.7) then gives the impedance parameters of the lossless transmission line as

$$Z_{11} = Z_{22} = -jZ_o \cot \beta\ell$$
$$Z_{12} = Z_{21} = -jZ_o \csc \beta\ell \tag{7.14}$$

In the foregoing transmission line formulation, we terminate the transmission line and its equivalent circuit with a load impedance Z_L which could be of any value to derive the transmission line's equivalent T-network. We now generalize the procedure by extending the concept of impedance termination to derive the T-equivalent circuit for any two-port network. To that end, we represent a general two-port network, characterized by $[Z]$ or $[Y]$, by a T-equivalent circuit as shown in Figure 7.10. We define Z_{1O}, Z_{2O} and Z_{1S}, Z_{2S} as the input impedance at port 1, 2 with port 2, 1 open and short circuited, respectively. From Figure 7.10(b), we can write

$$Z_{1O} = Z_1 + Z_2$$
$$Z_{2O} = Z_2 + Z_3$$
$$Z_{1S} = Z_1 + \frac{Z_2 Z_3}{Z_2 + Z_3}$$
$$Z_{2S} = Z_3 + \frac{Z_1 Z_2}{Z_1 + Z_2} \tag{7.15}$$

from which, we can derive

$$Z_1 = Z_{1O} - \sqrt{Z_{2O}(Z_{1O} - Z_{1S})}$$
$$Z_3 = Z_{2O} - \sqrt{Z_{2O}(Z_{1O} - Z_{1S})}$$
$$Z_2 = \sqrt{Z_{2O}(Z_{1O} - Z_{1S})} \tag{7.16}$$

which leads to a conclusion that we can model any two-port RF circuit by a T-equivalent circuit whose elements can be determined from the input impedances at ports 1 and 2 of the circuit when ports 2 and 1 are open

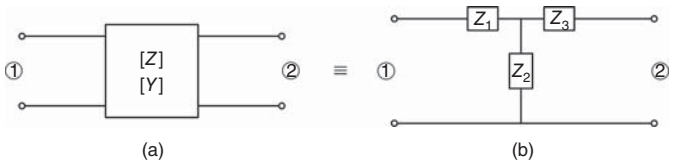


Figure 7.10. Two-port network (a) and its T-equivalent circuit (b).

or short, respectively. These input impedances can be determined from the impedance or admittance matrix of the two-port circuits. This result is useful for RF circuit modeling and can be extended to the modeling of multiport RF networks. It is noted that the presented formulation employs a T-network as the equivalent circuit for a general two-port circuit just to illustrate how an RF circuit can be modeled or synthesized using impedance parameters. Other kinds of equivalent circuits, such as a π-network, can also be used as stated earlier.

7.4.2 π-Network Representation of Two-Port RF Circuits

We now consider a π-network as shown in Figure 7.11. Similar to the formulation of the T-network, we can derive the admittance parameters of the π-network as

$$\begin{aligned} Y_{11} &= Y_1 + Y_2 \\ Y_{12} &= Y_{21} = -Y_2 \\ Y_{22} &= Y_2 + Y_3 \end{aligned} \tag{7.17}$$

Solving (7.17) gives the π-network elements in terms of the admittance parameters as

$$\begin{aligned} Y_1 &= Y_{11} + Y_{12} \\ Y_2 &= -Y_{12} \\ Y_3 &= Y_{22} + Y_{12} \end{aligned} \tag{7.18}$$

which demonstrate that any two-port network represented by an admittance matrix can be modeled as a π-network as shown in Figure 7.12.

Equation (7.17) can be used to obtain the admittance parameters of any π-network. Consider a gap on transmission line and its equivalent circuit as shown in Figure 7.13; we can write the admittance parameters of the gap upon using (7.17) as

$$\begin{aligned} Y_{11} &= Y_{22} = j\omega(C_1 + C_2) \\ Y_{12} &= Y_{21} = -j\omega C_1 \end{aligned} \tag{7.19}$$

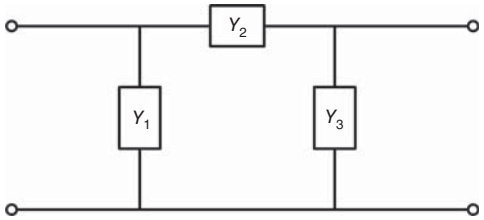


Figure 7.11. π-Network.

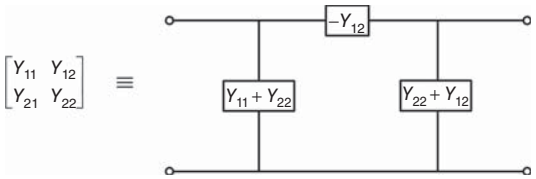


Figure 7.12. Representation of a two-port network characterized by an admittance matrix by a π-network.

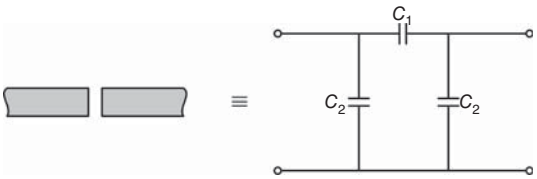


Figure 7.13. Transmission-line1 gap and its equivalent circuit.

Similar to the T-equivalent circuit of any two-port RF circuits discussed previously, we can model any two-port RF circuit by a π-equivalent circuit whose elements can be determined from the input admittances at ports 1 and 2 of the circuit when ports 2 and 1 are open or short, respectively. The input admittances can be determined from the admittance or impedance matrix of the two-port circuits. The result is useful for RF circuit modeling and can be extended to the modeling of multiport RF networks.

7.5 SCATTERING MATRIX

It is recognized that the impedance and admittance matrices are not very useful for the measurement of RF circuits, components and active devices due to the fact that the required open and short circuit are not possible at high frequencies, especially over wide operating frequency ranges. To resolve this issue, the scattering (S) matrix based on traveling waves [1] is typically employed in RF engineering.

7.5.1 Fundamentals of Scattering Matrix

For ease of understanding, the fundamentals of the S-matrix are first discussed for two-port networks and then extended to N-port networks.

7.5.1.1 Scattering Matrix for Two-Port Networks. We consider a two-port network inserted between transmission lines having the same characteristic impedance Z_o as shown in Figure 7.14.[1] We assume that Z_o is real; that is, the transmission lines are theoretically lossless. Lossless transmission lines are nonexistent, and so, in practice, low loss transmission lines need to be used at these ports in order for the results obtained here to remain accurate. We define the overall network as another two-port having ports at 1 and 2. This overall network is characterized by its impedance $[Z]$ or admittance $[Y]$ matrix. Using the admittance matrix, we can write:

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{7.20}$$

The voltage and current at each port each consists of two (wave) components traveling in opposite directions as

$$\begin{aligned} V_1 &= V_1^+ + V_1^- \\ V_2 &= V_2^+ + V_2^- \\ I_1 &= I_1^+ + I_1^- = \frac{1}{Z_o}(V_1^+ - V_1^-) \\ V_2 &= I_2^+ + I_2^- = \frac{1}{Z_o}(V_2^+ - V_2^-) \end{aligned} \tag{7.21}$$

[1] Adding a transmission line at each port of an RF circuit is normally done in practice and is in fact needed for measurement or connecting with other circuits.

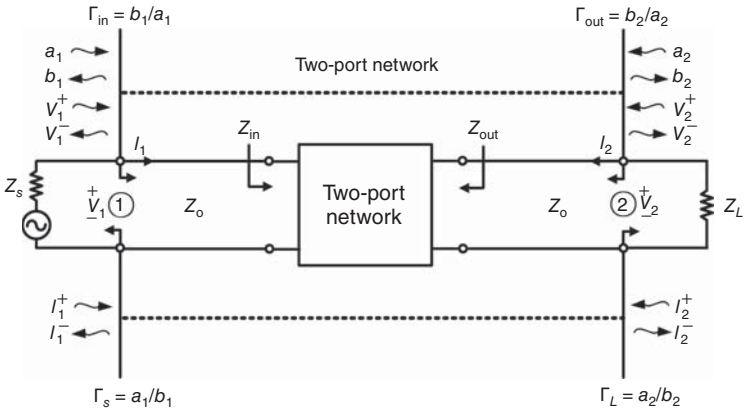


Figure 7.14. Two-port network with ports 1 and 2.

where the characteristic impedance Z_o of the transmission lines at ports 1 and 2 is given as $Z_o = \frac{V_1^+}{I_1^+} = -\frac{V_1^-}{I_1^-} = \frac{V_2^+}{I_2^+} = -\frac{V_2^-}{I_2^-}$. Substituting (7.21) into (7.20) and solving for V_1^- and V_2^- in terms of V_1^+ and V_2^+ yields

$$\begin{aligned} V_1^- &= S_{11}V_1^+ + S_{12}V_2^+ \\ V_2^- &= S_{21}V_1^+ + S_{22}V_2^+ \end{aligned} \tag{7.22}$$

where S_{11}, S_{12}, S_{21}, and S_{22} are functions of $Y_{ij}(i, j = 1, 2)$. Note that S_{11}, S_{12}, S_{21}, and S_{22} can also be obtained in terms of the impedance parameters Z_{ij} if the impedance matrix equation was used in (7.20). As can be seen, the parameters S_{ij} relate the scattered or reflected waves (V_1^-, V_2^-) to the incident waves (V_1^+, V_2^+), instead of the total voltages (V_1, V_2) and currents (I_1, I_2), at the two ports; they are thus called "scattering parameters."

We now divide (7.22) by $\sqrt{Z_o}$ to obtain

$$\begin{aligned} \frac{V_1^-}{\sqrt{Z_o}} &= S_{11}\frac{V_1^+}{\sqrt{Z_o}} + S_{12}\frac{V_2^+}{\sqrt{Z_o}} \\ \frac{V_2^-}{\sqrt{Z_o}} &= S_{21}\frac{V_1^+}{\sqrt{Z_o}} + S_{22}\frac{V_2^+}{\sqrt{Z_o}} \end{aligned} \tag{7.23}$$

which can be rewritten as

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_1 \end{aligned} \tag{7.24}$$

upon defining

$$\begin{aligned} a_1 &\triangleq \frac{V_1^+}{\sqrt{Z_o}} \quad a_2 \triangleq \frac{V_2^+}{\sqrt{Z_o}} \\ b_1 &\triangleq \frac{V_1^-}{\sqrt{Z_o}} \quad b_2 \triangleq \frac{V_2^-}{\sqrt{Z_o}} \end{aligned} \tag{7.25}$$

where a_i an $b_i (i = 1, 2)$ are the (normalized) incident and reflected voltages at port i, respectively, as designated in Figure 7.14. Equation (7.24) is the standard S-matrix equation used in RF engineering.

The S-parameters can be obtained from (7.24) as

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0}$$
$$S_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0}$$
$$S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0}$$
$$S_{22} = \left.\frac{b_2}{a_2}\right|_{a_1=0} \tag{7.26}$$

where $a_i = 0$ $(i = 1, 2)$ indicates that there is no signal entering port i, implying that port i is perfectly matched. It is apparent from (7.26) that S_{11} represents the reflection coefficient at port 1 when port 2 is terminated with a matched load; S_{12} represents the transmission coefficient from port 2 to port 1 (or the backward transmission coefficient, isolation or reverse voltage gain depending on the RF circuit being considered) when port 1 is terminated with a matched load, S_{21} represents the transmission coefficient from port 1 to port 2 (or the forward transmission coefficient), also commonly known as the insertion loss (IL) (for passive circuits such as a lumped-element low pass filter) or voltage gain (for active circuits like an amplifier) when port 2 is terminated with a matched load; and S_{22} represents the reflection coefficient at port 2 when port 1 is terminated with a matched load. Examination of (7.24) shows that it indeed describes precisely the signal behavior at each port. Each equation in (7.24) shows that the signal leaving one port is contributed by the signal reflected at that port and the signal transmitted from the other port.

We now see that the S-parameters can be measured by terminating an appropriate port with a matched load equal to the characteristic impedance of the transmission line at that port (typically 50 Ω). This termination is basically a resistive load and hence is easily realized in the RF range even across extremely wide frequency ranges. Additionally, considering active devices, terminating the terminals of these devices directly with resistive loads (for instance in device measurement) or indirectly through some circuit elements in active circuits (such as amplifier) minimize the possibility of unwanted oscillations in active devices or circuits, hence making measurement possible, accurate and reliable. Moreover, the S-parameters are defined based on traveling waves on lossless transmission lines, which have constant amplitude due to the lossless criterion. This is in contrast with the RF voltages and currents, whose amplitudes vary along the lines, used in the impedance and admittance matrices. This characteristic effectively facilitates the measurement of RF circuits since it can be performed by an appropriate instrument such as a vector network analyzer at locations away from the circuits, provided that low loss transmission lines (ideally lossless transmission lines) are used in between. Altogether, we can see that accurate S-parameters at RF can be measured easily, accurately and reliably. Typically, the measurement of the S-parameters is performed using a vector network analyzer. Examination of (7.26) reveals that the input and output ports of a two-port network does not need to be impedance-matched to the connecting transmission line; that is, the network's port impedances Z_{in} and Z_{out} do not need to be equal to Z_o. As long as the terminating impedances Z_S and Z_L are equal to Z_o, the incident voltages a_1 and a_2 would be equal to zero as required, which is completely independent with the network's port impedances. A remark needs to be made at this point that the S-parameters characterize the properties of a network itself regardless of the surrounding conditions, even though that they are defined when all ports are matched.

We consider again the two-port network as shown in Figure 7.14. The reflection coefficients at ports 1 and 2 are given, according to Eqs. (11.7) and (11.8) in Chapter 11, by

$$\Gamma_{\text{in}} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \tag{7.27}$$

and

$$\Gamma_{\text{out}} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \tag{7.28}$$

When $Z_L = Z_o$, port 2 is perfectly matched and $\Gamma_L = 0$, and accordingly Γ_{in} becomes

$$\Gamma_{\text{in}} = S_{11} \tag{7.29}$$

Similarly, Γ_{out} reduces to

$$\Gamma_{\text{out}} = S_{22} \tag{7.30}$$

when port 1 is perfectly matched. These results are of course expected according to (7.26). On the other hand, when $Z_L \neq Z_o$ and $Z_S \neq Z_o$, or equivalently, ports 2 and 1 are not matched, $\Gamma_{\text{in}} \neq S_{11}$ and $\Gamma_{\text{out}} \neq S_{22}$, respectively.

Taking the square of the magnitude of the incident and reflected voltages at ports 1 and 2 gives

$$\begin{aligned} |a_1^2| &= \frac{|V_1^+|^2}{Z_o} \quad |a_2^2| = \frac{|V_2^+|^2}{Z_o} \\ |b_1^2| &= \frac{|V_1^-|^2}{Z_o} \quad |b_2^2| = \frac{|V_2^-|^2}{Z_o} \end{aligned} \tag{7.31}$$

which represent the corresponding (peak) incident and reflected powers of the traveling power waves at ports 1 and 2. The corresponding *rms* powers are $\frac{1}{2}|a_i|^2$ and $\frac{1}{2}|b_i|^2$, where $i = 1, 2$. Furthermore, taking the square of the magnitude of the S-parameters in (7.26) and using the incident and reflected powers, we can obtain

$$\begin{aligned} |S_{11}|^2 &= \frac{|b_1|^2}{|a_1|^2} = \frac{\text{Reflected power at port 1}}{\text{Incident power at port 1}} = \text{Return loss at port 1} \\ |S_{12}|^2 &= \frac{|b_1|^2}{|a_2|^2} = \frac{\text{Reflected power at port 1}}{\text{Incident power at port 2}} \\ &= \begin{cases} \text{Insertion loss from port 2 to port 1 or isolation (for passive circuits)} \\ \text{Reverse power gain or isolation (for active circuits)} \end{cases} \\ |S_{21}|^2 &= \frac{|b_2|^2}{|a_1|^2} = \frac{\text{Reflected power at port 2}}{\text{Incident power at port 1}} \\ &= \begin{cases} \text{Insertion loss from port 1 to port 2 (for passive circuits)} \\ \text{Power gain (for active circuits)} \end{cases} \\ |S_{22}|^2 &= \frac{|b_2|^2}{|a_2|^2} = \frac{\text{Reflected power at port 2}}{\text{Incident power at port 2}} = \text{Return loss at port 2} \end{aligned} \tag{7.32}$$

These results are obtained under the matched conditions. They are important results characterizing the return loss (RL) and insertion loss (IL) (or gain), typically used in RF engineering, that we will discuss later.

7.5.1.2 Scattering Matrix for N-Port Network. We consider an N-port network as shown in Figure 7.15 and, similar to the previously consider two-port network in Figure 7.14, we assume that the terminating transmission line at each port is lossless and has equal (real) characteristic impedance of Z_o. The previous formulation for a two-port network can be extended to an N-port network to obtain the following S-matrix equations:

$$[b] = [S][a] \tag{7.33}$$

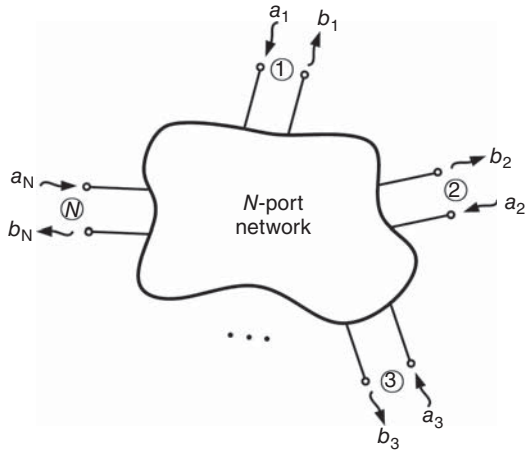


Figure 7.15. *N*-port network.

or

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ S_{N1} & S_{N2} & \cdots & S_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \tag{7.34}$$

where a's and b's are the normalized incident and reflected voltages, respectively, defined as

$$\begin{aligned} a_i|_{i=1 \text{ to } N} &\triangleq \frac{V_i^+}{\sqrt{Z_o}} \\ b_i|_{i=1 \text{ to } N} &\triangleq \frac{V_i^-}{\sqrt{Z_o}} \end{aligned} \tag{7.35}$$

with V_i^+ and V_i^- being the incident and reflected voltages at port i, respectively. The S-parameters of the N-port network are obtained from (7.34) as

$$\begin{aligned} S_{ii} &= \left.\frac{b_i}{a_i}\right|_{a_k=0(k\neq i)} \\ S_{ij} &= \left.\frac{b_i}{a_j}\right|_{a_k=0(k\neq j)} \end{aligned} \tag{7.36}$$

S_{ii} represents the reflection coefficient at port i when all ports different from i are terminated with a matched load, and S_{ij} represents the transmission coefficient from port j to port i when all ports except j are terminated with a matched load. The S-parameters of an N-port network represent the entire network itself and are defined when all ports are matched. That is, the S-parameters of a network remain the same even when the environment around it changes (i.e., when the load terminations at the network's ports are changed). For instance, we consider a three-port RF power divider characterized by its S-parameters and connect its ports to other RF components, then even if the power divider's ports are no longer matched (i.e., port terminations are changed from, say, 50 Ω), the S-parameters of the power divider remain the same. The only parameters that change are the transmission coefficient between the ports and the reflection coefficients at the ports of the power divider, which depend on the ports' terminating impedances. It is recalled that these coefficients are only equal to the corresponding S-parameters under the matched condition. It is apparent from (7.36)

that the S-parameters of multiport RF circuits can be measured by terminating all the unused ports with a matched termination. This matched termination is typically 50 Ω since a 50-Ω transmission line is typically used at each port.

The (peak) incident and reflected power at each port are $|a_i|^2$ and $|b_i|^2$, where $i = 1, 2, \ldots, N$, respectively. The corresponding *rms* powers are $\frac{1}{2}|a_i|^2$ and $\frac{1}{2}|b_i|^2$. We can also define the following power parameters obtained under the matched conditions:

$$
\begin{aligned}
|S_{ii}|^2 &= \frac{|b_i|^2}{|a_i|^2} = \frac{\text{Reflected power at port } i}{\text{Incident power at port } i} = \text{Return loss at port } i \\
|S_{ij}|^2 &= \frac{|b_i|^2}{|a_j|^2} = \frac{\text{Reflected power at port } i}{\text{Incident power at port } j} \\
&= \begin{cases} \text{Insertion loss from port } j \text{ to port } i \text{ or isolation} & \text{(for passive circuits)} \\ \text{Power gain or isolation} & \text{(for active circuits)} \end{cases}
\end{aligned} \tag{7.37}
$$

These parameters are widely used in RF engineering to describe the RL and IL (or gain) of RF circuits.

7.5.1.3 Scattering Matrix for Unequal Terminating Impedances. In the foregoing analysis, we assume that the characteristic impedance of the terminating transmission lines at each port is equal, which is typical in practice. For generalization, these results are now modified considering different terminating impedances at ports. We consider again an N-port network as shown in Figure 7.15 and assume that the lossless transmission line at each port has unequal (real) characteristic impedance; that is, $Z_{oi} \neq Z_{oj}$ for $i \neq j$ and $i, j = 1, 2, \ldots, N$, where Z_{oi} and Z_{oj} are the characteristic impedances of the transmission lines at ports i and j, respectively.

Following the previous formulation, we can derive the S-matrix equation as given in (7.34) directly from the impedance or admittance matrix equation, where the normalized incident and reflected voltages at port i are now defined by

$$
\begin{aligned}
a_i|_{i=1 \text{ to } N} &\triangleq \frac{V_i^+}{\sqrt{Z_{oi}}} \\
b_i|_{i=1 \text{ to } N} &\triangleq \frac{V_i^-}{\sqrt{Z_{oi}}}
\end{aligned} \tag{7.38}
$$

The S-parameters can still be obtained from (7.34), making use of (7.38), as

$$
\begin{aligned}
S_{ii} &= \left.\frac{b_i}{a_i}\right|_{a_k=0(k\neq i)} = \left.\frac{V_i^-}{V_i^+}\right|_{V_k^+=0(k\neq i)} = S_{ii}^e \\
S_{ij} &= \left.\frac{b_i}{a_j}\right|_{a_k=0(k\neq j)} = \sqrt{\frac{Z_{oj}}{Z_{oi}}}\left.\frac{V_i^-}{V_j^+}\right|_{V_k^+=0(k\neq j)} = \sqrt{\frac{Z_{oj}}{Z_{oi}}}S_{ij}^e
\end{aligned} \tag{7.39}
$$

where S_{ii}^e and S_{ij}^e represent the S-parameters when the network has the same terminating characteristic impedance. Equation (7.39) shows that reflection S-parameter S_{ii} is the same for N-port networks having equal and unequal terminating characteristic impedances, while the transmission S-parameter S_{ij} for an N-port network with unequal terminating characteristic impedances is equal to S_{ij} of the same network terminated with equal characteristic impedances scaled by a factor of $\sqrt{Z_{oj}/Z_{oi}}$. These results are useful for determining the S-parameters of N-port networks with arbitrary terminating characteristic impedances.

7.5.1.4 Properties of Scattering Parameters. We consider again an N-port network as shown in Figure 7.15 and discuss the properties of the S-parameters for two cases: reciprocal and lossless.

Reciprocal Network

A reciprocal network is defined as having identical transmission between one port to another port and vice versa. That is, in terms of S-parameters,

$$S_{ij} = S_{ji} \tag{7.40}$$

where $i, j = 1, 2, \ldots, N$. Applying (7.40) to the S-matrix in (7.34) gives

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ S_{N1} & S_{N2} & \cdots & S_{NN} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{21} & \cdots & S_{N1} \\ S_{12} & S_{22} & \cdots & S_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ S_{1N} & S_{2N} & \cdots & S_{NN} \end{bmatrix} \tag{7.41}$$

which proves that the S-matrix $[S]$ of a reciprocal network is equal to its transpose matrix $[S]^T$ as

$$[S] = [S]^T \tag{7.42}$$

Equation (7.42) implies that the S-matrix of a reciprocal network is always symmetrical. For commonly used two-port networks, we have for reciprocity:

$$S_{12} = S_{21} \tag{7.43}$$

Lossless Network

As mentioned at various places in this book, the term "lossless" is defined loosely; it merely implies that no resistive elements are used in circuits. In a lossless network, the power indeed does not dissipate, and hence the total power entering the network must be equal to the total power leaving the network according to the principle of energy conservation. Therefore, for a lossless network, we obtain

$$\sum_{i=1}^{N} |a_i|^2 = \sum_{i=1}^{N} |b_i|^2 \tag{7.44}$$

Equation (7.45) can be rewritten as

$$\sum_{i=1}^{N} a_i a_i^* = \sum_{i=1}^{N} b_i b_i^* \tag{7.45}$$

where * stands for the conjugate of a complex parameter. Equation (7.45) is equivalent to

$$[a]^T [a]^* = [b]^T [b]^* \tag{7.46}$$

Substituting (7.33) into (7.37) gives

$$[a]^T [a]^* = [a]^T [S]^T [S]^* [a]^* \tag{7.47}$$

Since $[a] \neq [0]$, (7.47) can be simplified to

$$[S]^T [S]^* = [I] \tag{7.48}$$

or

$$[S]^* = \{[S]^T\}^{-1} \tag{7.49}$$

where $[I]$ is the identity matrix. Equation (7.49) indicates that the S-matrix of a lossless network is a unitary matrix.

Equation (7.48) can be written in details as

$$\begin{bmatrix} S_{11} & S_{21} & \cdots & S_{N1} \\ S_{12} & S_{22} & \cdots & S_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ S_{1N} & S_{2N} & \cdots & S_{NN} \end{bmatrix} \begin{bmatrix} S_{11}^* & S_{12}^* & \cdots & S_{1N}^* \\ S_{21}^* & S_{22}^* & \cdots & S_{2N}^* \\ \vdots & \vdots & \cdots & \vdots \\ S_{N1}^* & S_{N2}^* & \cdots & S_{NN}^* \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{7.50}$$

which, after carrying out the matrix multiplication, gives

$$\begin{gathered} S_{11}S_{11}^* + S_{21}S_{21}^* + \cdots + S_{N1}S_{N1}^* = 1 \\ S_{12}S_{12}^* + S_{22}S_{22}^* + \cdots + S_{N2}S_{N2}^* = 1 \\ \cdots \\ S_{1N}S_{1N}^* + S_{2N}S_{2N}^* + \cdots + S_{NN}S_{NN}^* = 1 \\ S_{11}S_{12}^* + S_{21}S_{22}^* + \cdots + S_{N1}S_{N2}^* = 0 \\ S_{12}S_{11}^* + S_{22}S_{21}^* + \cdots + S_{N2}S_{N1}^* = 0 \\ \cdots \\ S_{1N}S_{1,N-1}^* + S_{2N}S_{2,N-1}^* + \cdots + S_{NN}S_{N,N-1}^* = 0 \end{gathered} \tag{7.51}$$

Equation (7.51) can be rewritten as

$$\begin{aligned} \sum_{n=1}^{N} |S_{ni}|^2 &= \sum_{n=1}^{N} S_{ni}S_{ni}^* = 1; \quad i = 1, 2, \ldots, N \\ \sum_{n=1}^{N} S_{ni}S_{nj}^* &= 0; \quad i, j = 1, 2, \ldots, N \text{ and } i \neq j \end{aligned} \tag{7.52}$$

For commonly used two-port networks, we have, assuming lossless,

$$\begin{gathered} |S_{11}|^2 + |S_{21}|^2 = 1 \\ |S_{22}|^2 + |S_{12}|^2 = 1 \\ S_{11}S_{12}^* + S_{21}S_{22}^* = 0 \\ S_{12}S_{11}^* + S_{22}S_{21}^* = 0 \end{gathered} \tag{7.53}$$

Applying the reciprocal property (7.43) and lossless condition (7.53), we obtain the following result for a lossless and reciprocal two-port network:

$$|S_{11}| = |S_{22}| \tag{7.54}$$

It is noted that practical RF circuits are lossy and so there is always power dissipated in RF circuits. In contrast to the lossless condition derived in (7.48), lossy N-port networks are subject to the following condition:

$$[S]^T[S]^* < [I] \tag{7.55}$$

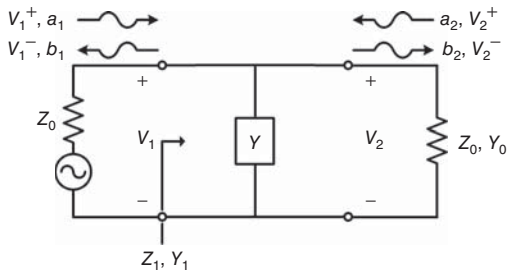


Figure 7.16. A shunt admittance.

7.5.2 Examples for Scattering Parameters

As the first example, we derive the S-parameters of an admittance Y connected in shunt as shown in Figure 7.16. The S-parameters S_{11} and S_{21} can be obtained from (7.26) as

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0} = \left.\frac{V_1^-}{V_1^+}\right|_{V_2^+=0} = \frac{Z_1 - Z_o}{Z_1 + Z_o} = \frac{Y_o - Y_1}{Y_o + Y_1} \tag{7.56}$$

or

$$S_{11} = \frac{Y}{Y + 2Y_o} \tag{7.57}$$

$$S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0} = \left.\frac{V_2^-}{V_1^+}\right|_{V_2^+=0} = \left.\frac{V_1^- + V_1^+}{V_1^+}\right|_{V_2^+=0} = S_{11} + 1 \tag{7.58}$$

making use of $V_1 = V_1^+ + V_1^- = V_2 = V_2^+ + V_2^-$ for the shunt element. Substituting (7.57) into (7.58) then gives

$$S_{21} = \frac{2Y_o}{Y + 2Y_o} \tag{7.59}$$

Since the network is symmetrical and reciprocal, we obtain $S_{22} = S_{11}$ and $S_{12} = S_{21}$.

In another example, we consider a lossless transmission line as shown in Figure 7.17. We can write based on traveling waves on transmission lines:

$$V_2^- = V_1^+ e^{-j\beta\ell} \tag{7.60}$$

which is equivalent to

$$b_2 = a_1 e^{-j\beta\ell} \tag{7.61}$$

and

$$b_1 = a_2 e^{-j\beta\ell} \tag{7.62}$$

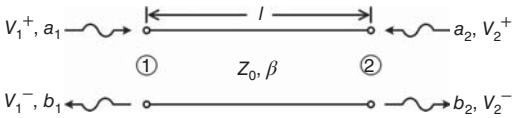


Figure 7.17. A transmission line.

Combine (7.61) and (7.62) yields the following S-matrix equation:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0 & e^{-j\beta\ell} \\ e^{-j\beta\ell} & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.63}$$

The result of $S_{21} = S_{12}$ indicates that the network is reciprocal as expected.

7.5.3 Effect of Reference-Plane Change on Scattering Matrix

While simulations can be performed right at the ports of RF circuits, components or active devices such as transistors, measurements of individual circuits, components or devices, or connections between them, often cannot be done directly at individual ports. In practice, each port of RF circuits, components or devices is typically connected to a transmission line (or in general an interface element) for interface purposes – for instance, transmission lines are used in a discrete RF band-pass filter for connecting to connectors, or on-wafer RF pads are used with a radio frequency integrated circuit (RFIC) or transistor for interfacing with on-wafer measurement instruments. As a result, the actual measurement or reference planes are different from those of the interested RF device under test (DUT), and hence the measured S-parameters are not the same as those of the DUT. In this section, we present a simple formulation to derive the S-parameters of a DUT from the measured S-parameters. This process is considered one of the "de-embedding" processes discussed in Chapter 15.

For ease in illustration purpose, we consider a two-port RF circuit (DUT) connected as shown in Figure 7.18. Let $[S]$ be the S-matrix of the DUT between ports 1 and 2, and $[S']$ be the measured S-matrix between ports $1'$ and $2'$, as noted in Figure 7.18. We can write

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{7.64}$$

between ports 1 and 2, and

$$\begin{bmatrix} b'_1 \\ b'_2 \end{bmatrix} = \begin{bmatrix} S'_{11} & S'_{12} \\ S'_{21} & S'_{22} \end{bmatrix} \begin{bmatrix} a'_1 \\ a'_2 \end{bmatrix} \tag{7.65}$$

between ports $1'$ and $2'$. Assume the transmission lines are lossless, we can write based on waves propagating on lossless transmission lines:

$$b'_1 = b_1 e^{-j\beta\ell_1} = b_1 e^{-j\theta_1} \tag{7.66}$$

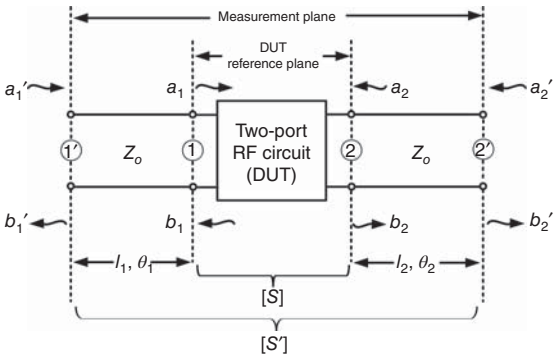


Figure 7.18. A two-port RF circuit with transmission line at each port.

where $\theta_1 = \beta\ell_1$ is the electrical length of the transmission line at port 1 with β being the phase constant. Taking the inverse of (7.66) and applying similar process for other voltage waves, we obtain

$$\begin{aligned} b_1 &= b_1' e^{j\theta_1} \\ b_2 &= b_2' e^{j\theta_2} \\ a_1 &= a_1' e^{-j\theta_1} \\ a_2 &= a_2' e^{-j\theta_2} \end{aligned} \tag{7.67}$$

where $\theta_2 = \beta\ell_2$ is the electrical length of the transmission line at port 2. Substituting (7.67) into (7.64) results in

$$\begin{bmatrix} b_1' \\ b_2' \end{bmatrix} = \begin{bmatrix} S_{11}e^{-j2\theta_1} & S_{12}e^{-j(\theta_1+\theta_2)} \\ S_{21}e^{-j(\theta_1+\theta_2)} & S_{22}e^{-j2\theta_2} \end{bmatrix} \begin{bmatrix} a_1' \\ a_2' \end{bmatrix} \tag{7.68}$$

Comparing (7.65) and (7.68) gives

$$\begin{aligned} \begin{bmatrix} S_{11}' & S_{12}' \\ S_{21}' & S_{22}' \end{bmatrix} &= \begin{bmatrix} S_{11}e^{-j2\theta_1} & S_{12}e^{-j(\theta_1+\theta_2)} \\ S_{21}e^{-j(\theta_1+\theta_2)} & S_{22}e^{-j2\theta_2} \end{bmatrix} \\ &= \begin{bmatrix} e^{-j\theta_1} & 0 \\ 0 & e^{-j\theta_2} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} e^{-j\theta_1} & 0 \\ 0 & e^{-j\theta_2} \end{bmatrix} \end{aligned} \tag{7.69}$$

or

$$[S'] = [\theta][S][\theta] \tag{7.70}$$

where

$$[\theta] = \begin{bmatrix} e^{-j\theta_1} & 0 \\ 0 & e^{-j\theta_2} \end{bmatrix} \tag{7.71}$$

Dividing (7.70) by $[\theta]$ twice leads to

$$[S] = [\theta]^{-1}[S'][\theta]^{-1} \tag{7.72}$$

which is equivalent to

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} S_{11}'e^{j2\theta_1} & S_{12}'e^{j(\theta_1+\theta_2)} \\ S_{21}'e^{j(\theta_1+\theta_2)} & S_{22}'e^{j2\theta_2} \end{bmatrix} \tag{7.73}$$

Equation (7.72) or (7.73) allows the actual S-parameters of a two-port DUT to be determined from those obtained through measurement.

The forgoing results can be extended for an N-port RF circuit as shown in Figure 7.19 to obtain the same relations in (7.70) and (7.72) between the S-parameters of an N-port RF circuit (DUT) and those measured at the measurement planes, where

$$[\theta] = \begin{bmatrix} e^{-j\theta_1} & 0 & \cdots & 0 \\ 0 & e^{-j\theta_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & e^{-j\theta_N} \end{bmatrix} \tag{7.74}$$

with $\theta_n = \beta\ell_n (n = 1, 2, \ldots, N)$ being the electrical length of the transmission line at port n.

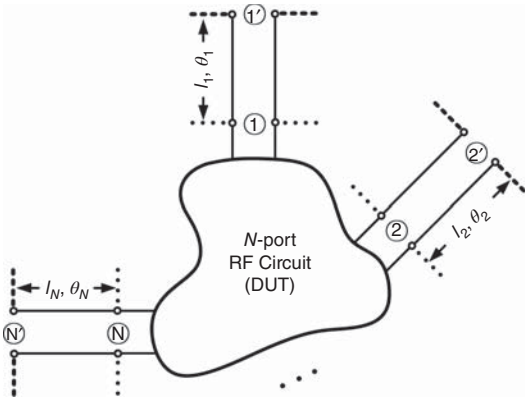


Figure 7.19. An N-port RF circuit with transmission line at each port. The dotted lines are the DUT reference planes corresponding to $[S]$ and the dashed lines are the measurement planes corresponding to $[S']$.

7.5.4 Return Loss, Insertion Loss, and Gain

RL, IL, and gain of RF circuits, indicated in (7.32) and (7.33), relate directly to the S-parameters and are typically used for characterizing RF circuits. For ease in illustrating the concept, we consider a general two-port RF network as shown in Figure 7.20 and define T_{21} and T_{12} as the (general) transmission coefficients between ports 1 and 2 and Γ_1 and Γ_2 as the (general) reflection coefficients looking into ports 1 and 2, respectively. As discussed before, these parameters are only equal to the corresponding S-parameters under the matched termination conditions. For instance, $S_{11} = \Gamma_1$ and $S_{21} = T_{21}$ when port 2 is perfectly matched.

7.5.4.1 Return Loss. Reflection is inevitable in practical RF circuits and this causes loss to the incident signal. This kind of loss is identified as the mismatch or reflection loss. However, in RF engineering, we do not use reflection loss to characterize the matching condition of RF circuits. Instead, we employ another parameter related to the reflection loss. This parameter is known as the RL and defined as

$$\mathrm{RL}_1 \triangleq 10 \log \frac{P_r}{P_{\mathrm{in}}} \quad \text{(dB)} \tag{7.75}$$

at port 1, which is equivalent to

$$\mathrm{RL}_1 = 10 \log |\Gamma_1|^2 \quad \text{(dB)} \tag{7.76}$$

Under the matched termination at port 2, the RL at port 1 becomes

$$\mathrm{RL}_1 = 10 \log |S_{11}|^2 \quad \text{(dB)} \tag{7.77}$$

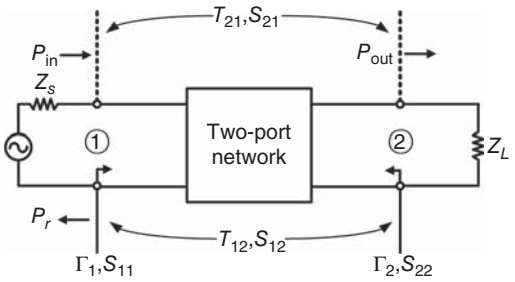


Figure 7.20. Two-port RF network. P_{in}, P_r, and P_{out} are the incident, reflected and transmitted power, respectively.

A remark needs to be made concerning the RLs given in (7.76) and (7.77). These RLs are in general different. The RL given by $10\log|S_{11}|^2$ in (7.77) is the RL under the matched condition and represents the RL (at port 1) of the RF network by itself which is fixed regardless of the electrical environment surrounding the network. On the other hand, the RL given by $10\log|\Gamma_1|^2$ in (7.76) is the RL under any termination condition and varies depending on the electrical environment around the network. For instance, when a designed RF circuit is placed in a system, the RL at port 1 of the RF circuit obtained from (7.76) changes depending on the impedance presented to port 2 of the RF circuit. Similarly, the RL of the RF network at port 2 is given as

$$\mathrm{RL}_2 = \begin{cases} 10\log|S_{22}|^2 & \text{for matched condition} \\ 10\log|\Gamma_2|^2 & \text{for general termination} \end{cases} \tag{7.78}$$

7.5.4.2 Insertion Loss and Gain. IL is inherent in RF circuits and is caused by the reflection and dissipation of power (assuming radiation is negligible). IL is used to characterize the loss between two separate ports and is defined (from the input port 1 to output port 2) according to Figure 7.20 as

$$\mathrm{IL}_{21} \triangleq 10\log\frac{P_{\mathrm{out}}}{P_{\mathrm{in}}} \quad \text{(dB)} \tag{7.79}$$

which is equal to

$$\mathrm{IL}_{21} = 10\log|T_{21}|^2 \quad \text{(dB)} \tag{7.80}$$

Under the matched termination at port 2, the IL (from port 1 to port 2) becomes

$$\mathrm{IL}_{21} = 10\log|S_{21}|^2 \quad \text{(dB)} \tag{7.81}$$

Similar to the RLs determined by Γ_1 and S_{11}, the ILs given in (7.80) and (7.81) are generally different. The IL given by (7.81) is the IL under the matched condition and represents the IL of the RF network by itself. It is constant with respect to the electrical environment around the network. On the other hand, the IL given by (7.80) is the IL under any termination condition and changes depending on the electrical environment surrounding the network. The IL from ports 2 to 1 is

$$\mathrm{IL}_{12} = \begin{cases} 10\log|S_{12}|^2 & \text{for matched condition} \\ 10\log|T_{12}|^2 & \text{for general termination} \end{cases} \tag{7.82}$$

As stated earlier, the two factors contributing to the IL are the losses due mismatch (or reflection) and power dissipation. To determine the mismatch loss for an RF circuit, we assume that it is lossless; that is, there is no dissipation and the only loss present is the mismatch loss. Under the lossless condition, $P_{\mathrm{out}} = P_{\mathrm{in}} - P_r$, and hence the mismatch loss (at port 1 or from ports 1 to 2) of an RF circuit in general is obtained as

$$\text{Mismatch loss (dB)} = 10\log\frac{P_{\mathrm{in}} - P_r}{P_{\mathrm{in}}} = 10\log(1 - |\Gamma_1|^2) \tag{7.83}$$

This mismatch loss is in fact equal to the IL of a lossless RF circuit. Using the *S*-parameters, the mismatch loss of an RF circuit (at port 1 or from ports 1 to 2) under the matched condition can hence be determined from (7.83) as

$$\text{Mismatch loss (dB)} = 10\log(1 - |S_{11}|^2) \tag{7.84}$$

To determine the dissipation loss which is hereafter referred to as the circuit loss, we again consider Figure 7.20 and let P_d be the dissipated power in the RF network. The actual power propagating into the two-port network (i.e., $P_{\text{in}} - P_r$) is partly dissipated in the network (P_d) and partly transmitted to port 2 (P_{out}). Therefore, we can write:

$$P_{\text{in}} - P_r = P_d + P_{\text{out}} \tag{7.85}$$

Dividing (7.85) by P_{in} leads to

$$\frac{P_d}{P_{\text{in}}} = 1 - \frac{P_r}{P_{\text{in}}} - \frac{P_{\text{out}}}{P_{\text{in}}} \tag{7.86}$$

Equation (7.86) can be rewritten using (7.75), (7.76), (7.79), and (7.80) as

$$\frac{P_d}{P_{\text{in}}} = 1 - |\Gamma_1|^2 - |T_{21}|^2 \tag{7.87}$$

which represents the normalized dissipated power that causes (circuit) loss to the signal from ports 1 to 2. The power, after dissipated, is transmitted to port 2. Using the same argument for the mismatch loss determined from the power transmitted after being reflected, we can state that the sum of the circuit loss and P_d/P_{in} would be equal to 1. This leads to

$$\text{Circuit loss} = 1 - \frac{P_d}{P_{\text{in}}} = |T_{21}|^2 + |\Gamma_1|^2 \tag{7.88}$$

upon using (7.87). Under the matched condition, (7.88) becomes the circuit loss (from ports 1 to 2) of a two-port network as

$$\text{Circuit loss (dB)} = 10\log(|S_{21}|^2 + |S_{11}|^2) \tag{7.89}$$

Similarly, the circuit loss from ports 2 to 1 of a two-port network under the matched condition is

$$\text{Circuit loss (dB)} = 10\log(|S_{12}|^2 + |S_{22}|^2) \tag{7.90}$$

An important note needs to be made at this point. In RF circuit design, both the losses due to mismatch and dissipation need to be minimized. However, while the mismatch loss can be improved through better circuit matching, it may be difficult to reduce the circuit loss significantly due to possible constraints imposed on a particular circuit. For instance, while the circuit loss may be reduced by using the thickest metal on the top-most metal layer in a complementary metal oxide silicon (CMOS) structure or employing a shield between the metal and the lossy silicon substrate, etc., these schemes may not be possible for some circuits. Nevertheless, assuming all design techniques have been exercised to reduce the circuit loss, the only possibility left is improving the matching. Therefore, in the design or measurement of an RF circuit, both the RL and IL need to be obtained instead of just the IL. A correct knowledge of the RL allows RF designers to determine accurately the contributions to the IL from the mismatch and circuit losses, from which proper actions can then be taken.

The gain (G) of RF networks is defined exactly the same as the IL by

$$G \triangleq 10\log\frac{P_{\text{out}}}{P_{\text{in}}} = 10\log|S_{21}|^2 \quad \text{(dB)} \tag{7.91}$$

The foregoing results can be extended for N-port RF networks as follows. The IL (from ports i to j) and RL (at port i) of an N-port RF network can be determined from its S-parameters (i.e., under matched terminations) as

$$\mathrm{IL}_{ji} = 10\log|S_{ji}|^2 \quad \text{(dB)} \tag{7.92}$$

$$\mathrm{RL}_{i} = 10\log|S_{ii}|^2 \quad \text{(dB)} \tag{7.93}$$

while all other (unused) ports are terminated with a matched load.

7.6 CHAIN MATRIX

Another matrix type that is popular among RF engineers is the chain matrix, commonly known as $ABCD$ matrix, for two-port circuits. We consider a two-port RF circuit with designated port currents and voltages as shown in Figure 7.21. The $ABCD$ matrix of a two-port circuit is defined in the following $ABCD$-matrix equation:

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} \tag{7.94}$$

where its parameters are obtained as

$$\begin{aligned} A &= \left.\frac{V_1}{V_2}\right|_{I_2=0} \\ B &= \left.\frac{V_1}{I_2}\right|_{V_2=0} \\ C &= \left.\frac{I_1}{V_2}\right|_{I_2=0} \\ D &= \left.\frac{I_1}{I_2}\right|_{V_2=0} \end{aligned} \tag{7.95}$$

As we can see from (7.95), in order to determine these parameters, port 2 needs to be open ($I_2 = 0$) or short ($V_2 = 0$) accordingly, which is very difficult, if not impossible, to achieve at RF as we discussed previously for the impedance and admittance matrices. The ABCD matrix, like the impedance and admittance matrices, is thus not useful for RF circuit measurement.

The ABCD matrix of two-port RF circuits has the following property:

$$\text{Reciprocal:} \quad AD - BC = 1 \tag{7.96}$$

$$\text{Symmetrical:} \quad A = D \tag{7.97}$$

$$\text{Lossless:} \quad \begin{matrix} A \text{ and } D\text{: Real} \\ B \text{ and } C\text{: Imaginary} \end{matrix} \tag{7.98}$$

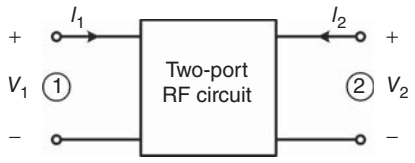


Figure 7.21. A two-port circuit.

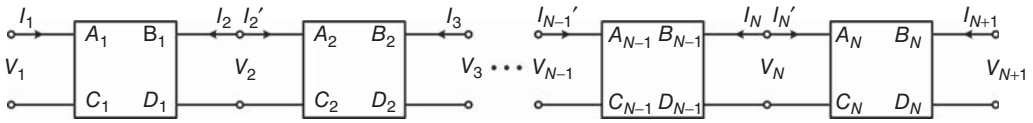


Figure 7.22. N Cascaded two-port RF circuits. A_i, B_i, C_i, D_i are the ABCD parameters of circuit i.

The $ABCD$ matrix is very useful for the evaluation of cascaded two-port circuits. We consider N two-port RF circuits connected in cascade as shown in Figure 7.22. Applying the $ABCD$ matrix formula in (7.94) to the network in Figure 7.22 gives

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} V_2 \\ I_2' \end{bmatrix} \tag{7.99}$$

$$\begin{bmatrix} V_2 \\ I_2' \end{bmatrix} = \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} V_3 \\ -I_3 \end{bmatrix} \tag{7.100}$$

$$\cdots$$

$$\begin{bmatrix} V_{N-1} \\ I_{N-1}' \end{bmatrix} = \begin{bmatrix} A_{N-1} & B_{N-1} \\ C_{N-1} & D_{N-1} \end{bmatrix} \begin{bmatrix} V_N \\ -I_N \end{bmatrix} = \begin{bmatrix} A_{N-1} & B_{N-1} \\ C_{N-1} & D_{N-1} \end{bmatrix} \begin{bmatrix} V_N \\ I_N' \end{bmatrix} \tag{7.101}$$

$$\begin{bmatrix} V_N \\ I_N' \end{bmatrix} = \begin{bmatrix} A_N & B_N \\ C_N & D_N \end{bmatrix} \begin{bmatrix} V_{N+1} \\ -I_{N+1} \end{bmatrix} \tag{7.102}$$

Substituting (7.102) into (7.101) and repeating the substitution of the resultant equation into the preceding equation until (7.99) lead to

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} \cdots \begin{bmatrix} A_{N-1} & B_{N-1} \\ C_{N-1} & D_{N-1} \end{bmatrix} \begin{bmatrix} A_N & B_N \\ C_N & D_N \end{bmatrix} \begin{bmatrix} V_{N+1} \\ -I_{N+1} \end{bmatrix} \tag{7.103}$$

which shows that the total $ABCD$ matrix of cascaded circuits is equal to the product of the $ABCD$ matrices of all individual circuits.

7.7 SCATTERING TRANSMISSION MATRIX

Scattering transmission matrix, $[T]$, combines the S-matrix and ABCD matrix for two-port networks. To formulate the scattering transmission matrix, we consider a two-port network as shown in Figure 7.23. Similar to the S-matrix, the scattering transmission matrix relates the incident and reflected waves between the input and output ports as

$$\begin{bmatrix} V_1^+ \\ V_1^- \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} V_2^- \\ V_2^+ \end{bmatrix} \tag{7.104}$$

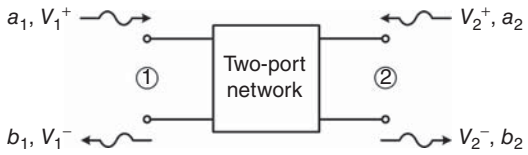


Figure 7.23. A two-port network.

or, in terms of normalized voltage waves,

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} b_2 \\ a_2 \end{bmatrix} \tag{7.105}$$

where a_1, a_2 and b_1, b_2 are the normalized incident and reflected voltages as given in (7.25). For cascaded two-port networks, we can also derive the following equation similar to that obtained in (7.103) for the $ABCD$ matrix:

$$[T] = [T_1][T_2] \cdots [T_{N-1}][T_N] \tag{7.106}$$

where $[T]$ is the overall scattering transmission matrix and $[T_n], n = 1, 2, \ldots, N$, is the scattering transmission matrix for network i.

For reciprocal two-port networks,

$$T_{11}T_{22} - T_{12}T_{21} = 1 \tag{7.107}$$

for normalized voltage waves, which correspond to normalized power waves; that is, $P_{\text{in},i} = \frac{1}{2}|a_i^+|^2$, $i = 1, 2$. For non-normalized voltage waves,

$$T_{11}T_{22} - T_{12}T_{21} \neq 1 \tag{7.108}$$

in general.

The definition of the scattering transmission matrix as given in (7.105) is not the only definition available. Other definitions for this matrix can be employed – for instance,

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \tag{7.109}$$

However, we should be aware that some of the definitions may not be applicable for all RF networks since they may cause potential problems in the characterization of some networks – for instance, causing unrealistically infinite gain for an assumed unilateral active device.

7.8 CONVERSION BETWEEN TWO-PORT PARAMETERS

While the measurement of RF circuits is always performed based on the S-parameters, the S-matrix as well as other matrices can be employed together or separately in analyzing RF circuits. As such, the conversion between the parameters of different matrices is useful. In this section, we discuss such conversion for two-port networks.

7.8.1 Conversion from [*Z*] to [*ABCD*]

We consider again a two-port network as shown in Figure 7.21. We can write the following equations using the impedance $[Z]$ and $[ABCD]$ matrix:

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{7.110}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2 \tag{7.111}$$

and

$$V_1 = AV_2 - BI_2 \tag{7.112}$$

$$I_1 = CV_2 - DI_2 \tag{7.113}$$

From (7.110), we obtain

$$I_1 = \frac{1}{Z_{21}} V_2 - \frac{Z_{22}}{Z_{21}} I_2 \tag{7.114}$$

Comparing (7.113) to (7.114) results in

$$C = \frac{1}{Z_{21}}$$

$$D = \frac{Z_{22}}{Z_{21}} \tag{7.115}$$

Substituting (7.114) into (7.110) gives

$$V_1 = \frac{Z_{11}}{Z_{21}} V_2 - \left(\frac{Z_{11}Z_{22} - Z_{12}Z_{21}}{Z_{21}} \right) I_2 \tag{7.116}$$

which, upon comparison with (7.112), leads to

$$A = \frac{Z_{11}}{Z_{21}}$$

$$B = \frac{Z_{11}Z_{22} - Z_{12}Z_{21}}{Z_{21}} \tag{7.117}$$

Using (7.24) and (7.105) for the scattering and scattering transmission parameters, respectively, we can derive the following relations between these parameters:

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{S_{21}} & -\dfrac{S_{22}}{S_{21}} \\ \dfrac{S_{11}}{S_{21}} & \dfrac{S_{12}S_{21} - S_{11}S_{22}}{S_{21}} \end{bmatrix} \tag{7.118}$$

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} \dfrac{T_{21}}{T_{11}} & \dfrac{T_{11}T_{22} - T_{12}T_{21}}{T_{11}} \\ \dfrac{1}{T_{11}} & -\dfrac{T_{12}}{T_{11}} \end{bmatrix} \tag{7.119}$$

Following a similar procedure, we can derive the conversion formulas between the other parameters of a two-port network. Table 7.1 summarizes the conversion equations for a two-port network assuming the same load or terminating impedance Z_o at both ports [2].

TABLE 7.1. Conversion Between Parameters of Two-Port Networks $Z_1 = Z_{11}Z_{22} - Z_{12}Z_{21}$; $Z_2 = (Z_{11} + Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}$; $Y_1 = Y_{11}Y_{22} - Y_{12}Y_{21}$; $Y_2 = (Y_{11} + Y_o)(Y_{22} + Y_o) - Y_{12}Y_{21}$; $A_1 = A + B/Z_o + CZ_o + D$; $S_1 = (1 - S_{11})(1 - S_{22}) - S_{12}S_{21}$; $S_2 = (1 + S_{11})(1 + S_{22}) - S_{12}S_{21}$; $Y_o = 1/Z_o$; Z_o: Load or Terminating Impedance

Parameters	$[S]$	$[Z]$	$[Y]$	$[ABCD]$
S_{11}	S_{11}	$\frac{(Z_{11} - Z_o)(Z_{22} + Z_o) - Z_{12}Z_{21}}{Z_2}$	$\frac{(Y_o - Y_{11})(Y_o + Y_{11}) + Y_{12}Y_{21}}{Y_2}$	$\frac{A + B/Z_o - CZ_o - D}{A_1}$
S_{12}	S_{12}	$\frac{2Z_{12}Z_o}{Z_2}$	$-\frac{2Y_{12}Y_o}{Y_2}$	$\frac{2(AD - BC)}{A_1}$
S_{21}	S_{21}	$\frac{2Z_{21}Z_o}{Z_2}$	$-\frac{2Y_{21}Y_o}{Y_2}$	$\frac{2}{A_1}$
S_{22}	S_{22}	$\frac{(Z_{11} + Z_o)(Z_{22} - Z_o) - Z_{12}Z_{21}}{Z_2}$	$\frac{(Y_o + Y_{11})(Y_o - Y_{22}) + Y_{12}Y_{21}}{Y_2}$	$\frac{-A + B/Z_o - CZ_o + D}{A_1}$
Z_{11}	$Z_o\frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21}}{S_1}$	Z_{11}	$\frac{Y_{22}}{Y_1}$	$\frac{A}{C}$
Z_{12}	$\frac{2Z_oS_{12}}{S_1}$	Z_{12}	$-\frac{Y_{12}}{Y_1}$	$\frac{AD - BC}{C}$
Z_{21}	$\frac{2Z_oS_{21}}{S_1}$	Z_{21}	$-\frac{Y_{21}}{Y_1}$	$\frac{1}{C}$
Z_{22}	$Z_o\frac{(1 - S_{11})(1 + S_{22}) + S_{12}S_{21}}{S_1}$	Z_{22}	$\frac{Y_{11}}{Y_1}$	$\frac{D}{C}$
Y_{11}	$Y_o\frac{(1 - S_{11})(1 + S_{22}) + S_{12}S_{21}}{S_2}$	$\frac{Z_{22}}{Z_1}$	Y_{11}	$\frac{D}{B}$
Y_{12}	$\frac{-2Y_oS_{12}}{S_2}$	$-\frac{Z_{12}}{Z_1}$	Y_{12}	$\frac{BC - AD}{B}$
Y_{21}	$\frac{-2Y_oS_{21}}{S_2}$	$-\frac{Z_{21}}{Z_1}$	Y_{21}	$-\frac{1}{B}$
Y_{22}	$Y_o\frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21}}{S_2}$	$\frac{Z_{11}}{Z_1}$	Y_{22}	$\frac{A}{B}$
A	$\frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21}}{2S_{21}}$	$\frac{Z_{11}}{Z_{21}}$	$-\frac{Y_{22}}{Y_{21}}$	A
B	$Z_o\frac{(1 + S_{11})(1 + S_{22}) - S_{12}S_{21}}{2S_{21}}$	$\frac{Z_1}{Z_{21}}$	$-\frac{1}{Y_{21}}$	B
C	$\frac{(1 - S_{11})(1 - S_{22}) - S_{12}S_{21}}{2Z_oS_{21}}$	$\frac{1}{Z_{21}}$	$-\frac{Y_1}{Y_{21}}$	C
D	$\frac{(1 - S_{11})(1 + S_{22}) - S_{12}S_{21}}{2S_{21}}$	$\frac{Z_{22}}{Z_{21}}$	$-\frac{Y_{11}}{Y_{21}}$	D

REFERENCES

1. Hewlett Packard, S-Parameter Design, Application Note 154, pp. 1–5, Apr. 1972.
2. D. Pozar, *Microwave Engineering*, Third Edition, p. 187, John Wiley & Sons, Inc., New York, 2005.

PROBLEMS

7.1 Derive Eq. (7.16).

7.2 Derive Eqs. (7.27) and (7.28) using the *S*-matrix equation instead of the signal flow graph as done in Appendix A11.

7.3 Assume a two-port network characterized by an impedance matrix [*Z*] or admittance matrix [*Y*] is represented by an equivalent π-network as shown in Figure 7.11. Following the formulation presented in Section 7.4.1, derive the elements of the π-equivalent circuit in terms of Y_{1O}, Y_{2O} and Y_{1S}, Y_{2S} defined as the input admittance at port 1, 2 with port 2, 1 open and short circuited, respectively.

7.4 We consider a lossless transmission line having characteristic impedance Z_o and length ℓ, as shown in Figure 7.8(a). Derive its impedance matrix directly using the impedance-matrix equation. Compare the results with those obtained in Eq. (7.14).

7.5 Derive the admittance parameters given in Eq. (7.17) for a π-network as shown in Figure 7.11.

7.6 Derive the admittance matrix of a lossless transmission line having characteristic impedance Z_o, phase constant β, and length ℓ.

7.7 Derive the impedance parameters for a circuit as shown in Figure P7.1(a). The transformer's parameters in Figure P7.1(b) are defined as

$$\frac{V_1}{V_2} = \frac{n_1}{n_2} \qquad \frac{I_1}{I_2} = -\frac{n_2}{n_1}$$

7.8 Consider the circuit as shown in Figure P7.2. The transformer's parameters are given Problem 7.7.

a) Derive the impedance parameters.

b) Assume $Z_{11} = j2\ \Omega$, $Z_{12} = j/\sqrt{2}\ \Omega$, $Z_{22} = -j0.25\ \Omega$. Find Z_1, n, and the electrical length of the transmission line.

7.9 Derive the *S*-parameters of a series impedance *Z* as shown in Figure P7.3. You would need to derive the *S*-parameters directly using the *S*-matrix equation – do not derive other parameters like the impedance parameters and then convert to the *S*-parameters using the conversion formulas.

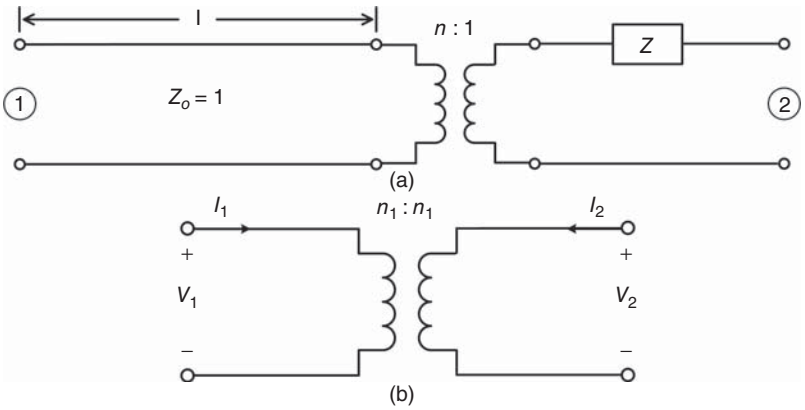


Figure P7.1.

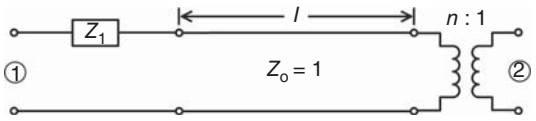


Figure P7.2.

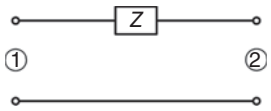


Figure P7.3.

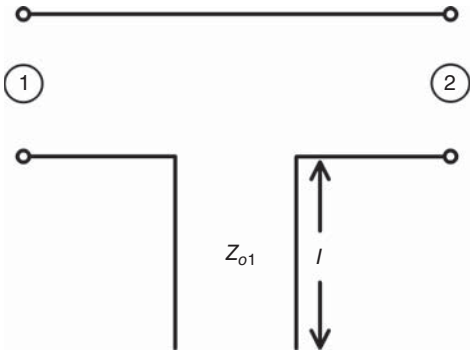


Figure P7.4.

7.10 Find the *S*-parameters of a lossless open-circuited stub connected in series as shown in Figure P7.4.

7.11 Repeat Problem 7.10 for a lossless short-circuited stub.

7.12 Find the *S*-parameters of a shunt lossless short-circuited stub as shown in Figure P7.5. You would need to derive the *S*-parameters directly using the *S*-matrix equation – do not derive other parameters like the admittance parameters and then convert to the *S*-parameters using the conversion formulas.

7.13 Repeat Problem 7.12 for a lossless open-circuited stub.

7.14 A transmission-line discontinuity usually encountered in RF circuits is a junction between two transmission lines of different characteristic impedances Z_{o1} and Z_{o2} as shown in Figure P7.6. This junction can be viewed as a two-port network whose (terminal) reference planes are right at the junction. Derive the [*S*] matrix characterizing the junction.

7.15 Derive the *S*-parameters of a T-network as shown in Figure 7.5 when ports 1 and 2 are terminated with different loads; that is, $Z_{L1} = Z_{o1} \neq Z_{L2} = Z_{o2}$.

7.16 Derive the *S*-parameters of a π-network as shown in Figure 7.11 when ports 1 and 2 are terminated with different loads; that is, $Z_{L1} = Z_{o1} \neq Z_{L2} = Z_{o2}$.

7.17 Derive the *S*-parameters of a lossy transmission line between ports 1 and 2 as shown in Figure P7.7. Obtain the results when the line is lossless.

7.18 Derive the inequality (7.55) for a lossy *N*-port network.

7.19 The following *S*-parameters are measured for a CMOS RFIC amplifier at 10 GHz using a vector network analyzer:

$$S_{11} = 0.2\angle 60^\circ \qquad S_{12} = 0.1\angle -50^\circ$$

$$S_{21} = 6.5\angle 120^\circ \qquad S_{22} = 0.15\angle 95^\circ$$

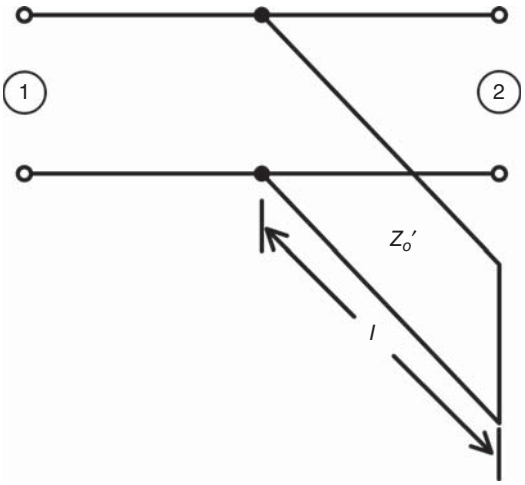


Figure P7.5.

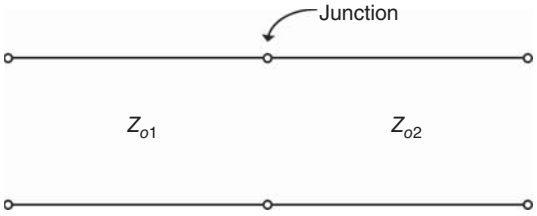


Figure P7.6.

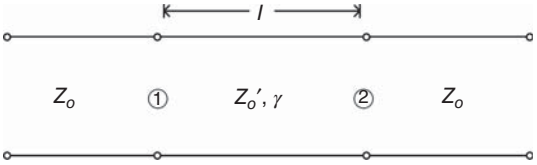


Figure P7.7.

a) Calculate the gain of the amplifier in dB.
b) Calculate the amplifier's input and output RLs in dB.
c) Calculate the amplifier's isolation in dB.

7.20 An RFIC component has the following measured *S*-parameters:

f (GHz)	S_{11}	S_{12}	S_{21}	S_{22}
10	$0.1\angle 20°$	$0.96\angle 40°$	$0.96\angle 40°$	$0.1\angle 20°$
15	$0.93\angle -45°$	$0.07\angle 120°$	$0.07\angle 120°$	$0.93\angle -45°$

where 1 and 2 represent the component's input and output ports, respectively. The component is assumed to be terminated in matched loads. Calculate the following parameters at 10 and 15 GHz:

a) RL and mismatch (or reflection) loss in dB at the input and output.
b) Reflection coefficient and voltage standing wave ratio (VSWR) at the input.

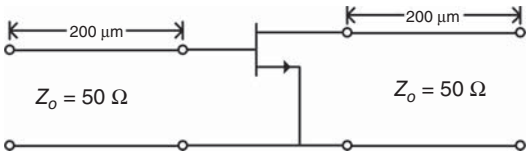


Figure P7.8.

c) IL in dB.
d) Circuit loss in dB.

7.21 The measured *S*-parameters at 20 GHz of a 0.25-μm metal oxide semiconductor transistor (MOSFET) connected with two 50-Ω microstrip lines, as shown in Figure P7.8, are

$$S_{11} = 0.718\angle -38^\circ \quad S_{12} = 0.029\angle 129^\circ$$
$$S_{21} = 1.675\angle 126.6^\circ \quad S_{22} = 0.751\angle 10.2^\circ$$

We assume the microstrip lines are realized using the CMOS profile shown in Figure P15.2. The top conductor is on M6. The bottom conductor (ground plane) is on M1 and is assumed infinitely large. Determine the *S*-parameters of the MOSFET alone.

7.22 Consider the following *S*-matrix of a three-port RF circuit:

$$\begin{bmatrix} 0.1 & 0.1e^{j\theta_1} & 0.1 \\ 0.1e^{j\theta_1} & e^{j\theta_2} & 0.2e^{-j\theta_3} \\ 0.1 & 0.2e^{-j\theta_3} & 0.2e^{j\theta_4} \end{bmatrix}$$

where θ_1, θ_2, and θ_3 are arbitrary angles.
a) Is this RF circuit lossless? Provide your rationale.
b) Is this RF circuit reciprocal? Provide your rationale.
c) Is this RF circuit perfectly matched? Provide your rationale.
d) Looking into port 2, what do you expect to see electrically? Provide your rationale.
e) Assume $\theta_4 = -30^\circ$ and $Z_o = 50\ \Omega$, describe the electrical element (type and value) you expect to see when looking into port 3.

7.23 Derive the [*ABCD*] matrix of a shunt admittance *Y* as shown in Figure 7.16. You would need to derive the *ABCD*-parameters directly using the *ABCD*-matrix equation – do not derive other parameters like the *S*-parameters and then convert to the *ABCD*-parameters using the conversion formulas.

7.24 Derive the [*ABCD*] matrix of a series impedance *Z* as shown in Figure P7.3. You would need to derive the *ABCD*-parameters directly using the *ABCD*-matrix equation – do not derive other parameters like the impedance parameters and then convert to the *ABCD*-parameters using the conversion formulas.

7.25 Find the *ABCD* matrix of a shunt lossless short-circuited stub as shown in Figure P7.5.

7.26 Consider an RF circuit consisting of a lossless transmission line and a lossless short-circuit stub connected in series as shown in Figure P7.9. Find the *ABCD* matrix of the RF circuit.

7.27 We consider a two-port RF circuit with transmission lines having different characteristic impedances $Z_{o1} = 10\ \Omega$ and $Z_{o2} = 50\ \Omega$ connected at ports 1 and 2, respectively. Assume that the *S*-parameters of

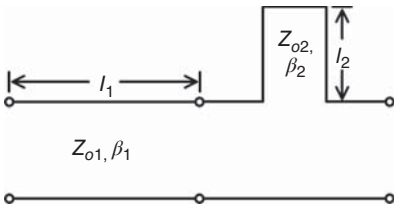


Figure P7.9.

the RF circuit at 10 GHz are

$$S_{11} = 0.6\angle 5^\circ \qquad S_{12} = 0.866\angle 40^\circ$$
$$S_{21} = 0.8\angle 30^\circ \quad S_{22} = 0.51\angle -10^\circ$$

a) Calculate the *S*-parameters of the same two-port RF circuit when its ports are connected with transmission lines having the same characteristic impedance of 50 Ω.
b) Determine the ABCD parameters of the RF circuit in Part (a).

7.28 Derive (7.96)–(7.98) for the reciprocal, symmetrical and lossless properties of *ABCD* matrix.

7.29 We consider a two-port network consisting of a parallel-coupled transmission line connected as shown in Figure P7.10. Two ports of the two coupled lines are tied to form port 1, while the remaining port of the upper line is left open and that of the lower line is used as port 2. We assume that the widths of the coupled lines are identical and so the transmission line supports two propagating modes, namely even and odd mode. Correspondingly, there are even- and odd-mode characteristic impedances (Z_{oe} and Z_{oo}) and electrical lengths (θ_e and θ_o).

a) Derive the impedance matrix of the two-port network.
b) Derive the chain matrix of the two-port network.
c) Use the obtained chain matrix to derive an equivalent network of the two-port network. This equivalent network consists of a transmission line of characteristic impedance $Z_{oe}/2$ and electrical length θ_e in series with a short-circuited transmission line of characteristic impedance $Z_{oo}/2$ and electrical length θ_o. Draw this equivalent network.
d) Use the obtained chain matrix in Part (b) to derive another equivalent network of the two-port network. This equivalent network consists of an open-circuited transmission line having electrical length θ_e connected in shunt with a transmission line having the same electrical length θ_e. Derive the characteristic impedances of the constituent transmission lines. Besides the fact that the shunt transmission line in this equivalent circuit is much easier to be realized than the series transmission line in the equivalent circuit in Part (c), through an examination of the characteristic impedances, do you think this equivalent circuit is better than the one in Part (c) for circuit analysis and synthesis? Provide your rationale.

7.30 Derive the scattering transmission matrix of a lossless transmission line having characteristic impedance Z_o, phase constant β, and length ℓ.

7.31 Derive Eqs. (7.118) and (7.119).

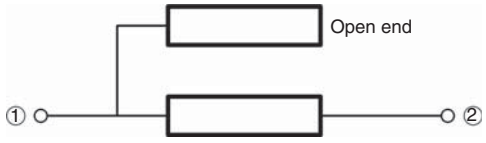


Figure P7.10.

7.32 We consider the following scattering transmission matrix of a two-port network:

$$\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}$$

Derive the conversion formulas between this scattering transmission matrix and the S-matrix.

7.33 A scattering transmission matrix can also be defined by relating the reflected voltages to the incident voltage waves as

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Derive the conversion formulas from the S-parameters to the scattering transmission parameters. Do you see any problem with the resultant equations? Can these equations be used for any RF networks?

7.34 Derive the reciprocal condition (7.107) for the scattering transmission matrix.

7.35 Prove that the reciprocal property described in (7.107) for the scattering transmission matrix does not always hold for non-normalized voltage waves.

7.36 Derive the conversion formulas between the S- and $ABCD$-parameters of two-port networks.

CHAPTER 8

RF PASSIVE COMPONENTS

Passive components such as directional couplers, power dividers, and filters are essential for radio frequency (RF) circuits and systems. They can be implemented as stand-alone components in RF systems, such as a band-pass filter in systems to define the bandwidth of the transmitting or receiving signal, or an integral part of RF circuits such as a 180° hybrid in balanced mixers. RF passive components, along with RF active components (which may also consist of some passive components) such as amplifiers, constitute radio frequency integrated circuit (RFIC) engineering. RFIC engineers thus should have sufficient knowledge on the analysis and design of passive RF components. In this chapter, we will present the analysis and design of various basic RF passive components including directional couplers, hybrids, power dividers, and filters. It is noted that the materials are also applicable to non-monolithic RF circuits such as microwave integrated circuit (MIC).

8.1 CHARACTERISTICS OF MULTIPORT RF PASSIVE COMPONENTS

RF passive components, such as active components, can have single or multiport. Multiport passive components such as three- and four-port have unique properties that dictate their overall performance limitations. RF designers need to know these properties so that they do not try to obtain the performance that is inherently not achievable. This section presents the characteristics of three- and four-port passive components, which can serve as the basic for extension to passive components comprising five or more ports. These characteristics are useful for the design of three- and four-port circuits as they indicate the intrinsic performance that is expected for these circuits.

8.1.1 Characteristics of Three-Port Components

There are various three-port passive components used in RF circuits and systems such as diplexers, power dividers, power combiners, circulators, switches, single directional couplers, etc. Figure 8.1 shows an example of a 0.18-μm complementary metal oxide semiconductor (CMOS) RFIC three-port single-pole double-throw (SPDT) switch operating from 22 to 29 GHz. In this SPDT switch, signal at the input port 1 travels to either the output port 2 or output port 3.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

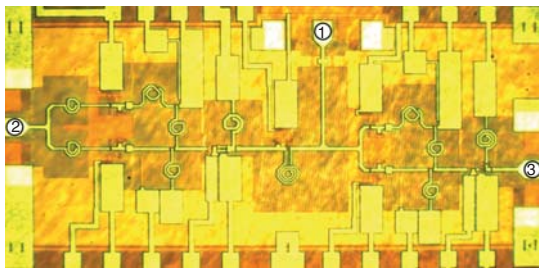


Figure 8.1. Photograph of a 0.18-μm CMOS RFIC three-port SPDT switch.

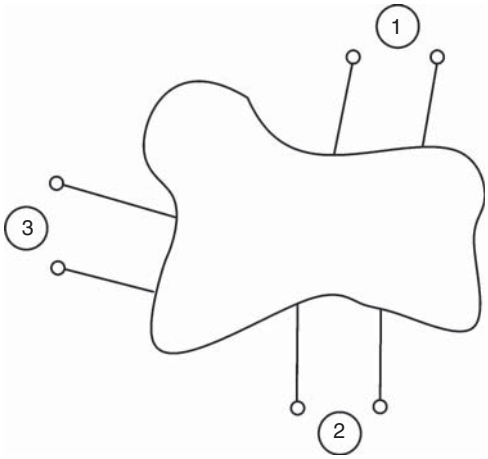


Figure 8.2. Three-port network.

A general three-port network as shown in Figure 8.2 can be described by the following scattering (S) matrix:

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} \tag{8.1}$$

Three-port networks have several properties which are different from two-port networks. Three-port networks cannot be matched, lossless,[1] and reciprocal simultaneously. If a three-port network is matched at three ports, it must be lossy (i.e., having resistive elements) or nonreciprocal. A three-port network, however, can be lossless (no resistive elements employed) and reciprocal with one or two ports matched. These properties are discussed as follows. These characteristics serve as the foundation and set the criteria for the design of any three-port passive circuit – one example is the design of the Wilkinson power divider that we will discuss in Section 8.4.

8.1.1.1 No Matched, Lossless, Reciprocal Three-Port Networks. We begin the analysis by assuming that a three-port network as shown in Figure 8.2 is perfectly matched at three ports, reciprocal, and lossless. For perfect match, the three-port network must have

$$S_{11} = S_{22} = S_{33} = 0 \tag{8.2}$$

[1]Lossless is used loosely here; it merely implies that no restive elements are used in the networks. Practical RF components, even without resistive circuit elements, are lossy since all circuit elements including dielectrics and conductors are non-perfect.

For reciprocity, its S-parameters satisfy

$$\begin{aligned} S_{12} &= S_{21} \\ S_{13} &= S_{31} \\ S_{23} &= S_{32} \end{aligned} \tag{8.3}$$

The lossless condition follows $[S]^T[S]^* = [I]$ in Eq. (7.48) for lossless networks in Chapter 7 (Scattering Parameters) and can be obtained as

$$\begin{aligned} |S_{11}|^2 + |S_{12}|^2 + |S_{13}|^2 &= 1 \\ |S_{12}|^2 + |S_{22}|^2 + |S_{32}|^2 &= 1 \\ |S_{13}|^2 + |S_{23}|^2 + |S_{33}|^2 &= 1 \end{aligned} \tag{8.4}$$

and

$$\begin{aligned} S_{11}S_{12}^* + S_{12}S_{22}^* + S_{13}S_{32}^* &= 0 \\ S_{11}S_{13}^* + S_{21}S_{32}^* + S_{13}S_{33}^* &= 0 \\ S_{12}S_{11}^* + S_{22}S_{12}^* + S_{32}S_{13}^* &= 0 \\ S_{12}S_{13}^* + S_{22}S_{32}^* + S_{32}S_{33}^* &= 0 \end{aligned} \tag{8.5}$$

making use of (8.3). Utilizing (8.2)–(8.4), we can derive:

$$\begin{aligned} |S_{12}|^2 + |S_{13}|^2 &= 1 \\ |S_{12}|^2 + |S_{32}|^2 &= 1 \\ |S_{13}|^2 + |S_{32}|^2 &= 1 \end{aligned} \tag{8.6}$$

From (8.2), (8.3), and (8.5), we obtain

$$\begin{aligned} S_{13}S_{32}^* &= 0 \\ S_{12}S_{32}^* &= 0 \\ S_{12}S_{13}^* &= 0 \\ S_{32}S_{13}^* &= 0 \end{aligned} \tag{8.7}$$

Examination of (8.7) reveals that all the equations in (8.7) cannot be satisfied unless two parameters among S_{12}, S_{32}, and S_{13} are equal to zero. This condition, however, leads to the result that one equation in (8.6) is not satisfied; implying that match, reciprocity, and no loss cannot exist simultaneously for any three-port network. Therefore, matched, lossless, reciprocal three-port component cannot be realized. In other words, it is impossible to design a three-port passive reciprocal RFIC containing no resistive elements with all ports matched.

8.1.1.2 Lossy or Nonreciprocal Matched Three-Port Networks. A matched three-port network must be either lossy or nonreciprocal. That is, either resistive elements must be introduced into the circuit or the circuit would need to be nonreciprocal. The S-matrix of a matched, reciprocal, lossy three-port network can be described as

$$[S] = \begin{bmatrix} 0 & S_{12} & S_{13} \\ S_{12} & 0 & S_{23} \\ S_{13} & S_{23} & 0 \end{bmatrix} \tag{8.8}$$

where the S-parameters S_{12}, S_{13}, and S_{23} can be determined upon applying certain conditions for specific three-port components. An example of such a network is the Wilkinson power divider having a common resistor (lossy element) between the two output arms described in Section 8.4 whose S-parameters can be derived as

$$[S] = -\frac{j}{\sqrt{2}}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{8.9}$$

For a matched, lossless, nonreciprocal three-port network, the S-matrix is

$$[S] = \begin{bmatrix} 0 & S_{12} & S_{13} \\ S_{21} & 0 & S_{23} \\ S_{31} & S_{32} & 0 \end{bmatrix} \tag{8.10}$$

which, upon applying the lossless condition $[S]^T[S]^* = [I]$, gives

$$[S] = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{8.11}$$

or

$$[S] = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{8.12}$$

The S-matrix in (8.11) or (8.12) shows that a signal entering a matched, lossless, nonreciprocal three-port component can travel only in one direction. This is indeed the characteristic of three-port circulators as shown in Figure 8.3 which can be used to connect the output of a transmitter, the input of a receiver, and the input of an antenna together to allow the transmitter and receiver to share a single antenna.

8.1.1.3 Lossless, Reciprocal Three-Port Networks with Two Ports Matched. We now consider a lossless, reciprocal three-port network having two matched ports. For the analysis illustration purpose, we assume that ports 1 and 2 are matched. The scattering matrix for a reciprocal three-port network with ports 1 and 2 matched is given as

$$[S] = \begin{bmatrix} 0 & S_{12} & S_{13} \\ S_{12} & 0 & S_{23} \\ S_{13} & S_{23} & S_{33} \end{bmatrix} \tag{8.13}$$

The lossless condition can be written utilizing the lossless criterion $[S]^T[S]^* = [I]$ as

$$\begin{bmatrix} 0 & S_{12} & S_{13} \\ S_{12} & 0 & S_{23} \\ S_{13} & S_{23} & S_{33} \end{bmatrix}\begin{bmatrix} 0 & S_{12}^* & S_{13}^* \\ S_{12}^* & 0 & S_{23}^* \\ S_{13}^* & S_{23}^* & S_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8.14}$$

Expanding (8.14) gives

$$\begin{aligned} |S_{12}|^2 + |S_{13}|^2 &= 1 \\ |S_{12}|^2 + |S_{23}|^2 &= 1 \\ |S_{13}|^2 + |S_{23}|^2 + |S_{33}|^2 &= 1 \end{aligned} \tag{8.15}$$

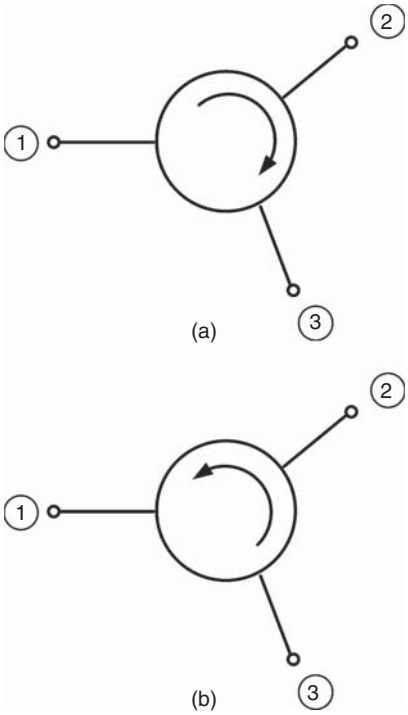


Figure 8.3. Circulator showing direction of signal propagation between ports 1, 2, and 3 (a) according to (8.11) or between ports 3, 2, and 1 (b) according to (8.12).

and

$$\begin{aligned} S_{13}^{*} S_{23} &= 0 \\ S_{12}^{*} S_{13} + S_{23}^{*} S_{33} &= 0 \\ S_{23}^{*} S_{12} + S_{33}^{*} S_{13} &= 0 \end{aligned} \tag{8.16}$$

Solving (8.15), we get

$$\begin{aligned} |S_{13}| &= |S_{23}| = 0 \\ S_{12} &= e^{j\theta} \\ S_{33} &= e^{j\phi} \end{aligned} \tag{8.17}$$

where θ is an arbitrary phase. The S-matrix of a lossless, reciprocal three-port with ports 1 and 2 matched can now be obtained from (8.13) and (8.17) as

$$[S] = \begin{bmatrix} 0 & e^{j\theta} & 0 \\ e^{j\theta} & 0 & 0 \\ 0 & 0 & e^{j\phi} \end{bmatrix} \tag{8.18}$$

The S-parameters S_{12} and S_{21} are equal to $e^{j\theta}$, implying that the transmission between ports 1 and 2 is perfect (conforming the lossless condition) with an arbitrary phase. Any transmission phase (θ) can be obtained by using an appropriate length for the terminating transmission line at either port 1 or 2 or both. The S-parameter S_{33} equaling $e^{j\phi}$ shows that, looking into port 3, there is an open, a short, or a reactance (capacitor or inductor).

As an example, it is possible to design a lossless (no resistive elements), reciprocal three-port with ports 1 and 2 matched, and $\theta = 2m\pi$ and $\phi = 2n\pi$, where m and n are integers including zero, which result in

$$[S] = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8.19}$$

It is noted from (8.19) that there is no transmission between ports 1–3 and 2–3, and total reflection would occur at port 3 which indeed represents an open circuit. This information is useful for the design and usage of such three-port component.

8.1.1.4 Lossless, Reciprocal Three-Port Networks with One Port Matched. The S-matrix for a lossless, reciprocal three-port component with only one port matched can also be derived using the lossless, reciprocal and one-port match criteria, demonstrating that such a network can be designed. An example is a power divider having perfect match at the input port 3 and equal power division at the output ports 1 and 2 described in the following S-matrix:

$$[S] = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \tag{8.20}$$

This device is a special 3-dB power divider resembling the H-plane waveguide junction well known in microwave engineering.

8.1.2 Characteristics of Four-Port Components

There are various four-port passive components used in RF circuits and systems such as ring hybrid, coupler, etc. Figure 8.4 shows an example of a four-port ring hybrid realized using a 0.25-μm CMOS process.

The S-matrix of a four-port network as shown in Figure 8.5 is given by

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{31} & S_{32} & S_{33} & S_{44} \end{bmatrix} \tag{8.21}$$

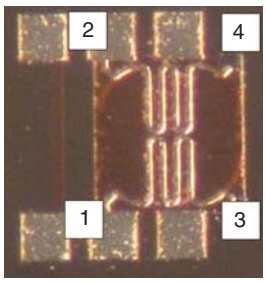


Figure 8.4. A 0.25-μm CMOS RFIC ring hybrid with four ports 1, 2, 3, and 4. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

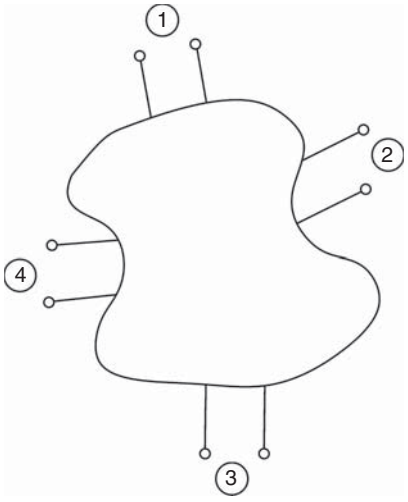


Figure 8.5. Four-port network.

Four-port networks have unique properties which are important for their design and usage, and they set the design criteria that one can impose on four-port circuits. These characteristics, as will be seen later, apply directly to the design of directional couplers. Contrary to three-port networks, four-port networks can be lossless (no resistive elements), reciprocal, and matched at all ports. Under the perfect match condition, a three-port network would have

$$S_{ii} = 0 \tag{8.22}$$

where $i = 1, 2, 3, 4$. For the reciprocity, the following condition is satisfied:

$$S_{ij} = S_{ji} \tag{8.23}$$

where $i, j = 1, 2, 3, 4$, while under the lossless criterion,

$$\begin{aligned} &\sum_{n=1}^{4} |S_{ni}|^2 = 1 \\ &\sum_{n=1}^{4} S_{ni} S_{nj}^* = 0; \quad i \neq j \end{aligned} \tag{8.24}$$

Applying the conditions in (8.22)–(8.24) yields the S-matrix for a lossless, reciprocal, matched four-port network as

$$[S] = \begin{bmatrix} 0 & 0 & S_{13} & S_{14} \\ 0 & 0 & S_{23} & S_{24} \\ S_{13} & S_{23} & 0 & 0 \\ S_{14} & S_{24} & 0 & 0 \end{bmatrix} \tag{8.25}$$

Examination of (8.25) reveals several facts that are useful for circuit design. The terminal or reference plane can be chosen for port 1 at an appropriate location along the terminating transmission line so that S_{14} is a real number; that is,

$$S_{14} = k_1 \tag{8.26}$$

where k_1 is real and positive. Similarly, the terminal plane at port 3 can be selected to produce

$$S_{13} = jk_2 \tag{8.27}$$

where k_2 is also real and positive. Combining (8.26) and (8.27), making use of (8.24) and (8.25), gives

$$k_1^2 + k_2^2 = 1 \tag{8.28}$$

Likewise, the terminal plane at port 2 can also be selected so that S_{23} is a real number. We can then derive that

$$S_{23} = S_{14} = k_1 \tag{8.29}$$

It is easy to prove that

$$S_{24} = S_{13} = jk_2 \tag{8.30}$$

The S-matrix of a lossless, reciprocal, matched four-port network then becomes

$$[S] = \begin{bmatrix} 0 & 0 & jk_2 & k_1 \\ 0 & 0 & k_1 & jk_2 \\ jk_2 & k_1 & 0 & 0 \\ k_1 & jk_2 & 0 & 0 \end{bmatrix} \tag{8.31}$$

Examination of (8.31) shows that a signal entering port 1 would travel to port 3 ($S_{31} \neq 0$) and port 4 ($S_{41} \neq 0$), but not to port 2 ($S_{21} = 0$), which is a unique property of directional couplers. A lossless, reciprocal, matched four-port network, whose S-parameters are given in (8.31), therefore represents an ideal four-port directional coupler. The transmission or coupling coefficients k_1 and k_2 of the coupler are related by (8.28). It is noted from S_{31} and S_{41} that the resultant phase difference between signals from 1 to 3 and 1 to 4 is 90°. This 90° phase difference, however, is not a unique property of directional couplers – depending on the design, in general, and the selection of terminal planes, in particular as described earlier, different phase can be obtained (e.g., 180°).

8.2 DIRECTIONAL COUPLERS

8.2.1 Fundamentals of Directional Couplers

Figure 8.6 shows some schematics for four-port dual and three-port single directional couplers. The ports are normally identified as the input, through (or direct), coupled, and decoupled (or isolated) ports. For instance, assuming the input port is 1, the through, coupled and decoupled ports are 4, 3 and 2, respectively, based on the schematics shown in Figure 8.6(a) and (b). Similarly, the through, coupled and decoupled ports, with respect to the input port 4, are 1, 2 and 3, respectively. For a single directional coupler as shown in Figure 8.6(c) and (d) with port 1 as the input port, the through and coupled ports are 2 and 3, respectively. The operation of dual directional couplers, assuming port 1 is the input port, is as follows. As a signal enters port 1 (input port), most of the signal would travel to port 4 (through port) and a small portion of the signal would go to port 3 (coupled port), while (ideally) none of the signal would appear at port 2 (decoupled port). For a single directional coupler, most of a signal entering port 1 (input port) would appear at port 2 (through port) while a fraction of it would go to port 3 (coupled port).

Directional couplers are characterized by four main parameters: insertion loss (IL), coupling, directivity, and isolation. They are also characterized by other typical parameters for RF passive components such as voltage standing wave ratio (VSWR) or return loss at the input, through and coupled ports, operating bandwidth, and variation of the coupling over the operating frequency range. Let P_i, P_t, P_c, and P_d to be the

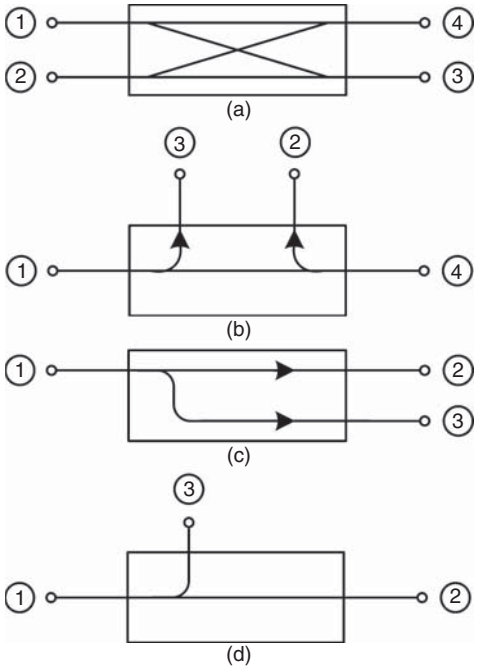


Figure 8.6. Schematics of dual (a, b) and single (c, d) directional couplers.

power at the input port (incident or input power), output power at the through port (through power), output power at the coupled port (coupled power), and output power at the decoupled port (decoupled power), respectively. The insertion loss, coupling (C), directivity (D), and isolation (I) in decibel (dB) are defined as

$$\mathrm{IL} = 10\log\left(\frac{P_t}{P_i}\right) \tag{8.32}$$

$$C = 10\log\left(\frac{P_c}{P_i}\right) \tag{8.33}$$

$$D = 10\log\left(\frac{P_d}{P_c}\right) \tag{8.34}$$

$$I = 10\log\left(\frac{P_d}{P_i}\right) \tag{8.35}$$

The directivity measures how well a coupler can isolate the coupled and decoupled ports. For ideal couplers, there is no power at the decoupled port, resulting in infinite directivity. The isolation indicates the degree of isolation between the input and decoupled ports. We can derive using (8.33)–(8.35) the relation between the coupling, isolation and directivity as

$$I = D + C \tag{8.36}$$

which shows that the isolation in a coupler is equal to the sum of its coupling and directivity in decibels.

In normal couplers, most of the input signal appears at the through port while only a fraction of it appears at the coupled port. However, there are special couplers, in which the input power splits equally with phase difference at the through and coupled ports – for instance, ring hybrid (or rat-race hybrid) and quadrature hybrid with 180° and 90° out-of-phase between the through and coupled ports, respectively. As mentioned earlier, Eq. (8.31) describes the scattering matrix of an ideal four-port directional coupler.

8.2.2 Parallel-Coupled Directional Couplers

Directional couplers based on the coupling between two or more parallel transmission lines are some of the most basic couplers that find popular use in RF circuits. In principle, there are two kinds of coupling in parallel transmission lines: edge- and broadside-coupling. Edge-coupling is the coupling along the edge of the coupled lines while broadside-coupling is the coupling across the full or partial width of the coupled lines. Broadside-coupling is suitable for weak, moderate and strong coupling and requires the use of multiple metal layers which are readily available in CMOS structures. On the other hand, edge-coupling needs only a single metal layer and is applicable only for weak and moderate coupling. This section presents an analysis for a very simple directional coupler consisting of two (identical and hence symmetrical) transmission lines in parallel. This serves not only as an example for studying the characteristics of directional couplers but also as a basis for the analysis of other couplers possessing structural symmetry.

8.2.2.1 Even- and Odd-Mode Analysis. The analysis of parallel-coupled directional couplers is based on the even- and odd-mode formulation [1]. The even- and odd-mode analysis can simplify significantly the analysis of RF circuits, in general, and passive RF circuits, in particular, that possess (physical) symmetry with respect to a certain plane in the circuits. We illustrate the overall analysis by considering an RF network as shown in Figure 8.7(a). We assume that the network has $2N$ ports that are symmetrical with respect to the central symmetry plane. The analysis is divided into three steps. In the first step, we apply equal potentials to each port of a symmetrical-port pair as shown in Figure 8.7(b). These excitations result in an open circuit, also called magnetic wall (MW), along the symmetry plane. The network is referred to as operating under the

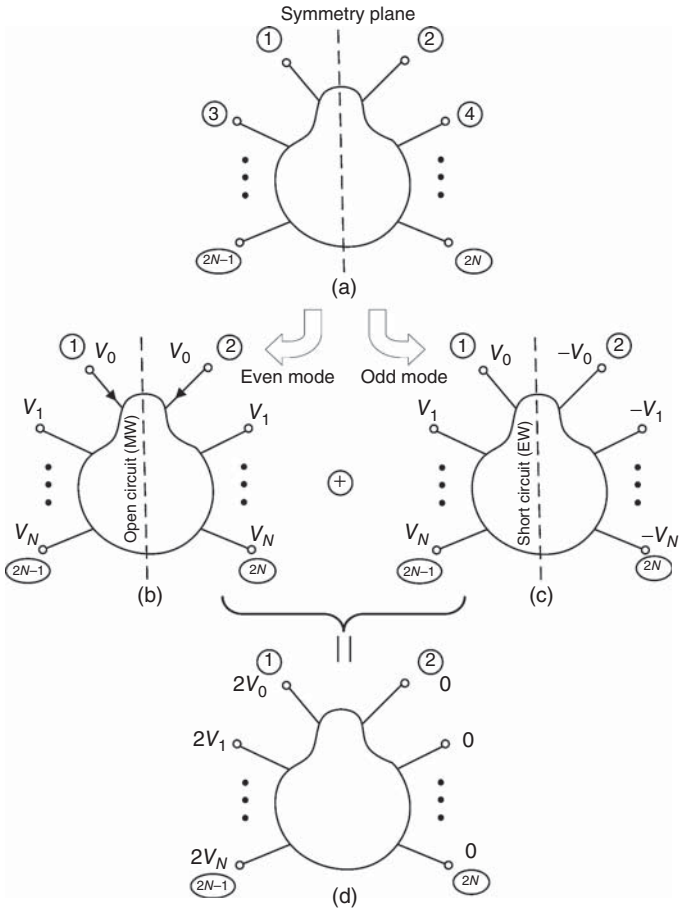


Figure 8.7. Even- and odd-mode analysis formulation: (a) symmetrical RF network, (b) even-mode operation, (c) odd-mode operation, and (d) superimposition of even- and odd-mode for final results.

"even mode." In the second step, we apply opposite potentials to each port of a symmetrical pair, as shown in Figure 8.7(c), which results in a short circuit, also called electrical wall (EW) at the symmetry plane. Under this condition, the network is operated under the "odd mode." In the final step, the even- and odd-mode structures are superimposed, as shown in Figure 8.7(d), and the previous even- and odd-mode analysis results are combined to derive the final results for the RF network.

8.2.2.2 Analysis of Parallel-Coupled Directional Couplers. Figure 8.8 shows a directional coupler formed by two parallel-coupled transmission lines. The two transmission lines are symmetrical with respect to the central plane AA. We assume that the terminating impedance at each port or the characteristic impedance of a (single) transmission line connected at each port (terminating transmission line) is Z_o. As outlined in the general even- and odd-mode analysis, the analysis proceeds with separate even-mode analysis and odd-mode analysis which are then combined to produce the final result.

Even-Mode Analysis

In the even-mode analysis, we apply the same potential $A_1/2$ to ports 1 and 3 and $A_2/2$ to ports 2 and 4. Figure 8.9(a) shows the coupler under the even-mode excitation which results in an open circuit or a MW along the symmetry plane. Note that $a_1^e = a_3^e = A_1/2$ and $a_2^e = a_4^e = A_2/2$; for simplicity, A_1 and A_2 are typically set to 1. Due to the open circuit, the two coupled transmission lines function as two isolated single transmission lines – each with an open-circuit plane on one side – as shown in Figure 8.9(b). These separate uncoupled transmission lines between ports 1, 2 and 3, 4 are shown in Figure 8.9(c) and (d). Z_{oe} is the characteristic impedance of the single transmission line under the even-mode operation (i.e., with an open circuit on one side of the line) which is commonly known as the even-mode characteristic impedance of coupled lines; β_e is the even-mode phase constant; and $\theta_e = 2\pi/\lambda_e = \lambda_{\text{air}}/\sqrt{\varepsilon^e_{\text{reff}}}$, with λ_e and λ_{air} being the even-mode wavelength and wavelength in air, respectively, is the electrical length of the single transmission line under the even-mode operation or the even-mode electrical length of the coupled lines. The scattering-matrix equations

Figure 8.8. Parallel-coupled directional coupler. If 1: input, then 2: through, 3: coupled, and 4: decoupled.

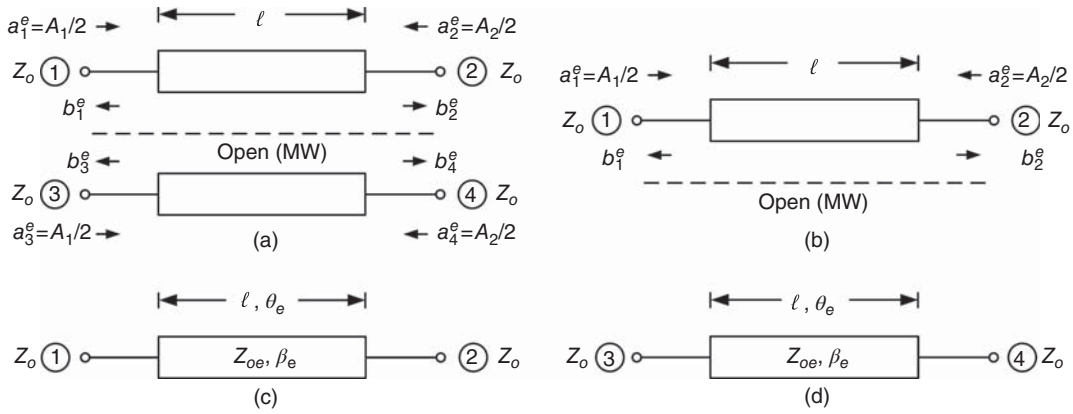


Figure 8.9. Coupler (a) and its equivalence (b), (c), (d) under even-mode excitation. *a*'s and *b*'s represent the respective incident and reflected voltage waves. ℓ is the physical length.

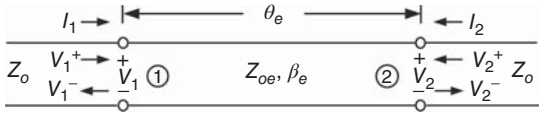


Figure 8.10. Single transmission line.

for these (two-port) transmission lines are

$$\begin{bmatrix} b_1^e \\ b_2^e \end{bmatrix} = \begin{bmatrix} S_{11}^e & S_{12}^e \\ S_{21}^e & S_{22}^e \end{bmatrix} \begin{bmatrix} a_1^e \\ a_2^e \end{bmatrix} \tag{8.37}$$

and

$$\begin{bmatrix} b_3^e \\ b_4^e \end{bmatrix} = \begin{bmatrix} S_{11}^e & S_{12}^e \\ S_{21}^e & S_{22}^e \end{bmatrix} \begin{bmatrix} a_3^e \\ a_4^e \end{bmatrix} \tag{8.38}$$

making use of the fact that these transmission lines have the same *S*-parameters. These *S*-parameters can be derived considering a single transmission line as shown in Figure 8.10.

The reflected voltages at ports 1 and 2 can be written as

$$V_1^- = V_2^+ e^{-j\theta_e} \tag{8.39}$$

$$V_2^- = V_1^+ e^{-j\theta_e} \tag{8.40}$$

The parameters of the transmission line's $ABCD$ matrix can be obtained with the help of (8.39) and (8.40) as

$$A = \left.\frac{V_1}{V_2}\right|_{I_2=0} = D = \frac{V_1^+ + V_1^-}{V_2^+ + V_2^-} = \frac{V_2^- e^{j\theta_e} + V_2^+ e^{-j\theta_e}}{2V_2^+} = \frac{e^{j\theta_e} + e^{-j\theta_e}}{2} = \cos\theta_e \tag{8.41}$$

making use of $V_2^+ = V_2^-$ for open circuit at port 2 and the symmetry property,

$$B = \left.\frac{V_1}{-I_2}\right|_{V_2=0} = Z_o' \frac{V_1^+ + V_1^-}{2V_2^-} = Z_o' \frac{V_2^- e^{j\theta_e} + V_2^+ e^{-j\theta_e}}{2V_2^-} = Z_o' \frac{e^{j\theta_e} - e^{-j\theta_e}}{2} = jZ_o' \sin\theta_e \tag{8.42}$$

utilizing $V_2^+ = -V_2^-$ for short circuit at port 2, $I_2 = I_2^+ + I_2^-$ and $Z_o' = V_2^+/I_2^+ = -V_2^-/I_2^-$, and

$$C = \left.\frac{I_1}{V_2}\right|_{I_2=0} = \frac{V_1^+ - V_1^-}{2Z_o' V_2^+} = \frac{V_1^+ - V_1^-}{2Z_o' V_2^+} = \frac{j\sin\theta_e}{Z_o'} \tag{8.43}$$

making use of $I_1 = I_1^+ + I_1^-$ and $Z_o' = V_1^+/I_1^+ = -V_1^-/I_1^-$. Applying the conversion formulas in Table 7.1 and using the symmetry and reciprocity properties then gives

$$\begin{aligned} S_{11}^e = S_{22}^e &= \frac{A + B/Z_o - CZ_o - D}{A + B/Z_o + CZ_o + D} = \frac{j\left(\frac{Z_{oe}}{Z_o} - \frac{Z_o}{Z_{oe}}\right)\sin\theta_e}{2\cos\theta_e + j\left(\frac{Z_{oe}}{Z_o} + \frac{Z_o}{Z_{oe}}\right)\sin\theta_e} = \frac{j(Z_{oe}^2 - Z_o^2)\sin\theta_e}{2Z_oZ_{oe}\cos\theta_e + j(Z_{oe}^2 + Z_o^2)\sin\theta_e} \\ S_{21}^e = S_{12}^e &= \frac{2}{A + B/Z_o + CZ_o + D} = \frac{2}{2\cos\theta_e + j\left(\frac{Z_{oe}}{Z_o} + \frac{Z_o}{Z_{oe}}\right)\sin\theta_e} = \frac{2Z_oZ_{oe}}{2Z_oZ_{oe}\cos\theta_e + j(Z_{oe}^2 + Z_o^2)\sin\theta_e} \end{aligned} \tag{8.44}$$

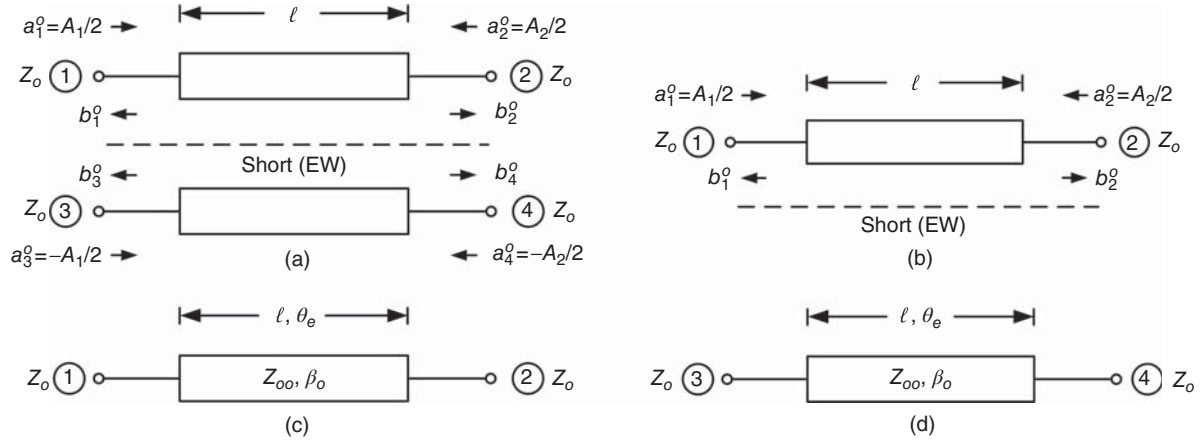


Figure 8.11. Coupler (a) and its equivalence (b), (c), (d) under odd-mode excitation.

Odd-Mode Analysis

In the odd-mode analysis, we apply the opposite potential $\pm A_1/2$ to ports 1 and 3 and $\pm A_2/2$ to ports 2 and 4. Figure 8.11(a) shows the coupler under the odd-mode excitation which results in a short circuit or an EW along the symmetry plane. Note that $a_1^o = -a_3^o = A_1/2$ and $a_2^o = a_4^o = -A_2/2$. Due to the short circuit, the two coupled transmission lines function as two isolated single transmission lines – each with a short-circuit plane on one side – as shown in Figure 8.11(b). These separate uncoupled transmission lines between ports 1, 2 and 3, 4 are shown in Figure 8.11(c) and (d) where Z_{oo} is the characteristic impedance of the single transmission line under the odd-mode operation (i.e., with a short circuit one on one side of the line) which is commonly known as the odd-mode characteristic impedance of coupled lines, and $\theta_o = 2\pi/\lambda_o = \lambda_{\text{air}}/\sqrt{\varepsilon_{\text{reff}}^o}$, with λ_o and λ_{air} being the odd-mode wavelength and wavelength in air, respectively, is the electrical length of the single transmission line under the odd-mode operation or the odd-mode electrical length of the coupled lines. It is recalled that, for inhomogeneous coupled transmission lines such as microstrip lines, $\varepsilon_{\text{reff}}^o \neq \varepsilon_{\text{reff}}^e$ and hence $\theta_o \neq \theta_e$. The S-matrix equations for these transmission lines under the odd mode are obtained as

$$\begin{bmatrix} b_1^o \\ b_2^o \end{bmatrix} = \begin{bmatrix} S_{11}^o & S_{12}^o \\ S_{21}^o & S_{22}^o \end{bmatrix} \begin{bmatrix} a_1^o \\ a_2^o \end{bmatrix} \tag{8.45}$$

and

$$\begin{bmatrix} b_3^o \\ b_4^o \end{bmatrix} = \begin{bmatrix} S_{11}^o & S_{12}^o \\ S_{21}^o & S_{22}^o \end{bmatrix} \begin{bmatrix} a_3^o \\ a_4^o \end{bmatrix} \tag{8.46}$$

where the S-parameters can be obtained from (8.44) with Z_{oo} and θ_o replacing Z_{oe} and θ_e, respectively, as

$$\begin{aligned} S_{11}^o = S_{22}^o &= \frac{j(Z_{oo}^2 - Z_o^2)\sin\theta_o}{2Z_oZ_{oo}\cos\theta_o + j(Z_{oo}^2 + Z_o^2)\sin\theta_o} \\ S_{12}^o = S_{21}^o &= \frac{2Z_oZ_{oo}}{2Z_oZ_{oo}\cos\theta_o + j(Z_{oo}^2 + Z_o^2)\sin\theta_o} \end{aligned} \tag{8.47}$$

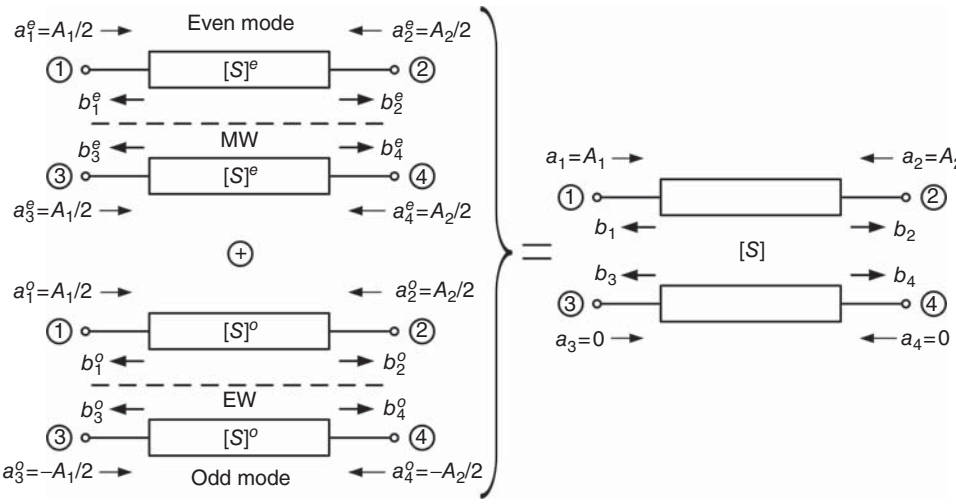


Figure 8.12. Superimposition of even and odd modes. $[S]^e$, $[S]^o$, and $[S]$ represent the S-matrix for the single transmission line under even-, odd-mode excitation, and for the two coupled lines, respectively.

Superimposition of Even and Odd Modes

In the final step, we superimpose the even and odd modes as illustrated in Figure 8.12 and obtain

$$\begin{aligned} a_1 &= a_1^e + a_1^o = A_1 \\ a_2 &= a_2^e + a_2^o = A_2 \\ a_3 &= a_3^e + a_3^o = 0 \\ a_4 &= a_4^e + a_4^o = 0 \end{aligned} \tag{8.48}$$

$$\begin{aligned} b_1 &= b_1^e + b_1^o \\ b_2 &= b_2^e + b_2^o \\ b_3 &= b_3^e + b_3^o \\ b_4 &= b_4^e + b_4^o \end{aligned} \tag{8.49}$$

It is noted that:

$$\begin{aligned} a_1 &= 2a_1^e = 2a_1^o = 2a_3^e = -2a_3^o \\ a_2 &= 2a_2^e = 2a_2^o = 2a_4^e = -2a_4^o \end{aligned} \tag{8.50}$$

The S-matrix equations for the coupler can be derived from (8.49), (8.37), (8.38), (8.45), (8.46), and (8.50) as

$$\begin{aligned} b_1 &= S_{11}^e a_1^e + S_{12}^e a_2^e + S_{11}^o a_1^o + S_{12}^o a_2^o = S_{11}^e \frac{a_1}{2} + S_{12}^e \frac{a_2}{2} + S_{11}^o \frac{a_1}{2} + S_{12}^o \frac{a_2}{2} = \frac{1}{2}(S_{11}^e + S_{11}^o)a_1 + \frac{1}{2}(S_{12}^e + S_{12}^o)a_2 \\ b_2 &= S_{21}^e a_1^e + S_{22}^e a_2^e + S_{21}^o a_1^o + S_{22}^o a_2^o = S_{21}^e \frac{a_1}{2} + S_{22}^e \frac{a_2}{2} + S_{21}^o \frac{a_1}{2} + S_{22}^o \frac{a_2}{2} = \frac{1}{2}(S_{21}^e + S_{21}^o)a_1 + \frac{1}{2}(S_{22}^e + S_{22}^o)a_2 \\ b_3 &= S_{11}^e a_3^e + S_{12}^e a_4^e + S_{11}^o a_3^o + S_{12}^o a_4^o = S_{11}^e \frac{a_1}{2} + S_{12}^e \frac{a_2}{2} - S_{11}^o \frac{a_1}{2} - S_{12}^o \frac{a_2}{2} = \frac{1}{2}(S_{11}^e - S_{11}^o)a_1 + \frac{1}{2}(S_{12}^e - S_{12}^o)a_2 \\ b_4 &= S_{21}^e a_3^e + S_{22}^e a_4^e + S_{21}^o a_3^o + S_{22}^o a_4^o = S_{21}^e \frac{a_1}{2} + S_{22}^e \frac{a_2}{2} - S_{21}^o \frac{a_1}{2} - S_{22}^o \frac{a_2}{2} = \frac{1}{2}(S_{21}^e - S_{21}^o)a_1 + \frac{1}{2}(S_{22}^e - S_{22}^o)a_2 \end{aligned} \tag{8.51}$$

or,

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} S_{11}^e + S_{11}^o & S_{12}^e + S_{12}^o \\ S_{21}^e + S_{21}^o & S_{22}^e + S_{22}^o \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{8.52}$$

$$\begin{bmatrix} b_3 \\ b_4 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} S_{11}^e - S_{11}^o & S_{12}^e - S_{12}^o \\ S_{21}^e - S_{21}^o & S_{22}^e - S_{22}^o \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{8.53}$$

Hence,

$$\begin{aligned} S_{11} &= \frac{1}{2}(S_{11}^e + S_{11}^o) \\ S_{12} &= \frac{1}{2}(S_{12}^e + S_{12}^o) \\ S_{21} &= \frac{1}{2}(S_{21}^e + S_{21}^o) \\ S_{22} &= \frac{1}{2}(S_{22}^e + S_{22}^o) \\ S_{31} &= \frac{1}{2}(S_{11}^e - S_{11}^o) \\ S_{32} &= \frac{1}{2}(S_{12}^e - S_{12}^o) \\ S_{41} &= \frac{1}{2}(S_{21}^e - S_{21}^o) \\ S_{42} &= \frac{1}{2}(S_{22}^e - S_{22}^o) \end{aligned} \tag{8.54}$$

from which, we can derive, making use of (8.44), (8.47) and $S_{11} = S_{33}$, $S_{22} = S_{44}$ from the symmetry property,

$$S_{11} = S_{22} = S_{33} = S_{44} = j\frac{1}{2}\left[\frac{(Z_{oe}^2 - Z_o^2)\sin\theta_e}{2Z_oZ_{oe}\cos\theta_e + j(Z_{oe}^2 + Z_o^2)\sin\theta_e} + \frac{(Z_{oo}^2 - Z_o^2)\sin\theta_o}{2Z_oZ_{oo}\cos\theta_o + j(Z_{oo}^2 + Z_o^2)\sin\theta_o}\right] \tag{8.55}$$

Equation (8.55) can be employed to derive some interesting and useful results for the coupler design. To that end, we assume the coupler is perfectly matched at its four ports; that is, $S_{ii} = 0$, $i = 1, 2, 3, 4$. This leads to, using (8.47),

$$\begin{aligned} &(Z_{oe}^2 - Z_o^2)\sin\theta_e[2Z_oZ_{oo}\cos\theta_o + j(Z_{oo}^2 + Z_o^2)\sin\theta_o] \\ &+ (Z_{oo}^2 - Z_o^2)\sin\theta_o[2Z_oZ_{oe}\cos\theta_e + j(Z_{oe}^2 + Z_o^2)\sin\theta_e] = 0 \end{aligned} \tag{8.56}$$

Equation (8.56) is satisfied when its real and imaginary parts are equal to zero, which results in

$$2Z_oZ_{oo}(Z_{oe}^2 - Z_o^2)\sin\theta_e\cos\theta_o + 2Z_oZ_{oe}(Z_{oo}^2 - Z_o^2)\sin\theta_o\cos\theta_e = 0 \tag{8.57}$$

and

$$(Z_{oe}^2 - Z_o^2)(Z_{oo}^2 + Z_o^2)\sin\theta_e\sin\theta_o + (Z_{oe}^2 + Z_o^2)(Z_{oo}^2 - Z_o^2)\sin\theta_e\sin\theta_o = 0 \tag{8.58}$$

respectively. We can then write from (8.58), assuming either θ_e or θ_o is different from $n\pi$, where $n = 0, 1, 2, \ldots$,

$$(Z_{oe}^2 - Z_o^2)(Z_{oo}^2 + Z_o^2) = -(Z_{oe}^2 + Z_o^2)(Z_{oo}^2 - Z_o^2) = 0 \tag{8.59}$$

from which, we obtain

$$Z_o^2 = Z_{oe}Z_{oo} \tag{8.60}$$

which is an important design equation. Substituting (8.60) into (8.57), we get, after some manipulations,

$$\sin\theta_e \cos\theta_o = \sin\theta_o \cos\theta_e \tag{8.61}$$

In order for (8.61) to be satisfied for any value of θ_e or θ_o including $\pi/2$, except $n\pi$, there must be[2]

$$\theta_e = \theta_o \tag{8.62}$$

which implies that the coupled lines must be homogenous having equal even- and odd-mode velocities – for example, the strip line discussed in Chapter 4. Equations (8.60) and (8.62) show that for two perfectly matched symmetrical coupled lines in an homogenous medium, the even- and odd-mode characteristic impedances are related by $Z_o = \sqrt{Z_{oe}Z_{oo}}$. These equations also imply that two inhomogeneous symmetric coupled lines such as microstrip coupled lines, for which the even- and odd-mode electrical lengths are different, can never be perfectly matched.

We now rewrite S_{21}^e, S_{12}^e and S_{11}^e, S_{22}^e in (8.44) by multiplying the numerator and denominator with Z_{oo} and letting $\theta_e = \theta_o = \theta$ according to (8.62) as

$$S_{21}^e = S_{12}^e = \frac{2Z_oZ_{oe}Z_{oo}}{2Z_oZ_{oe}Z_{oo}\cos\theta + j(Z_{oe}^2 + Z_o^2)Z_{oo}\sin\theta} \tag{8.63}$$

and

$$S_{11}^e = S_{22}^e = \frac{jZ_{oo}(Z_{oe}^2 - Z_o^2)\sin\theta}{2Z_oZ_{oe}Z_{oo}\cos\theta + jZ_{oo}(Z_{oe}^2 + Z_o^2)\sin\theta} \tag{8.64}$$

which, upon applying (8.60), become

$$S_{21}^e = S_{12}^e = \frac{2Z_o}{2Z_o\cos\theta + j(Z_{oe} + Z_{oo})\sin\theta} \tag{8.65}$$

and

$$S_{11}^e = S_{22}^e = \frac{j(Z_{oe} - Z_{oo})\sin\theta}{2Z_o\cos\theta + j(Z_{oe} + Z_{oo})\sin\theta} \tag{8.66}$$

Similarly, S_{21}^o, S_{12}^o and S_{11}^o, S_{22}^o in (8.47) can be rewritten, upon multiplying and dividing by Z_{oe} and applying (8.60), as

$$S_{21}^o = S_{12}^o = \frac{2Z_o}{2Z_o\cos\theta + j(Z_{oe} + Z_{oo})\sin\theta} \tag{8.67}$$

and

$$S_{11}^o = S_{22}^o = \frac{j(Z_{oo} - Z_{oe})\sin\theta}{2Z_o\cos\theta + j(Z_{oo} + Z_{oe})\sin\theta} \tag{8.68}$$

which show that $S_{21}^e = S_{12}^e = S_{21}^o = S_{12}^o$ and $S_{11}^e = S_{22}^e = -S_{11}^o = -S_{22}^o$.

[2]If it is assumed that either θ_e or θ_o is different from $(2n+1)\pi/2 (n = 0, 1, 2, \ldots)$ then (8.59) can be divided by $\cos\theta_o \cos\theta_e$, which results in $\tan\theta_e = \tan\theta_o$ and hence $\theta_e = \theta_o$. This, however, excludes the length of the coupled lines from being $\pi/2$.

Substituting (8.65) and (8.67) into S_{21} in (8.54) gives

$$S_{21} = S_{12} = \frac{2Z_o}{2Z_o \cos\theta + j(Z_{oe} + Z_{oo}) \sin\theta} \tag{8.69}$$

Equations (8.65), (8.67), and (8.69) demonstrate that the insertion loss between ports 1 and 2 of two parallel-coupled transmission lines is the same as that for each of the two isolated (even- and odd-mode) single transmission lines. The S-parameter S_{31} in (8.54) becomes, after applying (8.66) and (8.68),

$$S_{31} = \frac{j(Z_{oe} - Z_{oo}) \sin\theta}{2Z_o \cos\theta + j(Z_{oe} + Z_{oo}) \sin\theta} \tag{8.70}$$

Letting

$$C = \frac{Z_{oe} - Z_{oo}}{Z_{oe} + Z_{oo}} \tag{8.71}$$

we can write

$$1 - C^2 = 1 - \left(\frac{Z_{oe} - Z_{oo}}{Z_{oe} + Z_{oo}}\right)^2 = \frac{4Z_{oe}Z_{oo}}{(Z_{oe} + Z_{oo})^2} \tag{8.72}$$

and hence, upon using (8.60),

$$\sqrt{1 - C^2} = \frac{2Z_o}{Z_{oe} + Z_{oo}} \tag{8.73}$$

Using (8.71) and (8.73) to rewrite (8.69) and (8.70), and applying the symmetry property ($S_{11} = S_{33},\ S_{22} = S_{44},\ S_{21} = S_{43}$) and reciprocity property ($S_{ij} = S_{ji},\ i, j = 1, 2, 3, 4$), we obtain

$$S_{21} = S_{12} = S_{34} = S_{43} = \frac{\sqrt{1 - C^2}}{\sqrt{1 - C^2} \cos\theta + j \sin\theta} \tag{8.74}$$

$$S_{31} = S_{13} = S_{24} = S_{42} = \frac{jC \sin\theta}{\sqrt{1 - C^2} \cos\theta + j \sin\theta} \tag{8.75}$$

$$S_{41} = S_{14} = S_{23} = S_{32} = 0 \tag{8.76}$$

At the design frequency at which $\theta = \pi/2$, or the physical length of the coupled lines is a quarter-wavelength ($\ell = \lambda/4$), the transmission between ports 1 and 3 characterized by S_{31} reaches a maximum value of C as seen from (8.75). The constant C, as defined in (8.71) and commonly referred to as the coupling factor, therefore represents the maximum coupling between the two coupled lines at the design frequency. The S-parameters of the coupler at the design frequency can be determined from the perfect match and (8.74)–(8.76) as

$$\begin{aligned} S_{11} &= S_{22} = S_{33} = S_{44} = 0 \\ S_{21} &= S_{12} = S_{34} = S_{43} = -j\sqrt{1 - C^2} \\ S_{31} &= S_{13} = S_{24} = S_{42} = C \\ S_{41} &= S_{14} = S_{23} = S_{32} = 0 \end{aligned} \tag{8.77}$$

As can be seen in (8.77), the output voltage at port 2 is 90° out-of-phase with respect to the input voltage entering port 1, which sets the property of the parallel-coupled coupler as a 90° hybrid or quadrature coupler.

8.2.2.3 Design of Parallel-Coupled Directional Couplers. In typical design of directional couplers, the coupling is given as a design specification and, hence, the parallel-coupled coupler can be designed according to its even- and odd-mode characteristic impedances derived from (8.60) and (8.71) as

$$Z_{oe} = Z_o\sqrt{\frac{1+C}{1-C}} \tag{8.78}$$

$$Z_{oo} = Z_o\sqrt{\frac{1-C}{1+C}} \tag{8.79}$$

or, in term of the coupling in dB (C_{dB}):

$$Z_{oe} = Z_o\sqrt{\frac{1+10^{C_{\text{dB}}/20}}{1-10^{C_{\text{dB}}/20}}} \tag{8.80}$$

$$Z_{oo} = Z_o\sqrt{\frac{1-10^{C_{\text{dB}}/20}}{1+10^{C_{\text{dB}}/20}}} \tag{8.81}$$

Once Z_{oe} and Z_{oo} are known, the physical dimensions of the coupled lines (e.g., the width of the strips and spacing between them for parallel-coupled microstrip lines) can be determined. For homogeneous coupled lines, the physical length is obtained as $\ell = \lambda/4$. For inhomogeneous transmission lines, $\theta \simeq (\theta_e + \theta_o)/2$ or $\ell \simeq (\lambda_e + \lambda_o)/8$ can be used as an initial length in the design.

As can be inferred from (8.77), there is no output signal at the decoupled port 4 for input signal at port 1, which results in infinite directivity. Therefore, (ideal) directional couplers realized using two parallel-coupled homogeneous transmission lines such as strip lines, for which $\theta_e = \theta_o$ or $\varepsilon^e_{\text{reff}} = \varepsilon^o_{\text{reff}} = \varepsilon_r$, where ε_r is the relative dielectric constant of the dielectric, have infinite directivity theoretically which is always desirable. It is noted that homogeneous transmission lines such as strip lines can be readily formed in RFIC structures utilizing their inherent multi metal layers. On the other hand, when the coupled transmission lines are inhomogeneous such as microstrip lines, $\varepsilon^e_{\text{reff}} \neq \varepsilon^o_{\text{reff}}$ and hence $\theta_e \neq \theta_o$, resulting in $S_{41} \neq 0$, which leads to finite directivity. Therefore, with respect to directivity, directional couplers based on homogeneous parallel-coupled lines are more desirable than their counterparts using inhomogeneous parallel-coupled lines. The directivity for inhomogeneous parallel-coupled couplers, however, can be enhanced by improving the match between the even- and odd-mode phase velocities, leading to $\varepsilon^e_{\text{reff}} \simeq \varepsilon^o_{\text{reff}}$, by using techniques such as dielectric overlay in microstrip lines, in which a dielectric having equal or different relative dielectric constant from the main dielectric underneath the metal strips is placed over the strips to make the surrounding environment close to homogeneous as much as possible. This dielectric overlay technique can be easily implemented in RFIC structures with their multiple dielectrics.

Figure 8.13 shows some possible implementations of parallel-coupled couplers using transmission lines. The coupled metal strips can have equal or different width. Unequal-width couplers may provide both coupling and impedance transformation. As shown in Figure 8.13, a coupler can be realized using parallel coupled lines on the same surface such as microstrip and strip lines or on different surfaces such as broadside-coupled microstrip lines. The former is based on edge-coupling and is suitable only for loose coupling (typically less than −3 dB) in order for the gap between the coupled lines to be realizable; for strong coupling (greater than −3 dB), the gap becomes too small that it may not be physically realizable. The latter achieves the coupling via broadside-coupling and is particularly suitable for tight coupling. It, however, can also be used for moderate and weak coupling (e.g., −10 dB) by adjusting the overlapping between the two strips. Parallel-coupled couplers employing a single section can work only over a narrow bandwidth, typically less than 10% depending on the specifications such as coupling variation versus frequency. To achieve a broader bandwidth using the same coupled-line configuration, multiple sections such as that shown in Figure 8.14 need to be used.

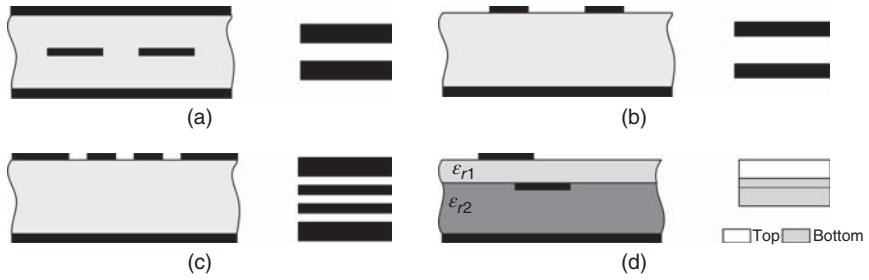


Figure 8.13. Parallel-coupled couplers (end and top view) implemented using transmission lines: (a) strip line, (b) microstrip line, (c) coplanar waveguide, and (d) broadside-coupled lines.

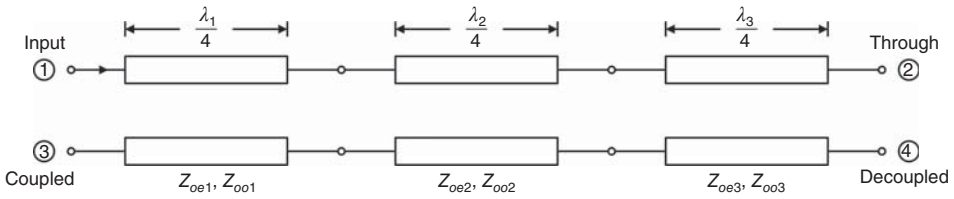


Figure 8.14. Three-section parallel-coupled coupler.

8.2.2.4 Lange Couplers. Directional couplers employing two parallel-coupled lines on the same surface suffer two important drawbacks: a narrow bandwidth and an unrealizable gap for tight coupling. Lange coupler [2] and its modified version known as the unfolded Lange coupler [3] overcome these constraints, making them attractive for broadband and tight-coupling design. Particularly, its geometry facilitates the compensation for the (undesired) unequal even- and odd-mode phase velocities encountered in inhomogeneous transmission lines such as microstrip lines and helps improve the operating bandwidth. Figure 8.15 shows the (conventional folded) Lange coupler and its unfolded version. They are essentially a quadrature coupler or hybrid having 90° phase difference between the through and coupled ports. Each has four parallel strips or fingers with one finger splitting into two halves (outer most finger halves) in the folded Lange coupler and is quarter-wavelength long. The interconnection between strips is typically done using air-bridges or cross-over wires. Similar to other parallel-coupled couplers, the coupling is based on the edge-coupling. However, in the Lange couplers, the coupling occurs along the edges of four fingers, which results in tight coupling and a very wide bandwidth. Coupling less than 3 dB and a bandwidth more than an octave can be

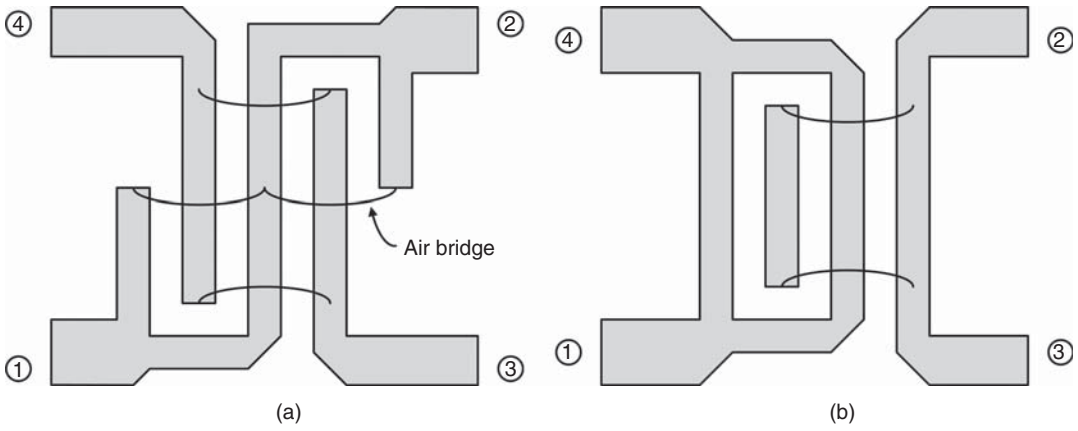


Figure 8.15. (a) Folded and (b) unfolded Lange couplers. The length of the coupled lines or fingers is $\lambda/4$. If 1: input, then 2: through, 3: coupled, and 4: decoupled.

achieved easily. The unfolded Lange coupler uses less number of interconnects and hence is easier to model and calculated accurately. The main problem of the Lange couplers, as with other parallel-coupled couplers, is the quarter-wavelength length, which takes much valuable space in RFIC chips, particularly in low RF range. Although meandering techniques can be used to shorten the overall length, sufficient space between meandered segments still needs to be maintained to avoid unwanted coupling between them, which makes the coupler's length not adequately small to be useful for RFIC implementation. In order for the Lange couplers to be useful for RFIC, the quarter-wavelength-dimension problem needs to be resolved.

Design of Lange Coupler

We assume that the four coupled lines used in the Lange coupler support only two modes (even and odd modes) as in two (symmetrical) parallel-coupled lines, instead of four modes generally occurring in four coupled lines. That is, the electrical properties of the four coupled lines are characterized in terms of the even- and odd-mode characteristic impedances and effective dielectric constants (or velocities). These are Z_{oei}, Z_{ooi}, $\varepsilon^e_{\text{reff}}$, and $\varepsilon^e_{\text{reff}}$, where $i = 1, 2, 3, 4$ corresponding to the ith line. We further assume that the four lines are identical having the same width and spacing between them, and the coupling occurs only between the two adjacent lines. Under these assumptions, the four coupled lines can be approximately characterized by the even- and odd-mode parameters of two "equivalent" parallel coupled lines. Assume the coupled lines are homogeneous (e.g., strip line), the even- and odd-mode phase velocities are equal, and the even- and odd-mode characteristic impedances of any two adjacent coupled lines in the four coupled lines can be approximately derived as [4]:

$$Z_{oe} \simeq \frac{4C_o - 3 + \sqrt{9 - 8C_o^2}}{2C_o\sqrt{(1 - C_o)/(1 + C_o)}} Z_o \tag{8.82}$$

and

$$Z_{oo} \simeq \frac{4 + 3 - \sqrt{9 - 8C_o^2}}{2C_o\sqrt{(1 + C_o)/(1 - C_o)}} Z_o \tag{8.83}$$

respectively, where Z_o is the characteristic impedance of the terminating line and

$$C_o \simeq \frac{3(Z_{oe}^2 - Z_{oo}^2)}{3(Z_{oe}^2 + Z_{oo}^2) + 2Z_{oe}Z_{oo}} \tag{8.84}$$

is the voltage coupling coefficient at the design center frequency. It is noted that the even- and odd-mode characteristic impedances in (8.82) and (8.83) are for every pair of two adjacent lines (assuming the pair is isolated from other nearby lines) of the four coupled lines, and they are different from those of the "equivalent" two parallel coupled lines of the four parallel coupled lines mentioned earlier. This is due to the fact that the surrounding of the coupled-line pair and the equivalent two parallel coupled lines are different, resulting in different even/odd-mode capacitance per unit length. For a given coupling (C_o) and terminating characteristic impedance (Z_o), the even- and odd-mode characteristic impedances of every two adjacent lines can be approximately calculated and, from which, the width and gap of the coupled lines, assuming they stand by themselves without any neighboring lines, can be estimated. These estimates can serve as good initial values for the design of the Lange coupler.

In general, the design of the Lange coupler, just like that for the parallel-coupled-line coupler described earlier, is relatively simple. In RFIC structures, however, the Lange coupler (as well as the parallel-coupled coupler and other RF components based on transmission lines) poses difficulty due to the fact that it does not lend itself to miniaturization needed in RFIC. Miniaturization of elements in a circuit is, in fact, a common issue for all RFIC that needs to be resolved in order to make the circuit useful for CMOS implementation. This issue is more pronounced for circuits based on transmission lines such as the Lange coupler whose length is a quarter-wavelength, which is relative long even at high frequencies. The following design implements multiple metal layers inherent in a CMOS structure to achieve a very compact Lange coupler suitable

for CMOS implementation [5]. This design also serves as an example to illustrate a design technique for achieving miniaturization for passive circuit elements and components employing transmission lines such as the parallel-coupled coupler. Figure 8.16 shows such a Lange coupler based on the unfolded configuration realized using multilayered broadside-coupled transmission lines. The broadside-coupled transmission lines facilitate not only significant size reduction through meandering, which is rather difficult with coupled lines having conductors on the same side of a dielectric substrate due to severe unwanted coupling between meandered sections too close to each other, but also tight coupling through coupling across entire or partial line width, which cannot be achieved with coupled lines having conductors on the same side that produce coupling along the edges of the lines.

A major consideration in the design of the CMOS broadside-coupled Lange coupler in particular, and CMOS broadside-coupled lines in general, is the treatment of the inhomogeneous and asymmetric nature of the coupled-line structure embedded in a multilayer dielectric. Furthermore, in order to obtain similar even- and odd-mode characteristic impedance for each finger pair in the multilayer broadside-coupled structure, different widths need to be used, owing to their differences in distance to the ground plane and in the oxide dielectric layers surrounding them. This further adds an asymmetric dimension along the broadside configuration.

The CMOS broadside-coupled Lange coupler design was carried out by individually calculating the effect of the ground plane (M1) on each finger. The coupling effect of the two adjacent fingers for the second (M4) and third (M3) fingers was also taken in account. It is evident from the cross-section of the structure, as shown in Figure 8.16, that the top finger (M5) is not particularly affected by the bottom two fingers (M3 and M2), and the bottom finger (M2) is not strongly affected by the top two fingers (M5 and M4). This fact was exploited to simplify the design by assuming the presence of only one finger (M4 or M3) adjacent to the top (M5) or bottom (M2) finger, respectively. A full wave electromagnetic (EM) analysis was then performed using IE3D. Table 8.1 shows the possible widths for each finger obtained by IE3D simulations. The values in parenthesis

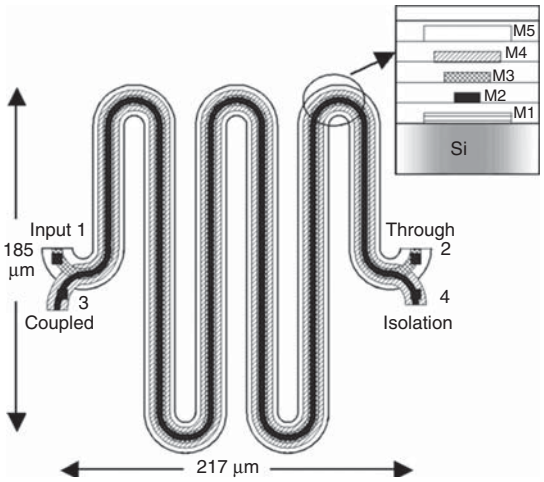


Figure 8.16. Layout of the unfolded broadside-coupled Lange coupler with cross-section shown in the inset. The four fingers are implemented on the top four metal layers (M2–M5). The bottom metal layer (M1) is used as the ground plane. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

TABLE 8.1. Coupling and Width Estimation from EM Analysis

Finger layer	Coupling layer	Width (μm)	Estimated coupling to adjacent layer (dB)
M5	M4	6–8 (7.8)	0.93
M4	M5, M3	4–7 (5.7)	0.6
M3	M4, M2	2–4 (3.4)	0.6
M2	M3	1–1.5 (1.0)	0.8

for the widths in Table 8.1 are the optimized numbers obtained after all four layers were incorporated in EM simulation and after the widths were adjusted for characteristic-impedance match and 90° phase difference conditions. Coupling (C) for each pair of adjacent fingers was calculated with another round of full wave EM analysis from S-parameters ($C = S_{ij}/S_{kj}$, where i, j, and k are port numbers). They indicate the tight coupling facilitated by the broadside topology. It must be noted that the coupling (as well as EM fields) for M4 and M3 is distributed unevenly between the two adjacent layers.

The four-finger broadside-coupled Lange coupler, shown in Figure 8.16, was fabricated on the TSMC 0.25-μm CMOS process [6]. The strip widths are indicated in Table 8.1, while the electrical length of each strip is estimated as 1238.5 μm based on the $\lambda/4$ length requirement. The distances between adjacent metal layers, including the ground-plane metal, are the same and dictated by the CMOS fabrication process. The structure was meandered in order to make it more compact. Circular corners were used instead of rectangular corners in order to avoid discontinuity and field crowding effects at high frequencies. The vias were connected through a meandering arc between alternate multilayers, that is, between M5–M3 and M4–M2, which could be easily de-embedded numerically as they introduce only a minor phase shift in the measured results. The fabricated structure is shown in Figure 8.17. The total chip size is just 217 × 185 μm demonstrating the advantage of using the multilayer broadside-coupled transmission lines.

For convenient two-port measurement, three different couplers were laid out for the two-port through, coupling, and isolation measurement. The results are shown in Figures 8.18–8.20. The measured results show a broadband performance and close concurrence with the simulated results. Measured performance exhibits around −3.1 dB through, −3.4 dB coupling, 11-dB isolation and more than 14 dB return loss across 15–35 GHz. The measured amplitude imbalance is about -3.1 ± 0.3 dB between 15 and 35 GHz, while the measured phase imbalance is $90 \pm 2°$ over the 25–35 GHz range. The excess loss is not significant and attributed mainly to the conductor losses of the fingers, especially those in the lower thin metal layers (M3 and M2), as well as to the nonideal nature of vias. It is obvious that the broadside coupled topology presents superior performance for the Lange coupler even on a lossy silicon substrate while occupying only a very small area.

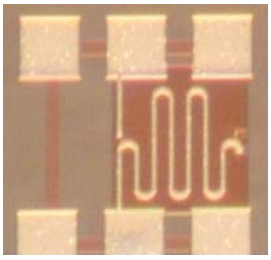


Figure 8.17. Die photograph of the fabricated Lange coupler. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

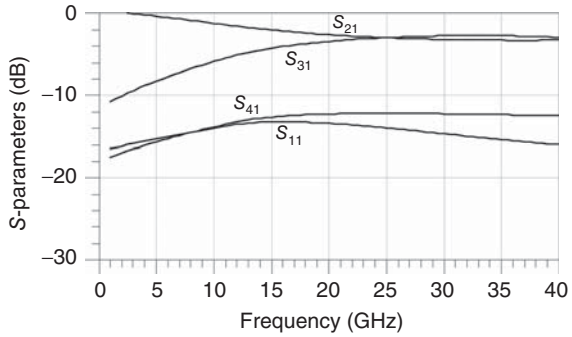


Figure 8.18. Simulated S-parameters of the Lange coupler. 1: input, 2: through, 3: coupled, and 4: isolated port. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

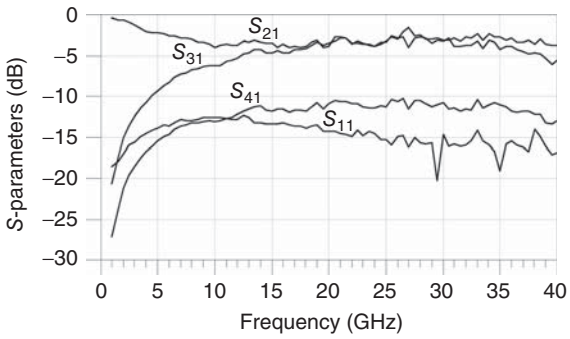


Figure 8.19. Measured *S*-parameters of the Lange coupler. 1: input, 2: through, 3: coupled, and 4: isolated port. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

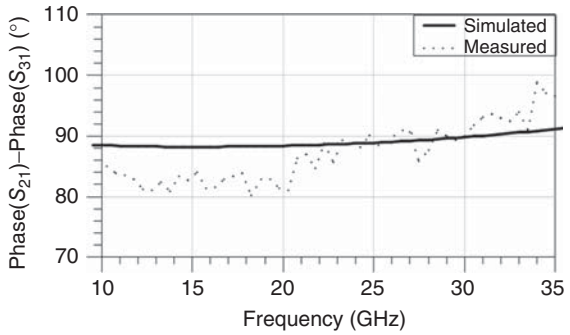


Figure 8.20. Measured and simulated phase difference between the through and coupled ports of the Lange coupler. 1: input, 2: through, 3: coupled, and 4: isolated port. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

8.3 HYBRIDS

We have learned that directional couplers are designed so that the signals at the two desired "through" and "coupled" output ports are related by their amplitudes via the "coupling" coefficient, but not on the phase. There is a particular class of couplers that relate not only the amplitude but also the phase between the output signals. This kind of couplers is known as "hybrid." Typical hybrids are designed to achieve 3-dB coupling with 90° or 180° phase difference between the output ports, referred to as 90° or quadrature hybrid and 180° hybrid, respectively. Well-known hybrids are the hybrid T (also known as magic T) and ring or rat-race hybrid (which is actually a special hybrid T), whose outputs have equal amplitude and in-phase or 180° out-of-phase, and branch-line directional coupler whose outputs are equal in amplitude and 90° out-of-phase.

8.3.1 Hybrid T

Hybrid T originates from the junction between different rectangular waveguides, from which bearing the name "hybrid-T junction" commonly known among microwave engineers. Figure 8.21 shows the hybrid T formed by the junction between waveguides along the E-plane (and hence E-plane arm) and H-plane (and hence H-plane arm) of a rectangular waveguide (between 3 and 4). It is recalled from Chapter 5 that waveguides can also be fabricated within CMOS structures and hence the hybrid T based on waveguides can be useful for RFIC if it is properly designed, laid out and used. The operation of the hybrid T can be inferred from the field distributions shown in Figure 8.21. When an input signal (assuming TE_{10} mode) enters port 1 (H-plane arm), the electric field of the signal splits equally with same phase between ports 3 and 4, and equally but opposite phase at port 2 (E-plane arm), as shown in Figure 8.21(b). As a result, the signals emerging from ports 3 and 4 have equal amplitude and phase, while there is no signal at port 2 due to signal cancelation. On the other hand, as seen in Figure 8.21(c), when an input signal (assuming TE_{10} mode) arrives at port 2,

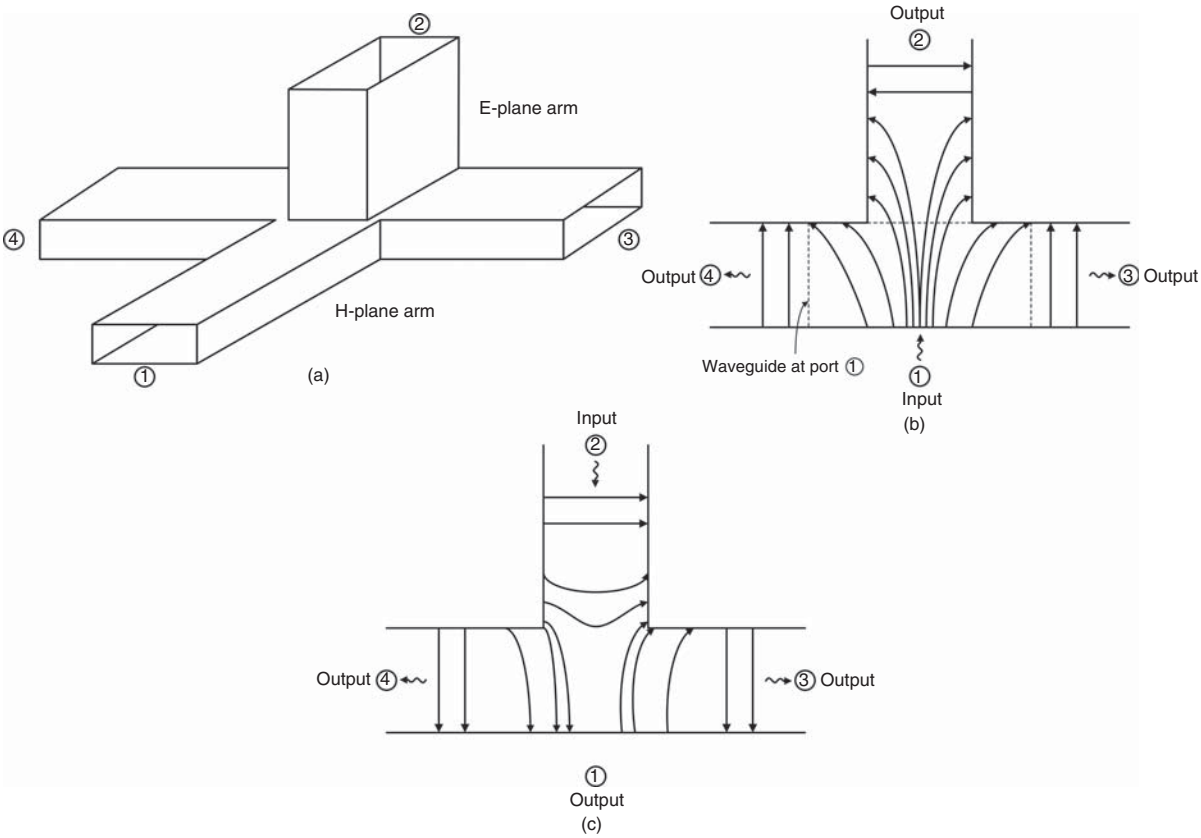


Figure 8.21. Hybrid T (a) and the field distributions with input signal at port 1 (b) and port 2 (c).

its electric field splits equally with opposite phase between ports 3 and 4, and equally but opposite phase at port 1. This results in 180° out of phase signals at ports 3 and 4 and no signal at port 1.

We consider a lossless, reciprocal, and matched hybrid T. We recall that the S-parameters of such a four-port network are described by (8.25). As described earlier, an input signal entering port 1 would split equally with 180° phase difference between ports 3 and 4; therefore $S_{42} = S_{24} = -S_{32} = -S_{23}$ considering the reciprocal property. The S-matrix from (8.25) can hence be rewritten as

$$[S] = \begin{bmatrix} 0 & 0 & S_{13} & S_{14} \\ 0 & 0 & S_{23} & -S_{23} \\ S_{13} & S_{23} & 0 & 0 \\ S_{14} & -S_{23} & 0 & 0 \end{bmatrix} \tag{8.85}$$

which contains three unknowns: S_{13}, S_{14}, and S_{23}. Enforcing the lossless condition in (8.24) leads to three equations in these unknowns that can be solved to obtain

$$|S_{13}| = |S_{14}| = |S_{23}| = \frac{\sqrt{2}}{2} \tag{8.86}$$

There is always a phase associate with each of the S-parameters – for instance, $S_{14} = \frac{\sqrt{2}}{2} e^{j\theta_{14}}$. By choosing proper terminal planes for ports 1 and 2, we can make S_{14} and S_{23} real, and hence the S-parameters of the

hybrid T can be described as

$$[S] = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \tag{8.87}$$

The S-matrix in (8.87) describes completely the operation of the hybrid T. When port 1 is the input port, ports 3 and 4 would be the output ports whose signals have equal magnitude and phase, and port 4 would contain no signal (isolated or decoupled port as defined in directional couplers). On the other hand, if the input port is port 2, then the output ports are 3 and 4 with an equal amplitude and 180° out of phase, and port 1 is the isolated or decoupled port.

8.3.2 Ring Hybrid

Figure 8.22 shows the ring hybrid (also known as rat-race ring or hybrid). The length between ports 1–2, 1–3, and 3–4 is one quarter-wavelength (or 90°) , and that between ports 2 and 4 is three quarter-wavelength (or 270°) at the design frequency. The ring hybrid, as mentioned earlier, is basically a hybrid T in which ports 1 and 2 can be called H- and E-plane arms, respectively. As such, its S-matrix can also be described by that of the hybrid T in (8.87). The operation of the ring hybrid can be inferred from the difference in the electrical lengths between the ports (0° and 180°). When an input signal arrives at port 1 (H-plane arm), the signal splits equally with same phase between ports 2 and 3, and vanishes at port 4. On the other hand, when an input signal enters port 2 (E-plane arm), it splits equally with 180° phase difference between ports 1 and 4, and no signal would appear at port 3. We can analyze the ring hybrid using the even- and odd-mode method as done for the parallel-coupled directional coupler. The analysis proceeds first with separate even- and odd-mode analysis, and then combining them to produce the final result.

8.3.2.1 Even-Mode Analysis. Figure 8.23(a) shows the ring hybrid with a symmetry plane along the center of the ring under the even-mode operation in which ports 1, 3 and 2, 4 are excited with equal potentials $A_1/2$ and $A_2/2$, respectively. This results in an open circuit or a MW along the symmetry plane. Due to the open circuit, the ring hybrid is divided into two independent identical halves with two open-circuited terminations – either of which can be used for the analysis of the ring hybrid. Figure 8.23(b) shows such a resultant circuit between ports 1 and 2. Figure 8.23(c) shows the equivalent circuit between ports 1 and 2 and between ports 3 and 4, where the shunt admittances at ports 1 (3) and 2 (4) are obtained as the input admittances of the $\lambda/8$ and $3\lambda/8$ open-circuited transmission lines, respectively. The S-matrix equations of these two-port networks are given in (8.37) and (8.38), respectively.

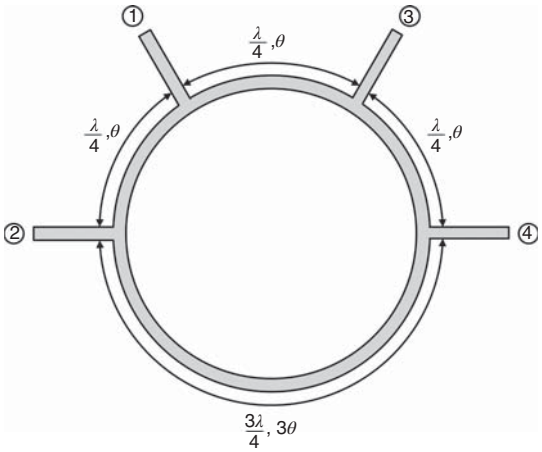


Figure 8.22. Ring hybrid or rat-race hybrid.

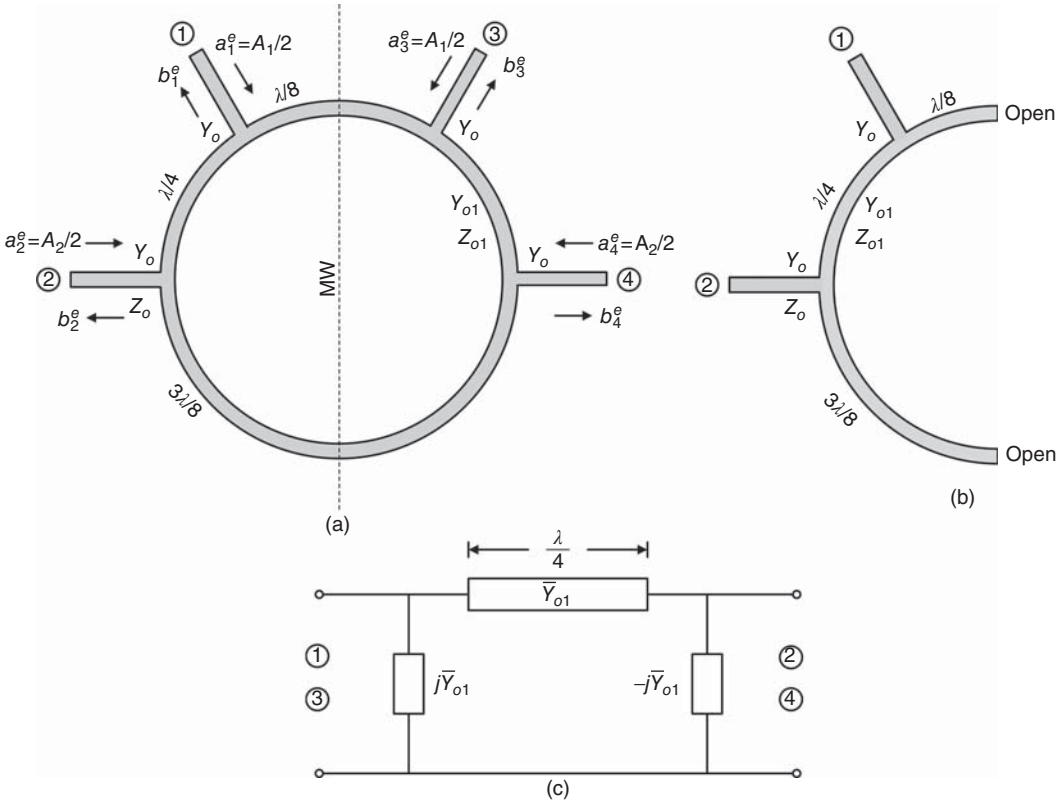


Figure 8.23. Ring hybrid under even-mode excitation (a), and the resultant half-circuit (b) and its equivalent circuit (c). *a*'s and *b*'s represent the incident and reflected voltage waves. Z_{o1} is the characteristic impedance of the ring's transmission line and $\overline{Y}_{o1} = Y_{o1}/Y_o = 1/Z_{o1}$ is the normalized characteristic admittance. Z_o is the terminating impedance or characteristic impedance of the connecting transmission line and $Y_o = 1/Z_o$ is the characteristic admittance.

The *ABCD* matrix of the network shown in Figure 8.23(c) can be obtained as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ j\overline{Y}_{o1} & 1 \end{bmatrix} \begin{bmatrix} 0 & \dfrac{j}{\overline{Y}_{o1}} \\ j\overline{Y} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j\overline{Y}_{o1} & 1 \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{j}{\overline{Y}_{o1}} \\ 2j\overline{Y}_{o1} & -1 \end{bmatrix} \tag{8.88}$$

Applying the conversion formulas between the *ABCD*- and *S*-parameters and the reciprocity property, we can derive the *S*-parameters as

$$\begin{aligned} S_{11}^e &= \frac{A+B-C-D}{A+B+C+D} = \frac{1+j/\overline{Y}_{o1}-2j\overline{Y}_{o1}+1}{1+j/\overline{Y}_{o1}+2j\overline{Y}_{o1}-1} = \frac{1-2\overline{Y}_{o1}^2-2j\overline{Y}_{o1}}{1+2\overline{Y}_{o1}^2} \\ S_{12}^e &= S_{21}^e = \frac{2(AD-BC)}{A+B+C+D} = \frac{-2j\overline{Y}_{o1}}{1+2\overline{Y}_{o1}^2} \\ S_{22}^e &= \frac{-A+B-C+D}{A+B+C+D} = \frac{-1+j/\overline{Y}_{o1}-2j\overline{Y}_{o1}-1}{1+j/\overline{Y}_{o1}+2j\overline{Y}_{o1}-1} = \frac{1-2\overline{Y}_{o1}^2+2j\overline{Y}_{o1}}{1+2\overline{Y}_{o1}^2} \end{aligned} \tag{8.89}$$

8.3.2.2 Odd-Mode Analysis. Figure 8.24(a) shows the ring hybrid under the odd-mode operation in which ports 1, 3 and 2, 4 are excited with opposite potentials $\pm A_1/2$ and $\pm A_2/2$, respectively. This results in a short

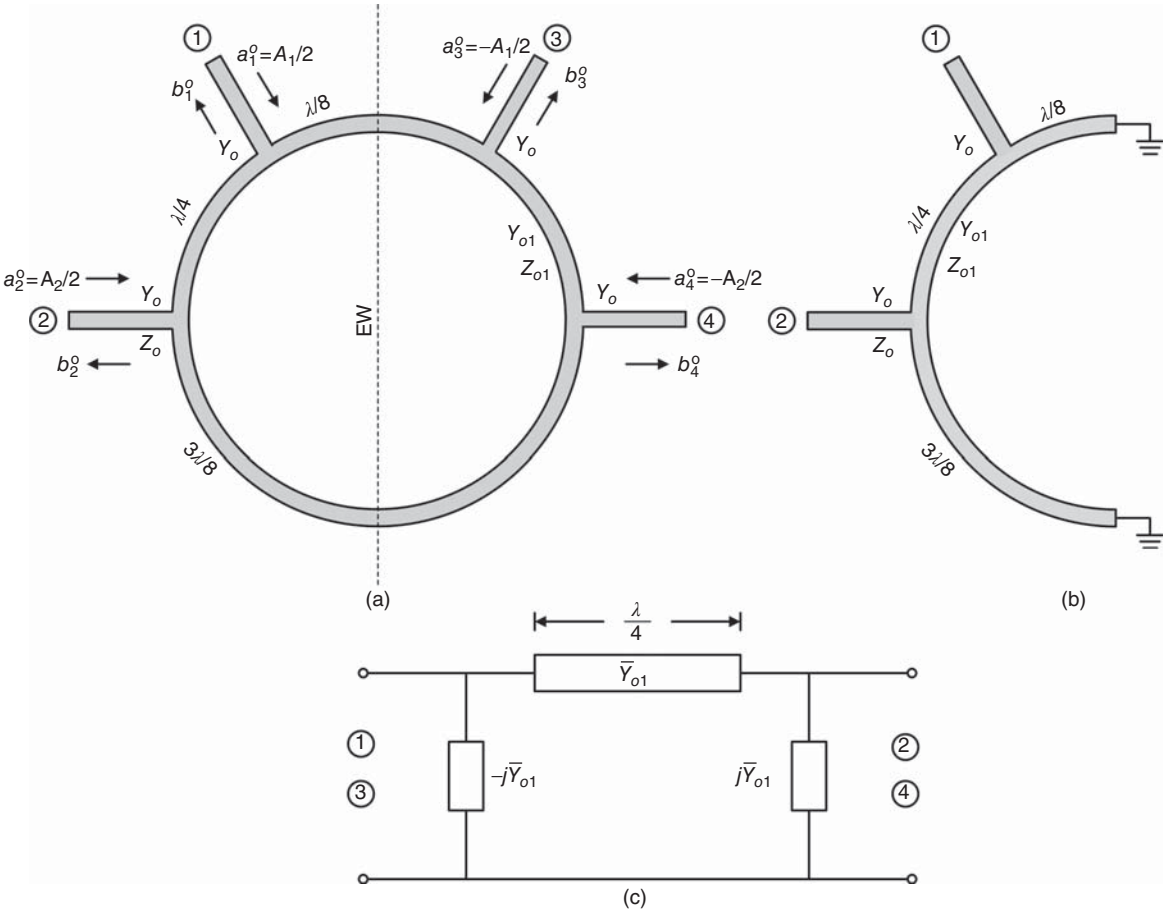


Figure 8.24. Ring hybrid under odd-mode excitation (a), and the resultant half-circuit (b) and its equivalent circuit (c).

circuit or an EW along the symmetry plane. Due to the short circuit, the ring hybrid is divided into two independent identical halves with two short-circuited terminations – either of which can be used for the analysis of the ring hybrid. Figure 8.24(b) shows such a resultant circuit between ports 1 and 2. Figure 8.24(c) shows the equivalent circuit between ports 1 and 2 and between ports 3 and 4, where the shunt admittances at ports 1 (3) and 2 (4) are obtained as the input admittances of the $\lambda/8$ and $3\lambda/8$ short-circuited transmission lines, respectively. The *S*-matrix equations of these two-port networks are given in (8.45) and (8.46), respectively.

The ABCD matrix of the network shown in Figure 8.24(c) can be obtained as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -j\overline{Y}_{o1} & 1 \end{bmatrix} \begin{bmatrix} 0 & \dfrac{j}{\overline{Y}_{o1}} \\ j\overline{Y}_{o1} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ +j\overline{Y}_{o1} & 1 \end{bmatrix} = \begin{bmatrix} -1 & \dfrac{j}{\overline{Y}_{o1}} \\ j\overline{Y}_{o1} & 1 \end{bmatrix} \tag{8.90}$$

Applying the conversion formulas between the *ABCD*- and *S*-parameters and the reciprocity property, we can derive the *S*-parameters as

$$S_{11}^o = \frac{A + B - C - D}{A + B + C + D} = \frac{-1 + j/\overline{Y}_{o1} - 2j\overline{Y}_{o1} - 1}{-1 + j/\overline{Y}_{o1} + 2j\overline{Y}_{o1} + 1} = \frac{1 - 2\overline{Y}_{o1}^2 + 2j\overline{Y}_{o1}}{1 + 2\overline{Y}_{o1}^2}$$

$$S_{12}^o = S_{21}^o = \frac{2(AD - BC)}{A + B + C + D} = \frac{-2j\overline{Y}_{o1}}{1 + 2\overline{Y}_{o1}^2}$$

$$S_{22}^{o} = \frac{-A + B - C + D}{A + B + C + D} = \frac{1 + j/\overline{Y}_{o1} - 2j\overline{Y}_{o1} + 1}{-1 + j/\overline{Y}_{o1} + 2j\overline{Y}_{o1} + 1} = \frac{1 - 2\overline{Y}_{o1}^{2} - 2j\overline{Y}_{o1}}{1 + 2\overline{Y}_{o1}^{2}} \tag{8.91}$$

8.3.2.3 Superimposition of Even and Odd Modes. Figure 8.25 shows the superimposition of the even- and odd-mode ring hybrid circuits in which the incident and reflected voltage waves are obtained as the summation of the mode voltage waves as in (8.48) and (8.49). Similar to the parallel-coupled directional coupler, the *S*-matrix equations for the ring hybrid are also given by (8.52) and (8.53), and likewise the *S*-parameters are given by (8.54). We can derive from (8.54), making use of (8.89), (8.91), $S_{11} = S_{33}$, $S_{22} = S_{44}$, $S_{12} = S_{34}$ from the symmetry property, and $S_{ij} = S_{ji}$ from the reciprocal property,

$$S_{11} = S_{22} = S_{33} = S_{44} = \frac{1 - 2\overline{Y}_{o1}^{2}}{1 + 2\overline{Y}_{o1}^{2}}$$

$$S_{12} = S_{21} = S_{13} = S_{31} = S_{34} = S_{43} = -S_{24} = -S_{42} = -\frac{2j\overline{Y}_{o1}}{1 + 2\overline{Y}_{o1}^{2}}$$

$$S_{14} = S_{41} = S_{32} = S_{23} = 0 \tag{8.92}$$

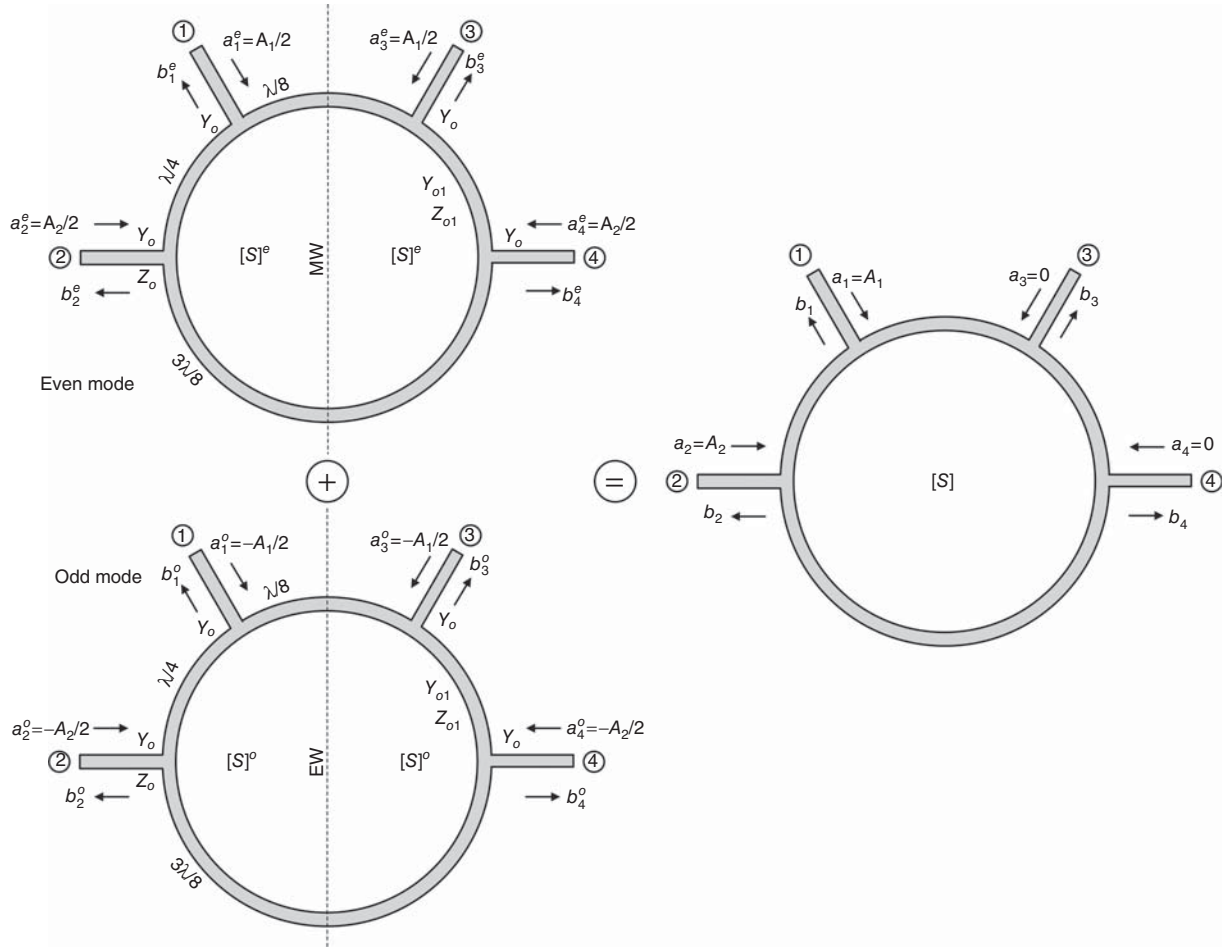


Figure 8.25. Superimposition of even and odd modes. $[S]^e$ and $[S]^o$ represent the *S*-matrix of half of the ring hybrid under the even- and odd-mode operation, respectively, and $[S]$ represents the *S*-matrix of the ring hybrid.

Setting $S_{ii} = 0, i = 1, 2, 3, 4$, for perfect match at all ports leads to

$$1 - 2\overline{Y}_{o1}^2 = 0 \tag{8.93}$$

or

$$Z_{o1} = Z_o\sqrt{2} \tag{8.94}$$

which is an important design equation for the ring hybrid. Substituting (8.94) into (8.92) gives the *S*-matrix for the ring hybrid as

$$[S] = -\frac{j}{\sqrt{2}}\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \tag{8.95}$$

The operation of the ring hybrid as mentioned earlier can now be seen clearly from (8.93). The foregoing analysis of the ring hybrid is obtained assuming the electrical length θ noted in Figure 8.23 is 90°, which is the electrical length at the design center frequency at which the performance is optimum. When the frequency is changed, θ changes as a function of frequency as

$$\theta = \frac{\pi}{2}\left(\frac{f}{f_o}\right) \tag{8.96}$$

where f_o is the design center frequency, and the performance of the ring hybrid varies accordingly. The operating bandwidth of the ring hybrid depends mainly on the specified amplitude and phase balance. The ring hybrid is typically used for a bandwidth around 15%.

8.3.2.4 Design of Ring Hybrid. As mentioned previously, one serious problem with transmission-line based RF circuits in CMOS structures is the transmission-line length. This problem is even more pronounced in the ring hybrid due to the use of a $3\lambda/4$-long transmission line in one arm to achieve the required 180° phase difference. The $3\lambda/4$-long transmission-line problem, however, can be overcome by introducing a 180° phase shifter in the $3\lambda/4$ arm. The phase shifter together with a $\lambda/4$-long transmission line produces the required 270° for the $3\lambda/4$ arm. Figure 8.26 illustrates the general concept of implementing a phase shifter in the ring hybrid. There are various ways to realize the phase shifter. The simplest technique that also yields a wide bandwidth for a 180° phase shift is based on exploiting the field distributions in transmission lines such

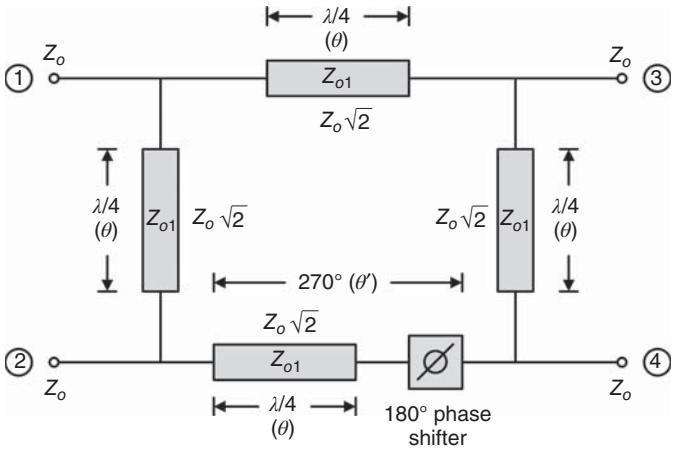


Figure 8.26. Ring hybrid with a 180° phase shifter in the $3\lambda/4$ arm. θ is an arbitrary electrical length and $\lambda/4$ is the nominal length. $Z_{o1} = \sqrt{2}Z_o$ when $\theta = \pi/2$.

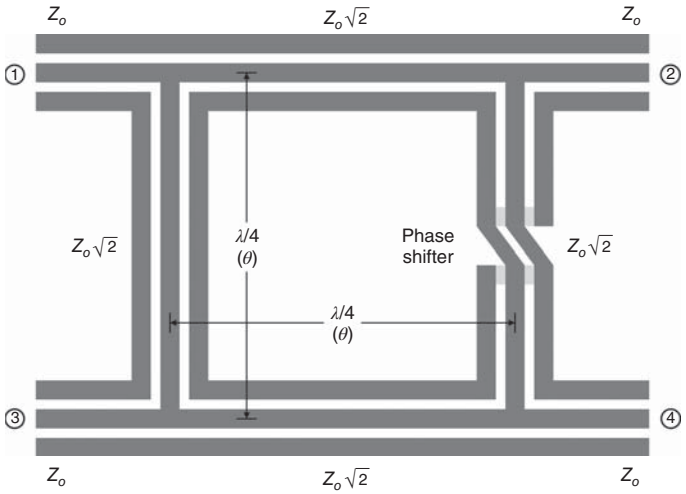


Figure 8.27. Ring hybrid with 180° phase shift obtained by interconnecting the ground and signal lines of CPW.

as that used to create a (180°) hybrid junction discussed in Section 4.12.6. Figure 8.27 shows a ring hybrid incorporating a broadband 180-deg phase shifter implemented by connecting the signal line of one coplanar waveguide (CPW) segment to the ground lines of another CPW segment [7, 8]. As pointed out in [7], the electrical length of each $\lambda/4$-arm needs not be 90° (i.e., $\theta \neq \pi/2$) when a phase shifter is used. This fact can be exploited to obtain a smaller ring circumference at the expense of the bandwidth. The characteristic impedance Z_{o1} of each arm can be determined from the characteristic impedance Z_o of the terminating transmission line and the electrical length θ of each arm as [7]

$$Z_{o1} = Z_o\sqrt{2(1 - \cot^2\theta)} \tag{8.97}$$

For $\theta = \pi/2$, this reduces to the widely known relation $Z_{o1} = \sqrt{2}Z_o$ typically used for the ring hybrid.

Besides implementing a 180° phase shifter in the $3\lambda/4$ arm, a ring hybrid design can also employ multiple metal layers available in CMOS/BiCMOS structures to achieve a very compact size. Figure 8.28 shows such

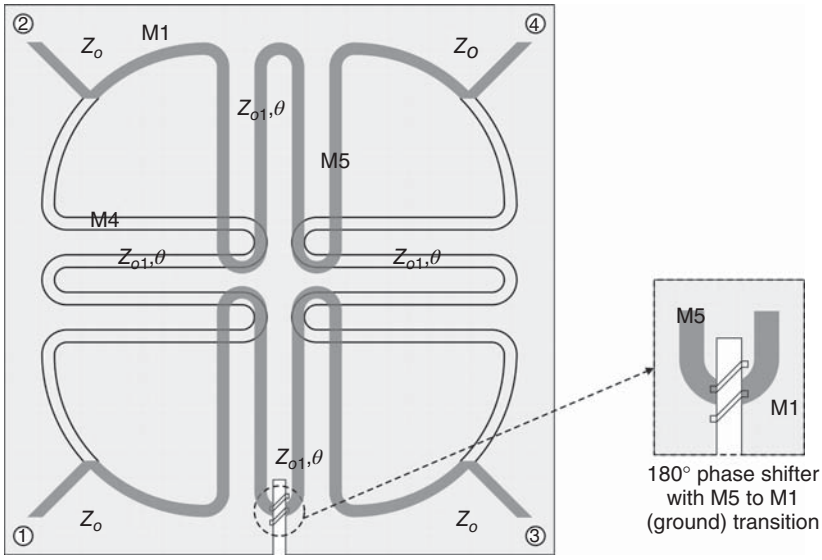


Figure 8.28. Layout of the ring hybrid with a 180° phase shifter in the CMOS structure shown in Figure 8.16. The ring's transmission line is located on both M4 and M5 and the ground plane is on M1. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

a design [5]. The multilayer conductor facilitates size reduction through meandering, which is rather difficult using conductors on the same side of a dielectric substrate due to severe unwanted coupling between meandered sections that are too close to each other. As mentioned earlier, the phase shifter compensates for the typical $3\lambda/4$ arm through a $\lambda/4$ arm. The structure was implemented in the same 5-metal layer TSMC 0.25 μm CMOS technology used for the four-finger broadside-coupled Lange coupler described previously. Though each arm of the ring hybrid is still $\lambda/4$ long, significant size reduction was obtained through simple meandering as the adjacent arms of the structure were implemented in different metal layers (M5 and M4). In order to maintain similar characteristic impedance in each arm, the widths of the transmission lines corresponding to the lower layer (M4) and upper layer (M5) must be different, owing to the different distances from the ground plane (M1). Figure 8.28 shows that each adjacent arm overlaps its neighbor, hence possibly causing unwanted coupling. This overlap spans for 47 μm for each turn and represents 12% of the total length of each arm. However, the transverse layout of the M5 and M4 bends makes the magnetic fields on the transverse transmission-line segments on M4 and M5 orthogonal to each other, effectively minimizing the magnetic coupling between these two segments and facilitating further miniaturization. The coupling between the overlapping segments caused by the electric field is also expected to be small because the cross sectional area of each line is not more than a few tens of microns. The design of the ring hybrid was based on estimating the characteristic impedance and electrical length of each arm of the ring according to (8.97). Using 50 Ω for the port impedance (Z_o) and the ring's characteristic impedance (Z_{o1}), the electrical length θ was determined as 54.7° from (8.97). However, after subsequent optimization through EM simulations using IE3D, the ring's characteristic impedance and electrical length were obtained as 48 Ω and 53°, respectively. The widths required to achieve these parameters were simulated from IE3D. The transmission line on the M5 layer has a width of 10 μm, while that on the M4 layer is 7.1 μm wide. Both of them have a similar length of 781.6 μm. The 180° phase shifter was realized with a 10-μm slit in the signal (M5) and ground (M1) planes with via-holes connecting M5 and M1 layers through M3.

For measurement convenience, three structures were fabricated to measure the through, coupling, and isolation. A die photo of a typical ring hybrid is shown in Figure 8.29. The dimensions of this structure are only 314 μm × 282 μm without RF pads. The measured and simulated results are shown in Figures 8.30–8.32. The through (2) and coupled (3) ports with respect to the input port (1) exhibit measured amplitudes of −3.1 to −3.18 and −5.1 to −5.7 dB with more than 17 dB isolation between the input (1) and decoupled (4) ports from 25 to 35 GHz, respectively. The excess loss at the coupled port is attributed mainly to the finite conductivity of the metal layers, the nonideal nature of vias, and the phase shifter. This excess loss is, however, a reasonable compromise for obtaining such extremely compact structures, owing to the significant emphasis placed on chip area in silicon technologies. The structure is very well matched around 28 GHz with the measured return loss remains below 10 dB between 20 and 35 GHz. The measured phase difference between the through and coupled ports presents a 180° response at 30 GHz with a ±5° imbalance across 25–35 GHz.

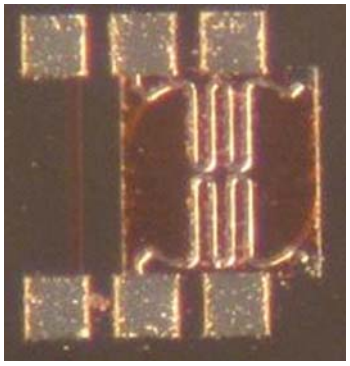


Figure 8.29. Die photograph of the ring hybrid coupler. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

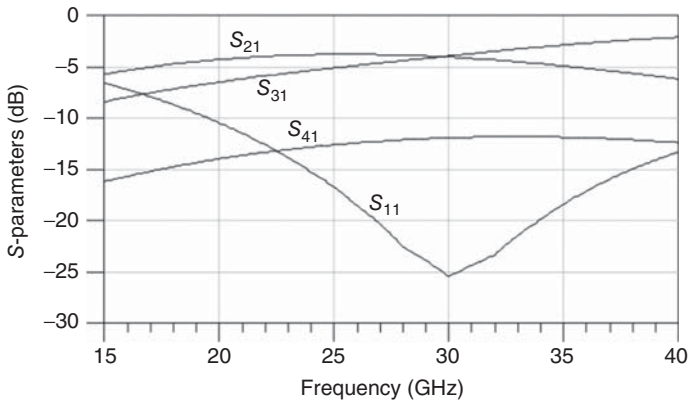


Figure 8.30. Simulated *S*-parameters of the ring hybrid. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

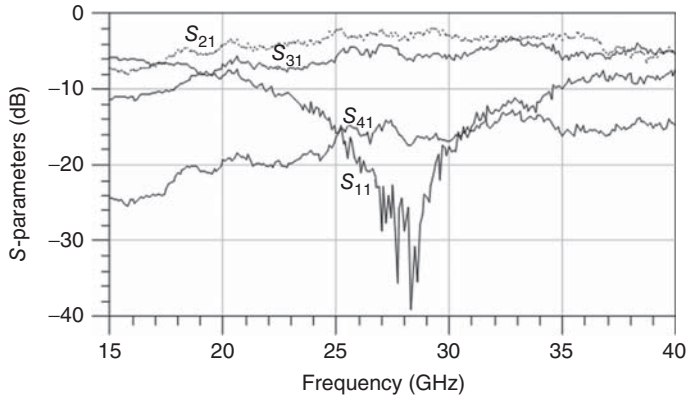


Figure 8.31. Measured *S*-parameters of the ring hybrid. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

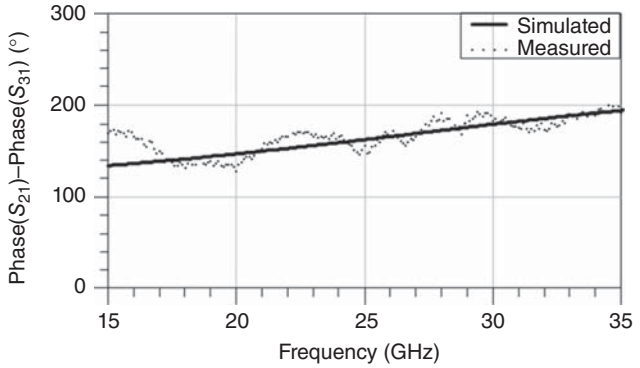


Figure 8.32. Simulated and measured phase difference between the coupled and through ports. (After Chirala and Nguyen [5]. Reprinted with permission of IEEE.)

8.3.3 Branch-Line Coupler

Figure 8.33 shows the branch-line coupler (also known as quadrature hybrid). The length between ports is a quarter-wavelength (or 90°) at the design frequency. The operation of the branch-line coupler can be inferred from the electrical length between ports. For instance, when an input signal arrives at port 1, the signal splits equally with 90° phase difference between ports 2 and 3, and no signal would appear at port 4. The branch-line coupler can be analyzed using the even- and odd-mode technique which consists of three parts: even-mode analysis, odd-mode analysis, and superposition of even- and odd-mode.

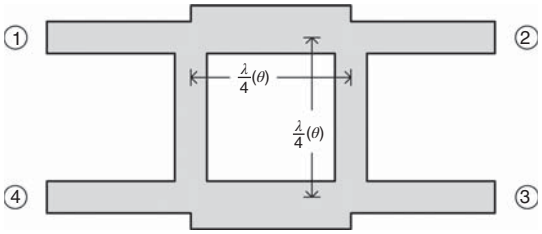


Figure 8.33. Branch-line coupler.

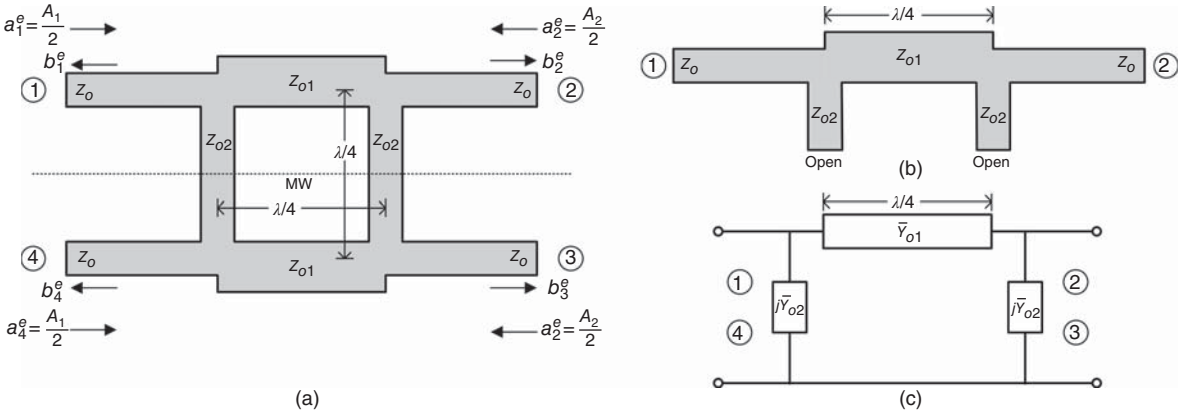


Figure 8.34. Branch-line coupler under even-mode excitation (a), and the resultant half-circuit (b) and its equivalent circuit (c). $Z_{o1}(Z_{o2})$ is the characteristic impedance of the horizontal (vertical) branch line and $\overline{Y}_{o1}(\overline{Y}_{o2})$ is the corresponding normalized characteristic admittance. Z_o is the terminating impedance or characteristic impedance of the connecting transmission line.

8.3.3.1 Even-Mode Analysis. Figure 8.34(a) shows the branch-line coupler with an (horizontal) symmetry plane along the center of the coupler under the even-mode operation in which ports 1, 4 and 2, 3 are excited with equal potentials $A_1/2$ and $A_2/2$, respectively. This results in an open circuit or a MW along the symmetry plane. Due to the open circuit, the branch-line coupler is divided into two independent identical halves with two open-circuited terminations. Figure 8.34(b) shows the circuit between ports 1 and 2. Figure 8.34(c) shows the equivalent circuit between ports 1 and 2 and between ports 4 and 3, where the shunt admittances at ports 1 (4) and 2 (3) are obtained as the input admittances of the $\lambda/8$ open-circuited transmission line. The S-matrix equations of these two-port networks are given in (8.37) and (8.38), respectively. Note that (8.38) still applies since ports 3 and 4 are symmetrical.

The ABCD matrix of the network shown in Figure 8.34(c) can be obtained as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ j\overline{Y}_{o2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \dfrac{j}{\overline{Y}_{o1}} \\ j\overline{Y}_{o1} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ j\overline{Y}_{o2} & 1 \end{bmatrix} = \begin{bmatrix} -\dfrac{\overline{Y}_{o2}}{\overline{Y}_{o1}} & -\dfrac{j}{\overline{Y}_{o1}} \\ j\left(\overline{Y}_{o1} - \dfrac{\overline{Y}_{o2}^2}{\overline{Y}_{o1}}\right) & -\dfrac{\overline{Y}_{o2}}{\overline{Y}_{o1}} \end{bmatrix} \tag{8.98}$$

from which, we can derive the S-parameters, making use of the symmetrical and reciprocity property, as

$$\begin{aligned} S_{11}^e = S_{22}^e &= \frac{A + B - C - D}{A + B + C + D} = \frac{-1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2}{1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2 + 2j\overline{Y}_{o2}} \\ S_{12}^e = S_{21}^e &= \frac{2(AD - BC)}{A + B + C + D} = \frac{-2\overline{Y}_{o1}}{2\overline{Y}_{o2} - j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)} \end{aligned} \tag{8.99}$$

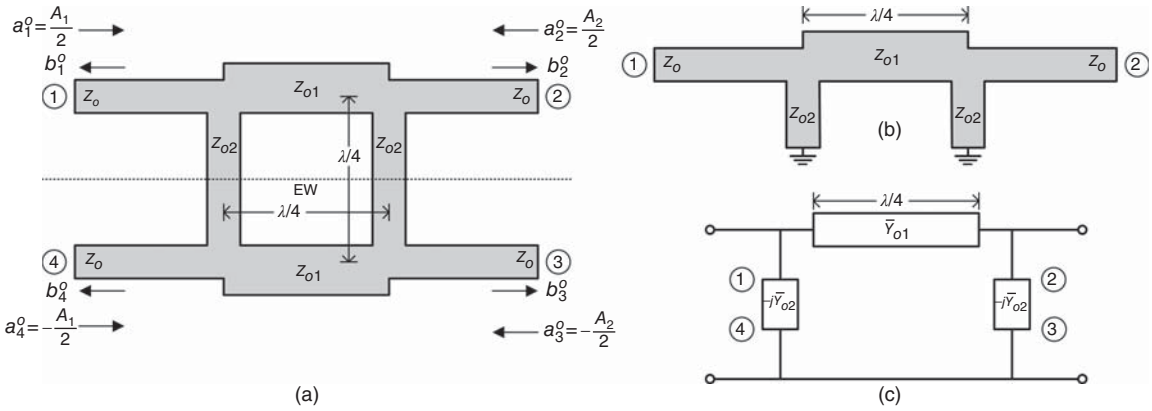


Figure 8.35. Branch-line coupler under odd-mode excitation (a), and the resultant half-circuit (b) and its equivalent circuit (c).

8.3.3.2 Odd-Mode Analysis. Figure 8.35(a) shows the branch-line coupler under the odd-mode operation in which ports 1, 4 and 2, 3 are excited with opposite potentials $\pm A_1/2$ and $\pm A_2/2$, respectively. This results in a short circuit or an EW along the symmetry plane. Due to the short circuit, the ring hybrid is divided into two independent identical halves with two short-circuited terminations. Figure 8.35(b) shows the circuit between ports 1 and 2. Figure 8.35(c) shows the equivalent circuit between ports 1 and 2 and between ports 4 and 3, where the shunt admittances at ports 1 (4) and 2 (3) are obtained as the input admittances of the $\lambda/8$ short-circuited transmission lines. The S-matrix equations of these two-port networks are given in (8.45) and (8.46), respectively.

The $ABCD$ matrix of the network shown in Figure 8.35(c) can be obtained as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -j\overline{Y}_{o2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \dfrac{j}{\overline{Y}_{o1}} \\ j\overline{Y}_{o1} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -j\overline{Y}_{o2} & 1 \end{bmatrix} = \begin{bmatrix} \dfrac{\overline{Y}_{o2}}{\overline{Y}_{o1}} & \dfrac{j}{\overline{Y}_{o1}} \\ j\left(\overline{Y}_{o1} - \dfrac{\overline{Y}_{o2}^2}{\overline{Y}_{o1}}\right) & \dfrac{\overline{Y}_{o2}}{\overline{Y}_{o1}} \end{bmatrix} \tag{8.100}$$

from which, we can derive the S-parameters, making use of the symmetrical and reciprocity property, as

$$\begin{aligned} S_{11}^o = S_{22}^o &= \frac{A + B - C - D}{A + B + C + D} = \frac{1 - \overline{Y}_{o1}^2 + \overline{Y}_{o2}^2}{1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2 - 2j\overline{Y}_{o2}} \\ S_{12}^o = S_{21}^o &= \frac{2(AD - BC)}{A + B + C + D} = \frac{2\overline{Y}_{o1}}{2\overline{Y}_{o2} + j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)} \end{aligned} \tag{8.101}$$

8.3.3.3 Superimposition of Even and Odd Modes. Figure 8.36 shows the superimposition of the even- and odd-mode branch-line-coupler circuits in which the incident and reflected voltage waves are obtained as the summation of the mode voltage waves as in (8.48) and (8.49). We can derive from (8.54) with ports 3 and 4 interchanged, making use of (8.99), (8.101), $S_{11} = S_{22} = S_{33} = S_{44}$, $S_{12} = S_{43}$, $S_{13} = S_{24}$, $S_{14} = S_{23}$ from the symmetry property, and $S_{ij} = S_{ji}$ from the reciprocal property,

$$S_{11} = S_{22} = S_{33} = S_{44} = \frac{j(1 + \overline{Y}_{o2}^2 - \overline{Y}_{o1}^2)}{2} \left[\frac{1}{-2\overline{Y}_{o2} + j\left(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2\right)} + \frac{1}{2\overline{Y}_{o2} + j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)} \right]$$

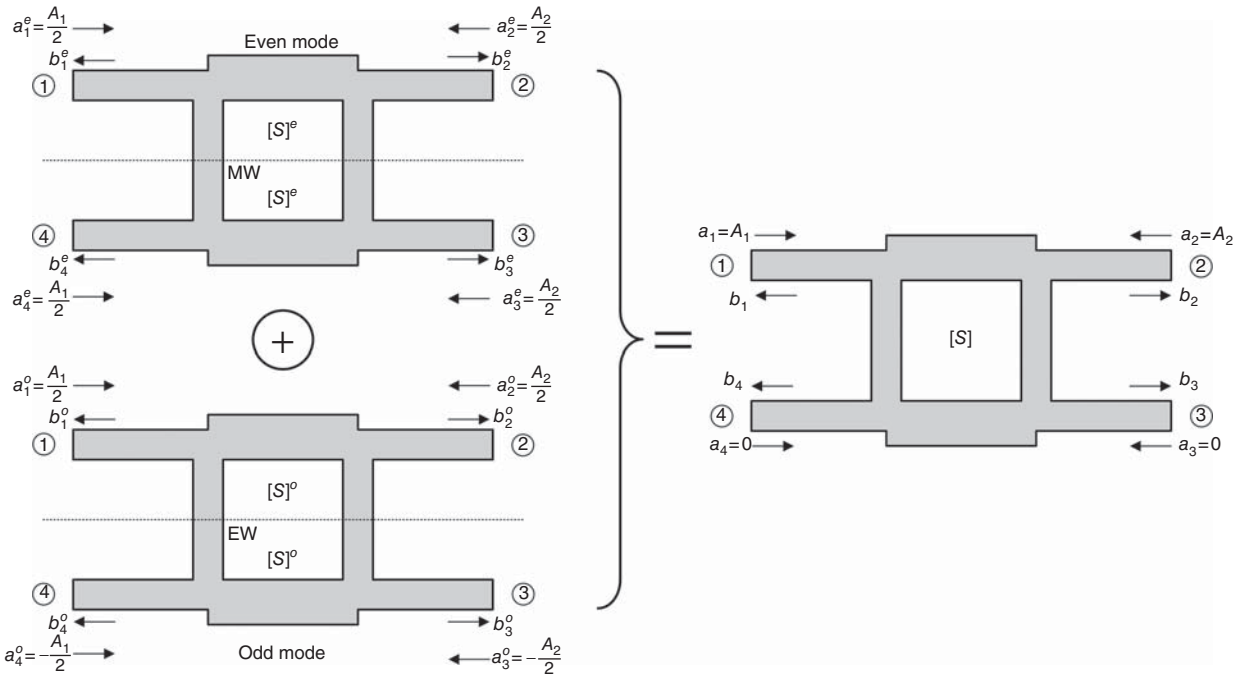


Figure 8.36. Superimposition of even and odd modes. $[S]^e$ and $[S]^o$ represent the S-matrix of half of the branch-line coupler under the even- and odd-mode operation, respectively, and $[S]$ represents the S-matrix of the branch-line coupler.

$$S_{12} = S_{21} = S_{34} = S_{43} = \overline{Y}_{o1}\left[\frac{1}{-2\overline{Y}_{o2} + j\left(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2\right)} + \frac{1}{2\overline{Y}_{o2} + j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)}\right]$$

$$S_{13} = S_{31} = S_{24} = S_{42} = \overline{Y}_{o1}\left[\frac{1}{-2\overline{Y}_{o2} + j\left(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2\right)} - \frac{1}{2\overline{Y}_{o2} + j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)}\right]$$

$$S_{14} = S_{41} = S_{32} = S_{23} = \frac{j(1 + \overline{Y}_{o2}^2 - \overline{Y}_{o1}^2)}{2}\left[\frac{1}{-2\overline{Y}_{o2} + j\left(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2\right)} - \frac{1}{2\overline{Y}_{o2} + j(1 + \overline{Y}_{o1}^2 - \overline{Y}_{o2}^2)}\right] \tag{8.102}$$

Under the perfect match condition, $S_{ii} = 0$, $i = 1, 2, 3, 4$, which, upon using (8.102), leads to

$$\frac{1}{Z_o^2} = \frac{1}{Z_{o1}^2} - \frac{1}{Z_{o2}^2} \tag{8.103}$$

Letting $Z_{o2} = Z_o$, we can then derive

$$Z_{o1} = \frac{Z_o}{\sqrt{2}} \tag{8.104}$$

which is an important design equation for the branch-line coupler. Substituting (8.104) into (8.102) gives the S-matrix for the branch-line coupler:

$$[S] = -\frac{1}{\sqrt{2}}\begin{bmatrix} 0 & j & 1 & 0 \\ j & 0 & 0 & 1 \\ 1 & 0 & 0 & j \\ 0 & 1 & j & 0 \end{bmatrix} \tag{8.105}$$

which clearly describes the operation of the branch-line coupler. It is noted that the foregoing analysis is obtained at the design center frequency corresponding to the electrical length θ noted in Figure 8.33 being 90°. As the frequency is changed from the center frequency, the performance degrades from the optimum performance since θ changes as a function of frequency as described in (8.96). As for the ring hybrid, the operating bandwidth of the branch-line coupler depends mainly on the specification of the amplitude and phase balance. The branch-line coupler is typically used for a bandwidth of around 10%.

8.4 POWER DIVIDERS

Power dividers, as the name implies, are used to split a signal from one port to other ports. In general, a power divider can have multiple ports and, when employing only passive elements, can also be used as a power combiner combining signals from different ports into one port – for instance, 1:3 power divider or combiner dividing a signal from the input port to three output ports or combining signals from the three input ports into one output port, respectively. Power dividers can be designed to have equal or unequal division. From a general perspective, power dividers can also be classified as directional couplers. One of the most fundamental and widely known power dividers is perhaps the two-way equal-division Wilkinson power divider [9] that is covered in this section.

Figure 8.37 shows the (two-way equal-division) Wilkinson power divider which splits an input signal into two equal output signals having 90° phase shift with respect to the input signal. The length between the input and output ports is a quarter-wavelength at the design frequency and there is a resistor R between the two output ports. We redraw the Wilkinson power divider as shown in Figure 8.38(a), where Z_o represents

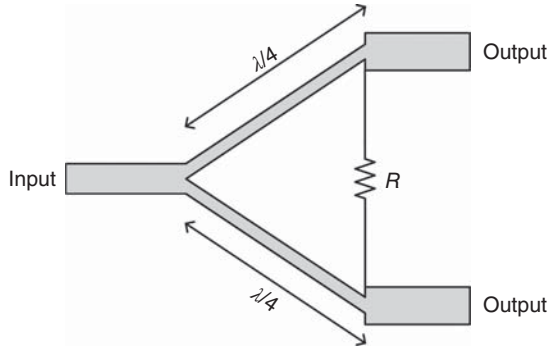


Figure 8.37. Wilkinson power divider.

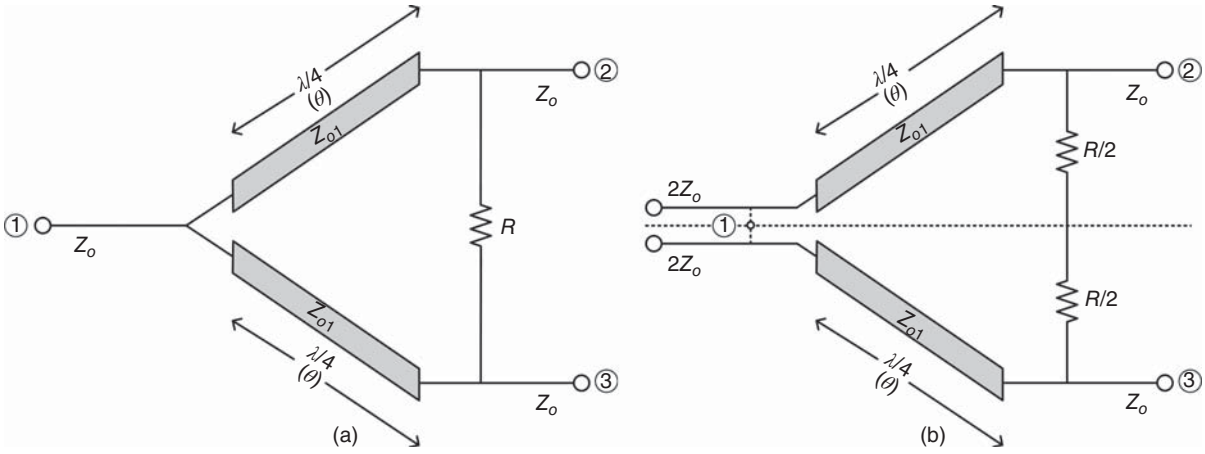


Figure 8.38. (a, b) Wilkinson power divider. Z_{o1} is the characteristic impedances of each branch. θ is the electrical length which is equal to 90° at the design frequency.

the characteristic impedance of the connecting transmission line or the port's terminating impedance. As shown in Figure 8.38(b), we further replace the resistor R with two resistors $R/2$ in series and the terminating impedance Z_o at port 1 by two terminating impedances $2Z_o$ in parallel. This is possible[3] since the impedance looking from port 1 to the terminating transmission line in Figure 8.38(a) is Z_o, assuming the transmission line is either infinitely long or terminated by a matched load of Z_o as usually done in the S-parameter analysis, while the impedance looking from port 1 to each of the terminating transmission lines having characteristic impedance $2Z_o$ in Figure 8.38(b) is $2Z_o$, assuming the transmission line is either infinitely long or terminated by a matched load of $2Z_o$. As Figure 8.38(b) shows, the power divider is perfectly symmetrical with respect to the central plane (dashed line), enabling us to apply the even- and odd-mode analysis based on the circuit's symmetry.

8.4.1 Even-Mode Analysis

Figure 8.39(a) shows the Wilkinson power divider under the even-mode operation in which ports 2 and 3 are excited with equal potential $A_2/2$. This results in an open circuit or a MW along the symmetry plane. It is noted that port 1 lies on the symmetry plane and hence the open circuit occurs by imposing only equal potential at ports 2 and 3. Due to the potentials applied at ports 2 and 3, there exists a potential at port 1 which can be decomposed into incident and reflected voltage waves. We let the normalized incident voltage $a_1^e = A_1$ and the corresponding reflected voltage is b_1^e as shown in Figure 8.39(a). Due to the open circuit, the power divider is divided into two independent identical halves – one of which between ports 1 and 2 is shown in Figure 8.39(b).

For generality, we use θ instead of 90° for the branch's electrical length at the design frequency. The input impedance $Z_{\text{in},1}$ at port 1 assuming the transmission line is lossless, is obtained as

$$Z_{\text{in},1} = Z_{o1}\frac{Z_o + jZ_{o1}\tan\theta}{Z_{o1} + jZ_o\tan\theta} \tag{8.106}$$

where Z_o is the load impedance at port 2. The resistor $R/2$ is not considered due to the open at its other terminal. Terminating port 2 with a matched load equal to Z_o, we can determine S_{11} as the reflection coefficient at port 1 as

$$S_{11}^e = \frac{Z_{\text{in},1}^e - 2Z_o}{Z_{\text{in},1}^e + 2Z_o} = \frac{-Z_oZ_{o1}\cos\theta + j(Z_{o1}^2 - 2Z_o^2)\sin\theta}{3Z_oZ_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} \tag{8.107}$$

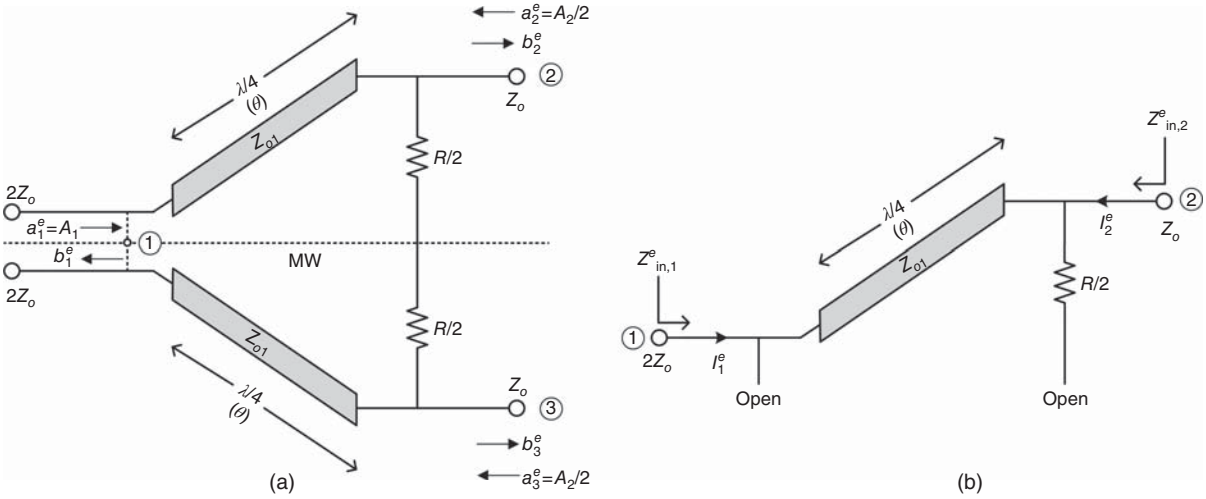


Figure 8.39. Wilkinson power divider under even-mode excitation (a) and the resultant half-circuit (b).

[3]In general, a transmission line with characteristic impedance Z_o is not equivalent to two transmission lines of $2Z_o$-characteristic impedance in parallel.

where $2Z_o$ is the load impedance at port 1. Similarly, the input impedance $Z^e_{\text{in},2}$ at port 2 is obtained as

$$Z^e_{\text{in},2} = Z_{o1}\frac{2Z_o + jZ_{o1}\tan\theta}{Z_{o1} + j2Z_o\tan\theta} \tag{8.108}$$

Terminating port 1 with a matched load equal to $2Z_o$, we can determine S^e_{22} S_{22} as the reflection coefficient at port 2 as

$$S^e_{22} = \frac{Z_{\text{in},2} - Z_o}{Z_{\text{in},2} + Z_o} = \frac{Z_oZ_{o1}\cos\theta + j(Z^2_{o1} - 2Z^2_o)\sin\theta}{3Z_oZ_{o1}\cos\theta + j(Z^2_{o1} + 2Z^2_o)\sin\theta} \tag{8.109}$$

Normalizing the voltage waves at ports 1 and 2 with respect to the corresponding port impedances $2Z_o$ and Z_o, respectively, and applying the S-matrix equation, we can write

$$S^e_{21} = \left.\frac{b^e_2}{a^e_1}\right|_{a^e_2=0} = \left.\frac{V^{e-}_2/\sqrt{Z_o}}{V^{e+}_1/\sqrt{2Z_o}}\right|_{V^{e+}_2=0} \tag{8.110}$$

Under the match condition at port 2 according to $V^{e+}_2 = 0$, $S^e_{11} = V^{e-}_1/V^{e+}_1$ and the total current I^e_1 at port 1 can be obtained as

$$I^e_1 = I^{e+}_1 + I^{e-}_1 = \frac{V^{e+}_1 - V^{e-}_1}{2Z_o} = \frac{V^{e+}_1(1 - S^e_{11})}{2Z_o} \tag{8.111}$$

making use of $2Z_o = V^{e+}_1/I^{e+}_1 = -V^{e-}_1/I^{e-}_1$ considering $2Z_o$ as the characteristic impedance of the terminating transmission line at port 1. Since the currents at ports 1 and 2 of the transmission line are related by $I^e_2 = -I^e_1e^{-j\theta}$, we can write using (8.111) and recognizing that $I^{e+}_2 = 0$ under the match condition at port 2,

$$I^e_2 = I^{e-}_2 = -I^e_1e^{-j\theta} = -\frac{V^{e+}_1(1 - S^e_{11})e^{-j\theta}}{2Z_o} \tag{8.112}$$

Making use of $Z_o = -V^{e-}_2/I^{e-}_2$ from the terminating transmission line of Z_o characteristic impedance at port 2, we can write from (8.112):

$$\frac{V^{e-}_2}{Z_o} = \frac{V^{e+}_1(1 - S^e_{11})e^{-j\theta}}{2Z_o} \tag{8.113}$$

Substituting (8.113) into (8.110) then gives

$$S^e_{21} = \frac{Z_o}{2Z_o}\sqrt{\frac{2Z_o}{Z_o}}(1 - S^e_{11})e^{-j\theta} = \frac{1}{\sqrt{2}}(1 - S^e_{11})e^{-j\theta} \tag{8.114}$$

which, upon using (8.107), becomes

$$S^e_{21} = \frac{2\sqrt{2}Z_o(Z_{o1}\cos\theta + jZ_o\sin\theta)e^{-j\theta}}{3Z_oZ_{o1}\cos\theta + j(Z^2_{o1} + 2Z^2_o)\sin\theta} \tag{8.115}$$

Again, using the voltage waves at ports 1 and 2 normalized to $2Z_o$ and Z_o, respectively, we can obtain S^e_{12} as

$$S^e_{12} = \left.\frac{b^e_1}{a^e_2}\right|_{a^e_1=0} = \left.\frac{V^{e-}_1/\sqrt{2Z_o}}{V^{e+}_2/\sqrt{Z_o}}\right|_{V^{e+}_1=0} \tag{8.116}$$

Under the match condition at port 1 according to $V_1^{e+} = 0$, $S_{22}^e = V_2^{e-}/V_2^{e+}$ and the total current I_2^e at port 2 is obtained as

$$I_2^e = I_2^{e+} + I_2^{e-} = \frac{V_2^{e+} - V_2^{e-}}{Z_o} = \frac{V_2^{e+}(1 - S_{22}^e)}{Z_o} \tag{8.117}$$

Since $I_1^e = -I_2^e e^{j\theta}$ and $I_1^{e+} = 0$ under the match condition at port 1, we can write using (8.117):

$$I_1^{e-} = -\frac{V_1^{e-}}{2Z_o} = -\frac{V_2^{e+}(1 - S_{22}^e)e^{j\theta}}{Z_o} \tag{8.118}$$

which, upon substituting into (8.116), gives

$$S_{12}^e = \sqrt{2}(1 - S_{22}^e)e^{j\theta} \tag{8.119}$$

Finally, substituting (8.109) into (8.119), we get

$$S_{12}^e = \frac{2\sqrt{2}Z_o(Z_{o1}\cos\theta + j2Z_o\sin\theta)e^{j\theta}}{3Z_oZ_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} \tag{8.120}$$

8.4.2 Odd-Mode Analysis

Figure 8.40(a) shows the Wilkinson power divider under the odd-mode operation in which ports 2 and 3 are excited with opposite potential $\pm A_2/2$. This results in a short circuit or an EW along the symmetry plane. Correspondingly, the potential at port 1 under the odd-mode excitation is zero and, hence, $a_1^o = -b_1^o$. It is noted that a_1^o can take on any value as long as the total potential at port 1 is zero. For the convenience of analysis, we let $a_1^o = -b_1^o = 0$ as noted in Figure 8.40(a); this does not affect the results since the potential at port 1 under the odd mode is always zero. Due to the short circuit, the power divider is divided into two independent identical halves – one of which between ports 1 and 2 is shown in Figure 8.40(b).

The input impedance $Z_{\text{in},2}^o$ at port 2 consists of the impedance looking into the short-circuited transmission line of characteristic impedance Z_{o1} and the resistance $R/2$ in parallel, and hence can be obtained as

$$Z_{\text{in},2}^o = \frac{RZ_{o1}\tan\theta}{2Z_{o1}\tan\theta - jR} \tag{8.121}$$

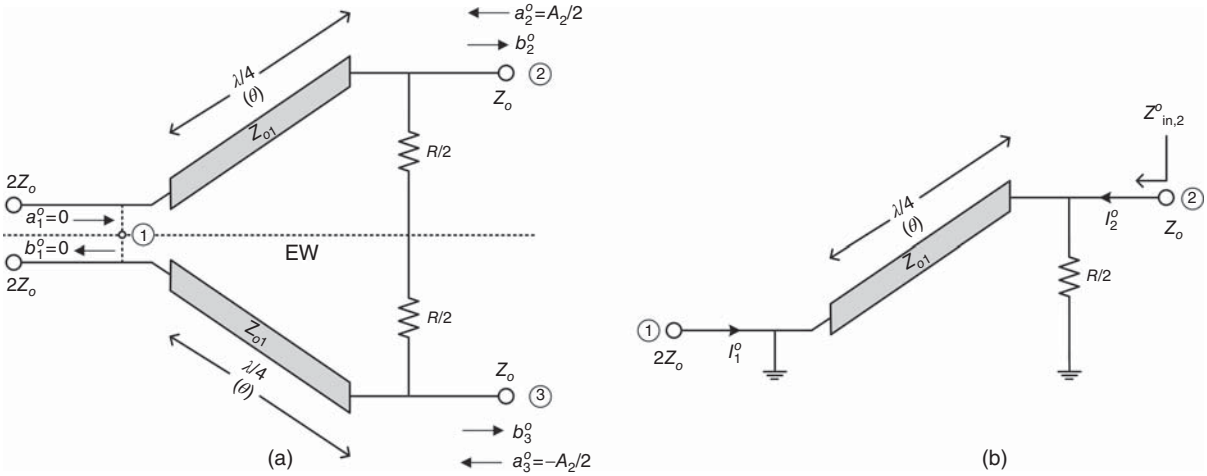


Figure 8.40. Wilkinson power divider under odd-mode excitation (a) and the resultant half-circuit (b).

The corresponding reflection coefficient is

$$S_{22}^o = \frac{Z_{\text{in},2}^o - Z_o}{Z_{\text{in},2}^o + Z_o} = \frac{Z_{o1}\sin\theta(R - 2Z_o) + jRZ_o\cos\theta}{Z_{o1}\sin\theta(R + 2Z_o) - jRZ_o\cos\theta} \tag{8.122}$$

while the transmission between ports 2 and 1 is

$$S_{12}^o = \left.\frac{b_1^o}{a_2^o}\right|_{a_1^o=0} = 0 \tag{8.123}$$

8.4.3 Superimposition of Even and Odd Modes

We now superimpose the even and odd modes as illustrated in Figure 8.41 to obtain

$$\begin{aligned} a_1 &= a_1^e + a_1^o = a_1^e = A_1 \\ a_2 &= a_2^e + a_2^o = 2a_2^e = 2a_2^o = 2a_3^e = -2a_3^o = A_2 \\ a_3 &= a_3^e + a_3^o = 0 \end{aligned} \tag{8.124}$$

$$\begin{aligned} b_1 &= b_1^e + b_1^o \\ b_2 &= b_2^e + b_2^o \\ b_3 &= b_3^e + b_3^o \end{aligned} \tag{8.125}$$

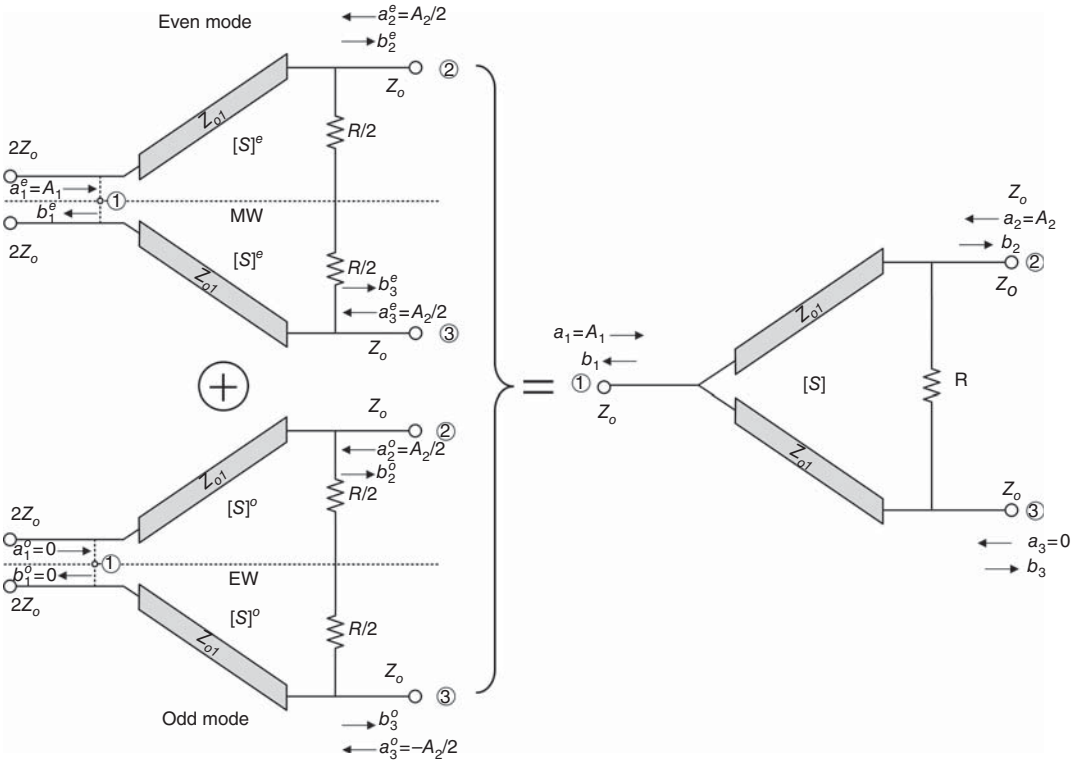


Figure 8.41. Superimposition of even and odd modes. $[S]^e$ and $[S]^o$ represent the S-matrix of half of the power divider under the even- and odd-mode operation, respectively, and $[S]$ represents the S-matrix of the power divider.

The S-matrix equations for the coupler can be derived from (8.125) and (8.124) as

$$\begin{aligned} b_1 &= S_{11}^e a_1^e + S_{12}^e a_2^e + S_{11}^o a_1^o + S_{12}^o a_2^o = S_{11}^e a_1 + \frac{1}{2}(S_{12}^e + S_{12}^o)a_2 \\ b_2 &= S_{21}^e a_1^e + S_{22}^e a_2^e + S_{21}^o a_1^o + S_{22}^o a_2^o = S_{21}^e a_1 + \frac{1}{2}(S_{22}^e + S_{22}^o)a_2 \\ b_3 &= S_{21}^e a_1^e + S_{22}^e a_3^e + S_{21}^o a_1^o + S_{22}^o a_3^o = S_{21}^e a_1 + \frac{1}{2}(S_{22}^e - S_{22}^o)a_2 \end{aligned} \tag{8.126}$$

or, in matrix form,

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11}^e & \frac{1}{2}\left(S_{12}^e + S_{12}^o\right) \\ S_{21}^e & \frac{1}{2}(S_{22}^e + S_{22}^o) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{8.127}$$

$$\begin{bmatrix} b_1 \\ b_3 \end{bmatrix} = \begin{bmatrix} S_{11}^e & \frac{1}{2}\left(S_{12}^e + S_{12}^o\right) \\ S_{21}^e & \frac{1}{2}(S_{22}^e - S_{22}^o) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \tag{8.128}$$

Hence,

$$\begin{aligned} S_{11} &= S_{11}^e \\ S_{12} &= \frac{1}{2}(S_{12}^e + S_{12}^o) \\ S_{21} &= S_{31} = S_{21}^e \\ S_{22} &= \frac{1}{2}(S_{22}^e + S_{22}^o) \\ S_{32} &= \frac{1}{2}(S_{22}^e - S_{22}^o) \end{aligned} \tag{8.129}$$

from which, we can derive upon using (8.107), (8.109), (8.115), (8.120), (8.122), and (8.123) and the symmetry between ports 2 and 3:

$$S_{11} = \frac{-Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 - 2Z_o^2)\sin\theta}{3Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} \tag{8.130}$$

$$S_{12} = S_{13} = \frac{\sqrt{2}Z_o(Z_{o1}\cos\theta + j2Z_o\sin\theta)e^{j\theta}}{3Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} \tag{8.131}$$

$$S_{21} = S_{31} = \frac{2\sqrt{2}Z_o(Z_{o1}\cos\theta + jZ_o\sin\theta)e^{-j\theta}}{3Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} \tag{8.132}$$

$$S_{22} = S_{33} = \frac{1}{2}\left[\frac{Z_o Z_{o1}\cos\theta + j\left(Z_{o1}^2 - 2Z_o^2\right)\sin\theta}{3Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} + \frac{Z_{o1}\sin\theta(R - 2Z_o) + jRZ_o\cos\theta}{Z_{o1}\sin\theta(R + 2Z_o) - jRZ_o\cos\theta}\right] \tag{8.133}$$

$$S_{32} = S_{23} = \frac{1}{2}\left[\frac{Z_o Z_{o1}\cos\theta + j\left(Z_{o1}^2 - 2Z_o^2\right)\sin\theta}{3Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 + 2Z_o^2)\sin\theta} - \frac{Z_{o1}\sin\theta(R - 2Z_o) + jRZ_o\cos\theta}{Z_{o1}\sin\theta(R + 2Z_o) - jRZ_o\cos\theta}\right] \tag{8.134}$$

Equations (8.130) and (8.133) can be used to derive useful design information for the Wilkinson power divider. Letting $S_{11} = 0$, we obtain the criterion for perfect match at port 1 as

$$-Z_o Z_{o1}\cos\theta + j(Z_{o1}^2 - 2Z_o^2)\sin\theta = 0 \tag{8.135}$$

which can be solved for Z_{o1} and θ. One of the solutions is

$$\theta = \frac{\pi}{2}$$
$$Z_{o1} = Z_o\sqrt{2} \tag{8.136}$$

which is the fundamental design formula for the Wilkinson power divider. Substituting (8.136) into (8.131), (8.132) gives

$$S_{12} = S_{13} = S_{21} = S_{31} = \frac{1}{\sqrt{2}} \tag{8.137}$$

If we further choose $R = 2Z_o$, then $S_{22} = S_{33} = 0$ and $S_{32} = S_{23} = 0$ as can be deduced from (8.133) and (8.134), respectively. Therefore, under the design criteria of $\theta = \pi/2$, $Z_{o1} = Z_o\sqrt{2}$ and $R = 2Z_o$, the S-matrix of the Wilkinson power divider becomes

$$[S] = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{8.138}$$

which completely describes the operation of an ideal Wilkinson power divider. The foregoing analysis is obtained at the design center frequency corresponding to the electrical length θ equal to 90°. When the frequency is deviated from the center frequency, the performance degrades from the optimum performance due to the variation of θ as a function of frequency as seen in (8.96). As for the ring hybrid and branch-line coupler, the operating bandwidth of the Wilkinson power divider primarily depends on the specification of the amplitude balance.

8.5 FILTERS

Filters, as the name implies, are used to pass desired signals at certain frequencies and reject unwanted signals at other frequencies. Filters are indispensable in RFICs, and they can be used as a standalone component or as an element in circuits such as mixers or in systems such as receivers. There are four types of filters: low pass filter, high pass filter, band-stop filter, and band-pass filter. Figure 8.42 shows some representative schematics of these filters and their frequency responses. Among these filters, the low pass filter is the most fundamental one and, in fact, its prototypes, as shown in Figure 8.51, serves as the model from which all the low pass, high pass, band-pass and band-stop filters are derived. In general, there are two kinds of filters: maximally flat (or Butterworth) and Chebyshev (or equal-ripple). The materials in this section are drawn from [10]. Similar to other RFICs, filters can be realized using lumped elements, transmission lines or a combination of these. Their performances are typically specified by the return loss, insertion loss and ripple level (for Chebyshev filters) in the pass band, and the rejection in the stop band.

8.5.1 Low Pass Filter

Figure 8.42(a) shows a schematic of low pass filters and Figure 8.43 shows the response of an ideal and actual low pass filter. The rejection in the stop band of an actual low pass filter is finite with finite selectivity, while that of an ideal low pass filter reaches infinity with an infinite slope. The characteristics and the circuit elements of low pass filters can be derived using a method based on the attenuation of the filters, which is basically a network synthesis method based on transfer functions, in which the transfer function is specified by the attenuation of the low pass filters.

In the attenuation method, a desired attenuation response as a function of frequency (frequency response) is used to approximate the response of an ideal low pass filter. Figure 8.44 illustrates this principle in which a

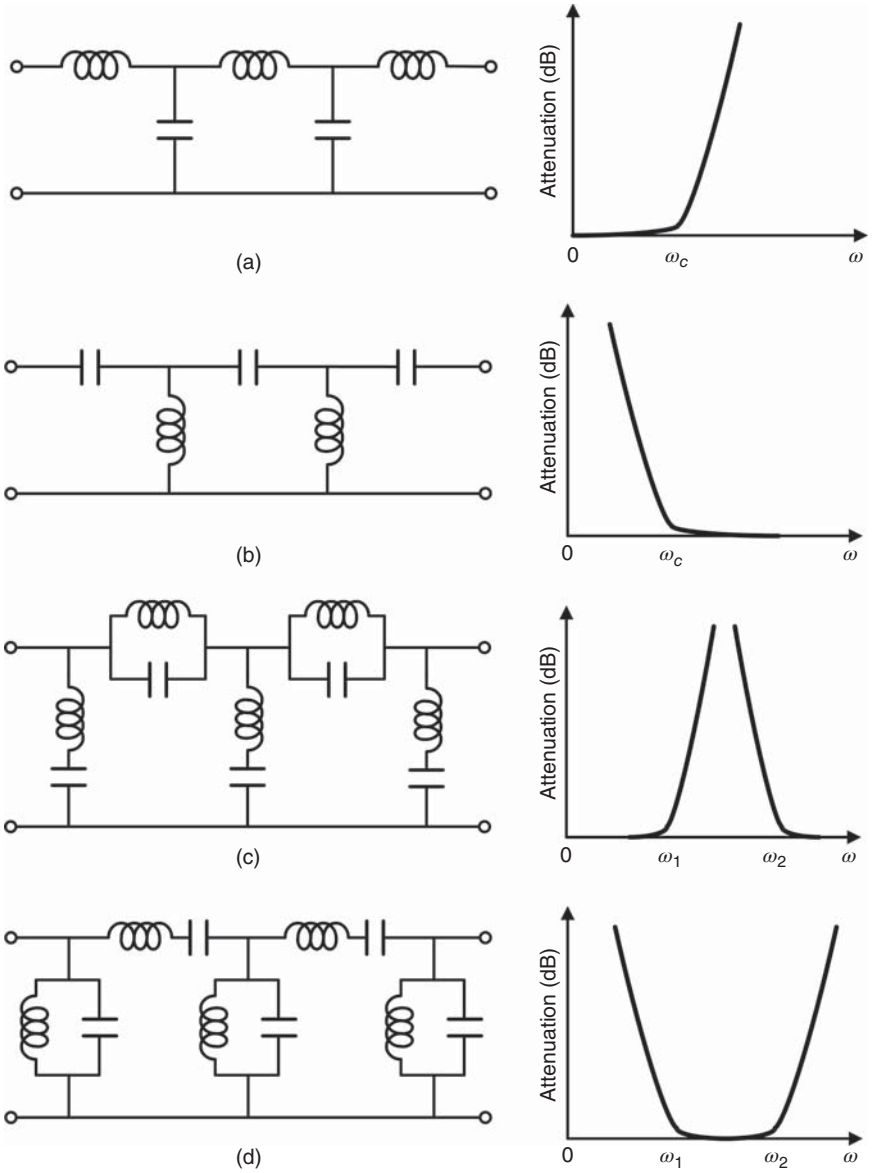


Figure 8.42. Schematic and sketch of frequency response of a (five-element) low pass filter (a), high pass filter (b), band-stop filter (b), and band-pass filter (d).

Chebyshev response is used to match the ideal low pass filter response. The attenuation for a low pass filter, or any passive network, assuming lossless (no resistive elements) is given as

$$A \triangleq |T|^2 = \frac{P_{\text{out}}}{P_{\text{in}}} = 1 - |\Gamma|^2 \tag{8.139}$$

where T and Γ are the (voltage) transmission and reflection coefficient, respectively, and P_{in} and P_{out} are the respective input and output power. This attenuation represents the loss due to mismatch (or reflection) of the filter. The transmission coefficient, reflection coefficient, or attenuation characterizes the low pass filter response and hence a desired transmission function, reflection function, or attenuation can be used in (8.139) for the design of low pass filters. This basically carries out a synthesis formulation to determine the filter elements. Although mathematically (8.139) indicates that any function for the transmission coefficient, reflection coefficient, or attenuation can be used to determine the filter elements, in practice, only certain

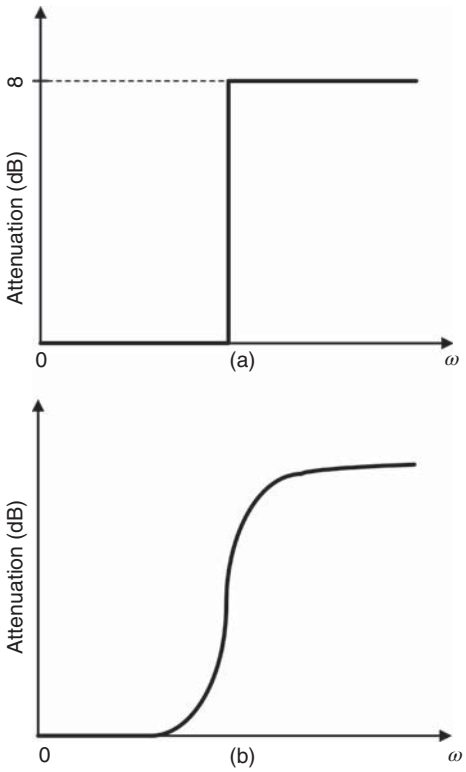


Figure 8.43. Ideal (a) and actual (b) response of low pass filters.

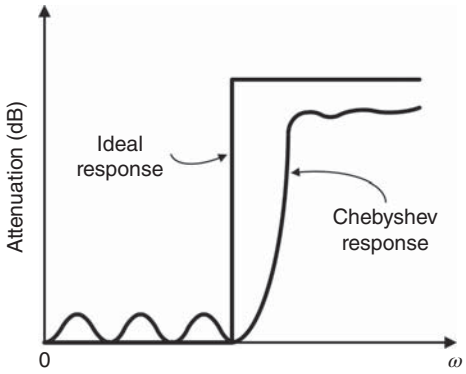


Figure 8.44. Responses of Chebyshev and ideal low pass filters.

functions can be used in order to lead to filter elements that are physically realizable. This imposes certain restrictions on these coefficients. For passive networks, $|\Gamma(\omega)| \leq 1$ and the necessary and sufficient condition for realizability of the filter elements is that the attenuation must be of the following form [4]:

$$A = 1 + \frac{P(\omega^2)}{Q(\omega^2)} \tag{8.140}$$

where $P(\omega^2)$ and $Q(\omega^2)$ are real polynomials in ω^2. Equations (8.139) and (8.140) hold the key for the low pass filter design in which any proper function [conforming (8.140)] can be employed to realize any desired filter response. The most fundamental filter responses are maximally flat and Chebyshev responses.

8.5.1.1 Attenuation in Maximally Flat Low Pass Filter. The attenuation in maximally flat low pass filter is defined by the following form:

$$A_m = 1 + \alpha\left(\frac{\omega}{\omega_c}\right)^{2n} \tag{8.141}$$

where n is the number of low pass filter elements representing the order of the low pass filter, ω_c is the cut-off frequency of the low pass filter, and α is a constant representing the maximum pass-band loss. The attenuation, as can be seen from (8.141), increases as the frequency is increased, reaching a maximum pass-band loss of $1 + \alpha$ at the cut-off frequency ω_c. It is a common practice to refer to the pass-band loss as insertion loss and stop-band loss as rejection and we will adapt these terminologies for filters. Letting α_{dB} be the maximum insertion loss (in dB) imposed on the filter, we can then obtain from (8.141):

$$\alpha = 10^{0.1\alpha_{\text{dB}}} - 1 \tag{8.142}$$

For the maximally flat low pass filters, the maximum insertion loss α or α_{dB} always occurs at the cutoff frequency. Another word, the cut-off frequency of the maximally flat low pass filters is defined at the point corresponding to the maximum insertion loss. Typically, this point is chosen at 3-dB loss. Setting $\alpha_{\text{dB}} = 3\,\text{dB}$ in (8.142) leads to $\alpha = 1$. In the stop band where $\omega > \omega_c$, the rejection increases as $2n$th power of the normalized frequency ω/ω_c. As the filter order n is increased, the rejection rises with increasing slope (sharper cutoff). At frequencies very high compared to the cutoff frequency, the rejection is approximately proportional to $\alpha(\omega/\omega_c)^{2n}$ as indicated by (8.141). Figure 8.45 illustrates the general behavior of the maximally flat low pass filters. Figure 8.46 displays the attenuation of the maximally flat low pass filters for different orders calculated using (8.141) with the cutoff frequency defined at the 3-dB point for insertion loss.

8.5.1.2 Attenuation in Chebyshev Low Pass Filter. The attenuation response of an nth order Chebyshev low pass filter is characterized by

$$A_c = 1 + \alpha T_n^2\left(\frac{\omega}{\omega_c}\right) \tag{8.143}$$

where T_n is the Chebyshev polynomial of degree n described as

$$T_n(x) = \begin{cases} \cos\left(n\cos^{-1}x\right), & x \le 1 \\ \cosh(n\cosh^{-1}x), & x \ge 1 \end{cases} \tag{8.144}$$

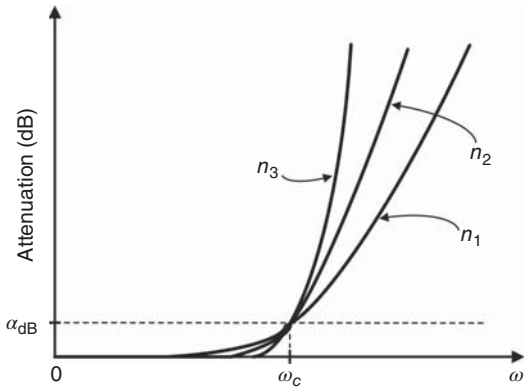


Figure 8.45. Response of the maximally flat low pass filters. $n_3 > n_2 > n_1$.

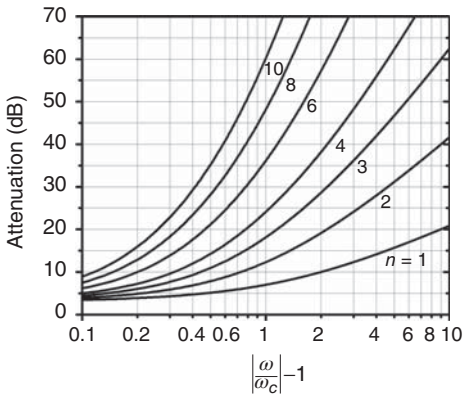


Figure 8.46. Attenuation of the maximally flat low pass filters with 3-dB maximum insertion loss.

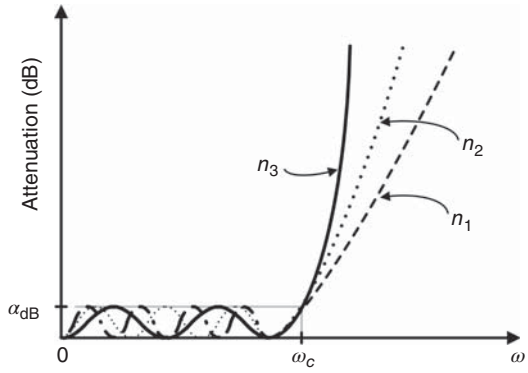


Figure 8.47. Response of the Chebyshev low pass filters. $n_3 > n_2 > n_1$.

Substituting (8.144) into (8.143) gives the attenuation of the Chebyshev low pass filters as

$$A_c = \begin{cases} 1 + \alpha \ \cos^2\left[n\cos^{-1}\left(\dfrac{\omega}{\omega_c}\right)\right], & \omega \le \omega_c \\ 1 + \alpha \ \cosh^2\left[n\cosh^{-1}\left(\dfrac{\omega}{\omega_c}\right)\right], & \omega \ge \omega_c \end{cases} \tag{8.145}$$

The insertion loss constant α is still given by (8.142) and α_{dB} is still the maximum insertion loss (in dB), but this maximum loss does not necessarily occur at the cut-off frequency. Examination of (8.145) shows that the insertion loss oscillates between the minimum value of 1 (or 0 dB) and maximum value of $1 + \alpha$ [or $\alpha_{\text{dB}} = 10\log(1 + \alpha)$] within the pass band, which is due to the nature of the Chebyshev polynomial for $\omega \le \omega_c$. Hence, α_{dB} is often referred to as the pass-band ripple for the Chebyshev filters. In the stop band where $\omega > \omega_c$, the rejection increases monotonically as the filter order n is increased, leading to increased rejection and attenuation slope. At $\omega \gg \omega_c$, the rejection is approximately equal to $\frac{\alpha_{\text{dB}}}{4}\left(\frac{2\omega}{\omega_c}\right)^{2n}$ as can be inferred from (8.145). Figure 8.47 illustrates the general behavior of the Chebyshev low pass filters. Figures 8.48 and 8.49 show the attenuation of the Chebyshev low pass filters for different orders calculated using (8.145) with the pass-band ripple of 0.1 and 1 dB.

Examination of the attenuation in the stop band for the maximally flat and Chebyshev low pass filters characterized in (8.141) and (8.145), respectively, shows that the rejection (dB) in the stop-band region of these filters are approximately related by

$$A_c(\text{dB}) \simeq A_m(\text{dB}) + 6.02(n - 1) \tag{8.146}$$

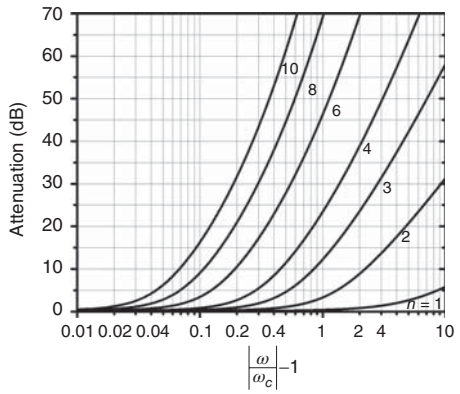


Figure 8.48. Attenuation of the Chebyshev low pass filters for 0.1-dB pass-band ripple.

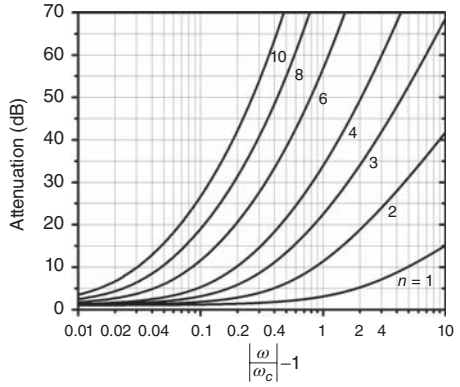


Figure 8.49. Attenuation of the Chebyshev low pass filters for 1-dB pass-band ripple.

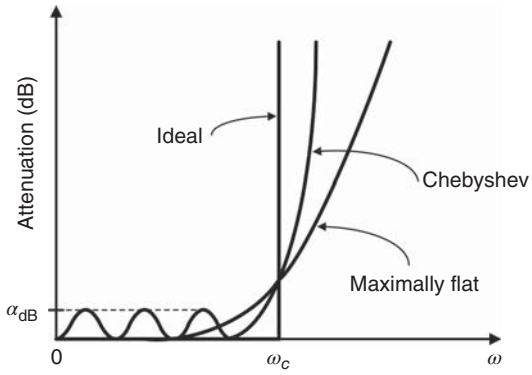


Figure 8.50. Comparison between Chebyshev and maximally flat low pass filter responses.

which demonstrates that, for given maximum insertion loss (α_{dB}) and filter order (n), the Chebyshev filter has much sharper rate of cutoff than the maximally flat filter and is the preferred filter with respect to the stop-band rejection. For instance, the added rejection is around 24 dB for $n = 5$, which is substantial. Figure 8.50 illustrate this phenomenon.

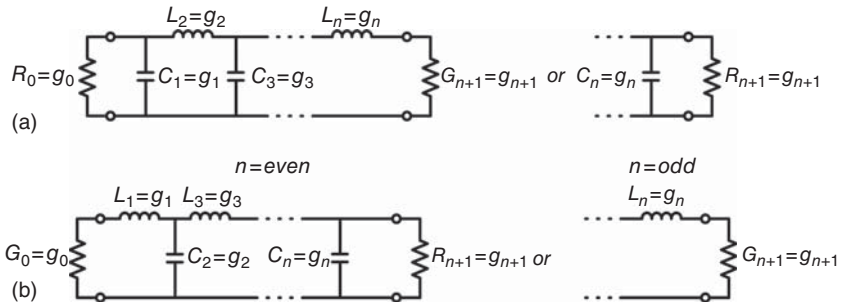


Figure 8.51. Low pass filter prototypes.

8.5.1.3 Low Pass Filter Prototypes

Topology of Prototypes

Low pass filter prototypes form the basis for all of the filters including low pass filters, from which these filters are derived, and as such are viewed as the most fundamental filter configurations. Figure 8.51 shows two principal low pass topologies which establish the low pass filter prototypes. These low pass filter prototypes are dual of each other and produce identical responses. In these prototypes, either end can be used as the source or load impedance due the reciprocity of the passive network. To generalize the results to be obtained for the elements of the low pass filter prototypes from the subsequent formulation, the prototype elements are designated with common variables (g's) as indicated in Figure 8.51 and explained below:

$$g_k(k = 1, 2, \ldots, n) = \text{inductance of a series inductor or capacitance of a shunt capacitor}$$

$$g_o = \begin{cases} \text{source resistance } R_o \text{ if } g_1 = C_1 \\ \text{source conductance } G_o \text{ if } g_1 = L_1 \end{cases}$$

$$g_{n+1} = \begin{cases} \text{load resistance } R_{n+1} \text{ if } g_n = C_n \\ \text{load conductance } G_{n+1} \text{ if } g_n = L_n \end{cases}$$

Prototype Elements

The elements (inductors and capacitors) in a prototype corresponding to a desired response can be determined by forcing the attenuation calculated from the prototype to the desired response (i.e., attenuation function) as mentioned earlier. Herein, we consider the two most commonly used responses: maximally flat and Chebyshev. Specifically, the reflection coefficient is determined from the prototype and used for calculating the attenuation as given in (8.139). This attenuation is set to be equal to the maximally flat or Chebyshev attenuation function described in (8.141) or (8.145), respectively. Corresponding terms in both sides of the resultant equation are then matched to determine the prototype element values $g_1, g_2, \ldots, g_n$. This procedure, although straight forward, is laborious and time consuming, particularly when the number of elements is greater than 4. A common procedure is to compute the element values using analytical equations for the maximally flat and Chebyshev low pass filter prototypes.

Maximally Flat Prototype Elements. We assume that the source and load terminations are resistors as typically the case and the insertion loss (or maximum pass-band loss) is 3 dB. We also assume that all values are normalized with respect to the source termination and the cutoff frequency ω_c'; that is, $g_o = 1$ and $\omega_c' = 1$ rad/s.

TABLE 8.2. Element Values for 3-dB Maximally Flat Low Pass Filter Prototypes ($g_o = 1$, $\omega_c' = 1$ rad/s, $\alpha_{dB} = 3$ dB)

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
1	2.0000	1.0000									
2	1.4142	1.4142	1.000								
3	1.0000	2.0000	1.0000	1.0000							
4	0.7654	1.8478	1.8478	0.7654	1.000						
5	0.6180	1.6180	2.0000	1.6180	0.6180	1.0000					
6	0.5176	1.4142	1.9318	1.9318	1.4142	0.5176	1.0000				
7	0.4450	1.2470	1.8019	2.0000	1.8019	1.2470	0.4450	1.0000			
8	0.3902	1.1111	1.6629	1.9615	1.9615	1.6629	1.1111	0.3902	1.0000		
9	0.3473	1.0000	1.5321	1.8794	2.0000	1.8794	1.5321	1.0000	0.3473	1.0000	
10	0.3129	0.9080	1.4142	1.7820	1.9754	1.9754	1.7820	1.4142	0.9080	0.3129	1.0000

The elements of the maximally flat low pass filter prototypes can be determined from the following equations:

$$\begin{aligned} g_o &= 1 \\ g_k &= 2 \sin\left[\frac{(2k-1)\,\pi}{2n}\right], \quad k = 1, 2, \ldots, n \\ g_{n+1} &= 1 \end{aligned} \tag{8.147}$$

where g_k is in henry (H) for inductance and farad (F) for capacitance. It is noted that both impedance and frequency normalization are used in (8.147), which is necessary for generalization of the results. The elements for actual filters can be easily obtained by de-normalizing or scaling as discussed later. Table 8.2 lists the element values for the 3-dB maximally flat filters with order of 1–10 calculated using (8.147).

Chebyshev Prototype Elements. As for the maximally flat filter prototypes, it is assumed that the source and load terminations are resistors, and $g_o = 1$ and $\omega_c' = 1$ rad/s. The elements of the Chebyshev low pass filter prototypes can be determined from the following equations:

$$g_o = 1$$

$$g_1 = \frac{2a_1}{p}$$

$$g_k = \frac{4a_{k-1}a_k}{a_{k-1}a_k}, \quad k = 1, 2, \ldots, n$$

$$g_{n+1} = \begin{cases} 1, & n \text{ odd} \\ \coth^2\left(\dfrac{q}{4}\right), & n \text{ even} \end{cases} \tag{8.148}$$

where

$$q = \ln\left(\coth\frac{\alpha_{dB}}{17.37}\right)$$

$$p = \sinh\frac{q}{2n}$$

$$a_k = \sin\left[\frac{(2k-1)\,\pi}{2n}\right]$$

$$b_k = p^2 + \sin^2\left(\frac{k\pi}{n}\right)$$

with α_{dB} being the pass-band ripple or insertion loss in decibels, g_k is in henry (H) for inductance and farad (F) for capacitance. The values calculated from (8.148) are normalized with respect to the source termination ($g_o = 1$) and cutoff frequency ($\omega_c' = 1\,\text{rad/s}$).

It is interesting to note that for the maximally flat prototypes, the source and load impedances are equal regardless whether the filter order n is odd or even while, for the Chebyshev prototypes, the source and load impedances are equal and unequal for odd and even order, respectively. Therefore, even-order Chebyshev filters possess impedance-transformation property, which may be exploited for certain impedance matching besides low pass filtering, or requires a matching network if equal source and load impedances are desired. Tables 8.3 and 8.4 list the element values for the Chebyshev filter prototypes for 0.1 and 1 dB pass-band ripples, respectively, and order of 1–10 calculated using (8.148).

Impedance and Frequency Scaling. In the foregoing low pass filter prototypes, we assume that the source or load impedance and cutoff frequency are normalized ($g_o = 1$ and $\omega_c' = 1$ rad/s). In actual low pass filters,

TABLE 8.3. Element Values for Chebyshev Low Pass Filter Prototypes with 0.1-dB Pass-Band Ripple ($g_o = 1$, $\omega_c' = 1$ rad/s, $\alpha_{\text{dB}} = 0.1$ dB)

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
1	0.3052	1.0000									
2	0.8430	0.6220	1.3554								
3	1.0315	1.1474	1.0315	1.0000							
4	1.1088	1.3061	1.7703	0.8180	1.3554						
5	1.1468	1.3712	1.9750	1.3712	1.1468	1.0000					
6	1.1681	1.4039	2.0562	1.5170	1.9029	0.8618	1.3554				
7	1.1811	1.4228	2.0966	1.5733	2.0966	1.4228	1.1811	1.0000			
8	1.1897	1.4346	2.1199	1.6010	2.1699	1.5640	1.9444	0.8778	1.3554		
9	1.1956	1.4425	2.1345	1.6167	2.2053	1.6167	2.1345	1.4425	1.1956	1.0000	
10	1.1999	1.4481	2.1444	1.6265	2.2253	1.6418	2.2046	1.5821	1.9628	0.8853	1.3554

TABLE 8.4. Element Values for Chebyshev Low Pass Filter Prototypes with 1-dB Pass-Band Ripple ($g_o = 1$, $\omega_c' = 1$ rad/s, $\alpha_{\text{dB}} = 1$ dB)

n	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}
1	1.0177	1.0000									
2	1.8219	0.6850	2.6599								
3	2.0236	0.9941	2.0236	1.0000							
4	2.0991	1.0644	2.8311	0.7892	2.6599						
5	2.1349	1.0911	3.0009	1.0911	2.1349	1.0000					
6	2.1546	1.1041	3.0634	1.1518	2.9367	0.8101	2.6599				
7	2.1664	1.1116	3.0934	1.1736	3.0934	1.1116	2.1664	1.0000			
8	2.1744	1.1161	3.1107	1.1839	3.1488	1.1696	2.9685	0.8175	2.6599		
9	2.1797	1.1192	3.1215	1.1897	3.1747	1.1897	3.1215	1.1192	2.1797	1.0000	
10	2.1836	1.1213	3.1268	1.1933	3.1890	1.1990	3.1738	1.1763	2.9824	0.8210	2.6599

however, both g_o and cutoff frequency are different from 1. This can be resolved by scaling the element values of the low pass prototypes to the desired impedance and cutoff frequency of the actual filters.

Impedance Scaling. In the impedance scaling, we assume the source impedance is to be transferred from 1 (of the prototype) to R_s (of the actual filter). The impedance-scaled inductance, capacitance, resistance, and conductance can be easily derived as:

$$\begin{aligned} L_k &= R_s L'_k \qquad C_k = \frac{C'_k}{R_s} \\ R &= R_s R' \qquad G = \frac{G'}{R_s} \end{aligned} \tag{8.149}$$

where the primed and unprimed parameters are for the prototype and actual filters, respectively.

Frequency Scaling. In the frequency scaling, the cutoff frequency is transformed from $\omega'_c = 1$ for the prototype to $\omega_c \neq 1$ for the actual filter. The transformation is described as flows.

Low pass filter prototype	Low pass filter
$\omega'_c = 1$ rad/s	ω_c
ω'	$\frac{\omega}{\omega_c}$
$j\omega' L'_k$	$j\frac{\omega}{\omega_c} L'_k$
$j\omega' C'_k$	$j\frac{\omega}{\omega_c} C'_k$

The frequency-scaled inductance and capacitance can then be obtained as

$$\begin{aligned} L_k &= \frac{L'_k}{\omega_c} \\ C_k &= \frac{C'_k}{\omega_c} \end{aligned} \tag{8.150}$$

Impedance and Frequency Scaling. Combing the impedance and frequency scaling leads to the following results:

Low pass filter prototype	Low pass filter
$R_o = 1\ \Omega, \omega'_c = 1$ rad/s	R_s, ω_c
R', G'	$R_s R', \frac{G'}{R_s}$
L'_k	$\frac{R_s L'_k}{\omega_c}$
C'_k	$\frac{C'_k}{\omega_c R_s}$

The low pass filter elements after the impedance and frequency scaling are now given as

$$L_k = \frac{R_s L'_k}{\omega_c}$$
$$C_k = \frac{C'_k}{\omega_c R_s} \tag{8.151}$$

It is noted that the frequency transformation from ω' to ω/ω_c serves as the basic transformation from the low pass filter prototype to a low pass filter. The low pass filter prototype can also be transferred to high pass, band-stop and band-pass filters using appropriate frequency transformation to be discussed later. Furthermore, (8.141) and (8.143) can be used to calculate the attenuation of any type of filters provided that a suitable normalized frequency ω/ω_c that maps the response of the low pass filter prototype to that of the corresponding filter is used.

8.5.1.4 Low Pass Filter Design

Lumped-Element Low Pass Filter Design

Lumped-element low pass filters can be designed directly from the low pass filter prototypes using the following procedure. First, the filter order or number of filter elements is determined from the filter specifications (maximum insertion loss, cutoff frequency, and stop-band rejection at a certain frequency) using the attenuation equations (8.141) and (8.143) or (8.145) for the maximally flat and Chebyshev filters, respectively, or appropriate attenuation curves such as those in Figures 8.46 and 8.48. Second, the element values $g_k (k = 0, 1, 2, \ldots, n+1)$ of the low pass filter prototype are calculated using (8.147) or (8.148), or appropriate tables listing these values. Third, the prototype values are converted to the values of the actual filter having $R_s \neq 1\ \Omega$, $\omega_c \neq 1$ rad/s using the foregoing impedance and frequency scaling equations. It is note that R_s is typically equal to 50 Ω.

As an example, we consider a maximally flat low pass filter having a maximum insertion loss of 3 dB, cutoff frequency of 20 GHz, and minimum rejection of 30 dB at 40 GHz. Using the attenuation equation in (8.141), we can obtain, for $\alpha_{\text{dB}} = 3$ dB, $f_c = 20$ GHz and $A_m = 30$ dB $= 1000$ at $f = 40$ GHz, $1000 = 1 + (4/2)^{2n}$, from which $n = 4.98$. Hence, the filter order must be equal or greater than 5.

On another example, we consider a Chebyshev low pass filter having 0.1-dB passband ripple, 20-GHz cutoff frequency, and a minimum rejection of 30 dB at 40 GHz. The source impedance R_s is assumed to be 50 Ω. Using the attenuation equation (8.145) or the attenuation curves in Figure 8.48, we see that the filter order n must be greater than 4, and hence 5 is chosen as the filter order. Using (8.148) or Table 8.3, we obtain $R_o = g_o = 1\ \Omega$, $C_1 = g_1 = 1.1468$ F, $L_2 = g_2 = 1.3712$ H, $C_3 = g_3 = 1.9750$ F, $L_4 = g_4 = 1.3712$, $C_5 = g_5 = 1.1468$, and $G_6 = g_6 = 1\ \text{℧}$. Scaling these values corresponding to $R_s = 50\ \Omega$ and $f_c = 20$ GHz, we then get the element values for the low pass filter as:

$$R_s = R'_o = 50\Omega \qquad G_L = G'_6 = \frac{g_6}{R_s} = 0.02\text{℧}$$
$$C'_1 = \frac{g_1}{\omega_c R_s} = 0.1825 \text{ pF} \qquad L'_2 = \frac{R_s g_2}{\omega_c} = 0.5456 \text{ nH} \qquad C'_3 = \frac{g_3}{\omega_c R_s} = 0.3143 \text{ pF}$$
$$L'_4 = \frac{R_s g_4}{\omega_c} = 0.5456 \text{ nH} \qquad C'_5 = \frac{g_5}{\omega_c R_s} = 0.1825 \text{ pF}$$

Figure 8.52 shows the schematic of the designed low pass filter.

Transmission-Line Low Pass Filter Design

The inductors and capacitors used in low pass filters can also be realized using transmission lines. In general, a series inductor is equivalent to a transmission line having high characteristic impedance (typically chosen

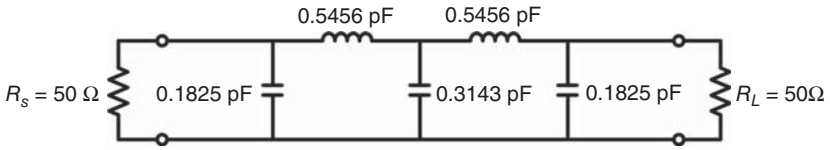


Figure 8.52. Designed Chebyshev low pass filter. R_s and R_L are the source and load impedances, respectively.

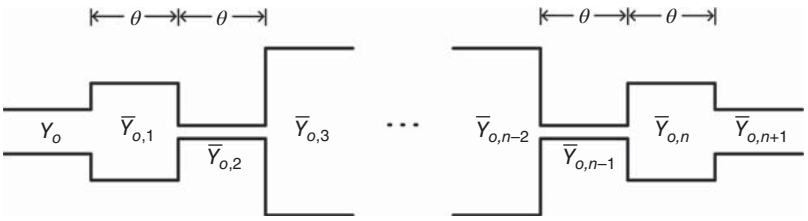


Figure 8.53. Transmission-line low pass filter. $\overline{Y}_{o,k}$ is the normalized characteristic admittance with respect to Y_o.

as high as possible) and short length (typically less than 1/8th of the wavelength at the cutoff frequency), while a shunt capacitor can be replaced with a transmission line having low characteristic impedance (typically chosen as low as possible) and short length (typically less than 1/8th of the wavelength at the cutoff frequency). A low pass filter realized using transmission lines thus consists of cascade of transmission lines having low and high characteristic impedances. This kind of filter is also referred to as semi-lumped element low pass filter.

Another kind of transmission-line low pass filters is shown in Figure 8.53, which consists of n cascaded commensurate transmission lines of varying characteristic admittances, each transmission line being one-quarter wavelength long at the stopband center frequency. The principal advantage of this filter over the semi-lumped element low pass filter is that it provides more linear phase response. The required normalized characteristic admittances for given filter specifications can be determined from the exact synthesis in [11]. However, because of the difficulty associated with this numerical technique, it is not very suitable for computer-aided design. Explicit formulas derived in [12] are more suitable for the filter design and given as follows.

Maximally Flat Low Pass Filter. The normalized characteristic admittances $\overline{Y}_{o,k}(k = 1, 2, \ldots, n)$ and $\overline{Y}_{o,n+1}$ of the transmission-line sections from 1 to $n + 1$ with respect to the characteristic admittance Y_o are obtained from

$$\overline{Y}_{o,k} = \begin{cases} G_k, & k \text{ odd} \\ \dfrac{1}{G_k}, & k \text{ even} \end{cases} \tag{8.152}$$

and

$$\overline{Y}_{o,n+1} = 1 \tag{8.153}$$

where

$$G_k = \frac{2\sin\left[(2k-1)\frac{\pi}{2n}\right]}{d}\left\{1 - \frac{\cos\left(\frac{\pi}{n}\right)d^2}{4\sin\left[(2k-3)\frac{\pi}{2n}\right]\sin\left[(2k+1)\frac{\pi}{2n}\right]}\right\} \tag{8.154}$$

with d being the bandwidth scaling factor.

Chebyshev Low Pass Filter. The normalized characteristic admittances $Y_{o,k}(k = 1, 2, \ldots, n)$ and $Y_{o,n+1}$ of the transmission lines with respect to the characteristic admittance Y_o are obtained from

$$\overline{Y}_{o,k} = \begin{cases} G_k, & k \text{ odd} \\ \dfrac{1}{G_k}, & k \text{ even} \end{cases} \tag{8.155}$$

and

$$\overline{Y}_{o,n+1} = \begin{cases} 1, & k \text{ odd} \\ \dfrac{\sqrt{1-\varepsilon^2}-\varepsilon}{\sqrt{1+\varepsilon^2}+\varepsilon}\tanh^2\left(\dfrac{n}{2}\sinh^{-1}y\right), & k \text{ even} \end{cases} \tag{8.156}$$

where

$$G_k = A_k\left[\frac{2\sin\left[(2k-1)\frac{\pi}{2n}\right]}{d} - \frac{d}{4}\left\{\frac{y^2+\sin^2\left(\frac{k\pi}{n}\right)}{\sin\left[(2k+1)\frac{\pi}{2n}\right]} + \frac{y^2+\sin^2\left[(k-1)\frac{\pi}{n}\right]}{\sin\left[(2k-3)\frac{\pi}{2n}\right]}\right\}\right], \quad k = 1 \text{ to } n \tag{8.157}$$

with

$$y = \sinh\left(\frac{1}{n}\sinh^{-1}\frac{1}{\varepsilon}\right) \tag{8.158}$$

$$\varepsilon = \sqrt{\alpha} = \sqrt{10^{0.1\alpha_{\text{dB}}}-1} \tag{8.159}$$

The insertion loss parameters α_{dB} and α are related as shown in (8.142), and

$$A_k = \frac{\left\{y^2+\sin^2\left[(k-2)\frac{\pi}{n}\right]\right\}\left\{y^2+\sin^2\left[(k-4)\frac{\pi}{n}\right]\right\}\cdots}{\left\{y^2+\sin^2\left[(k-1)\frac{\pi}{n}\right]\right\}\left\{y^2+\sin^2\left[(k-3)\frac{\pi}{n}\right]\right\}\cdots} \tag{8.160}$$

where the last term $y^2 + \sin^2(0)$ is replaced by y; e.g., $A_2 = \frac{y}{y^2+\sin^2(\pi/n)}$.

8.5.2 High Pass Filter Design

As mentioned earlier, the low pass filter prototypes serve as the basis from which the responses of all filters can be derived through proper frequency transformations. We have seen the transfer from a low pass filter prototype to an (actual) low pass filter via the frequency transformation from ω' (of low pass filter prototype) to ω/ω_c (of low pass filter) where $\omega_c' = 1$ rad/s is the low pass filter prototype's cutoff frequency.

The response of a high pass filter can be obtained from that of a low pass filter prototype by transforming ω' (of low pass filter prototype) to ω (of high pass filter) using the following form:

$$\frac{\omega'}{\omega_c'} = -\frac{\omega_c}{\omega} \tag{8.161}$$

where ω_c is the cutoff frequency of the high pass filter. Applying this frequency transformation to the low pass filter prototype, we can map its response onto another response as illustrated in Figure 8.54, which indeed represents a high pass filter. It is noted that ω' and ω, as defined in Figure 8.54, are the radian frequency in the low pass filter prototype and high pass filter domains, respectively.

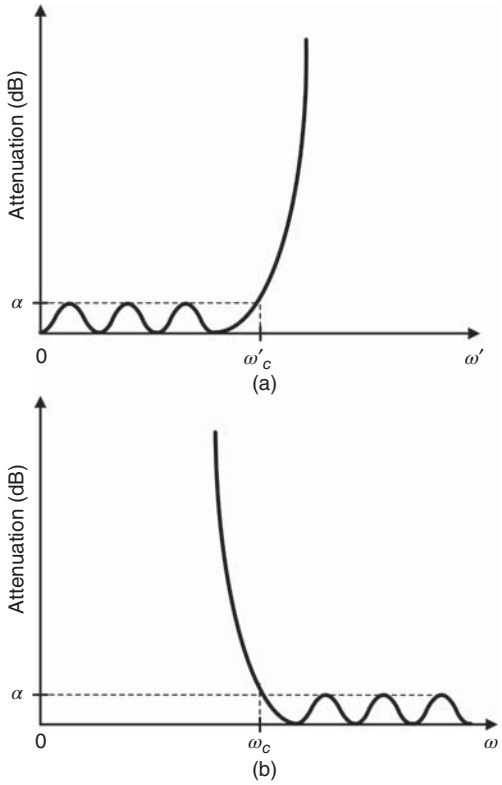


Figure 8.54. Frequency responses of the low pass filter prototype (a) and its corresponding high pass filter (b).

Applying the frequency mapping given in (8.161) to the inductors and capacitors making up the low pass filter prototypes, we obtain the following transformations:

Low pass filter prototype	High pass filter
$j\omega' L'_k$	$-j\frac{\omega_c\omega'_c}{\omega}L'_k = \frac{1}{j\omega(1/\omega_c\omega'_c L'_k)} = \frac{1}{j\omega C_k}$
$\frac{1}{j\omega' C'_k}$	$j\omega\left(\frac{1}{\omega_c\omega'_c C'_k}\right) = j\omega L_k$

It is apparent from these transformations that the inductors and capacitors in the low pass filter prototypes become the capacitors and inductors in high pass filters, respectively, through the frequency mapping (8.161) as

$$C_k = \frac{1}{\omega_c\omega'_c L'_k}$$

$$L_k = \frac{1}{\omega_c\omega'_c C'_k} \tag{8.162}$$

It is recalled that $\omega'_c = 1$ rad/s for the low pass filter prototypes. Figure 8.55 shows the schematic of a high pass filter transformed from the low pass filter prototype in Figure. 8.51(a) for n odd. Similarly, other high pass filters can be obtained from other low pass filter prototypes shown in Figure 8.51 through the

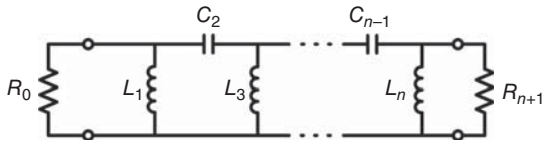


Figure 8.55. Schematic of a high pass filter.

frequency mapping (8.161) and the transformation equations (8.162). The impedance scaling as described earlier in (8.149) for the low pass filters is then applied to obtain the final element values for the high pass filters.

The design of a high pass filter can be proceeded from a low pass filter prototype as follows. First, the number of filter elements is determined by calculating the normalized low pass filter prototype frequency ω'/ω'_c given in the frequency transformation (8.161). This value is then used to determine the number of the low pass filter elements which is the same as that of the high pass filter elements as done for the low pass filter design. The attenuation characteristic of the high pass filter can be determined from (8.141) and (8.143) for the maximally flat and Chebyshev topology, respectively, upon applying the frequency mapping (8.161).

8.5.3 Band-Pass Filter Design

The response of a band-pass filter can be obtained from that of a low pass filter prototype by the transforming ω' (of low pass filter prototype) to ω (of band-pass filter) according to

$$\frac{\omega'}{\omega'_c} = \frac{1}{\Delta\omega}\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right) \tag{8.163}$$

where $\Delta\omega$ is the fractional bandwidth defined as

$$\Delta\omega = \frac{\omega_2 - \omega_1}{\omega_o} \tag{8.164}$$

and ω_o is the (design) center frequency defined as

$$\omega_o = \sqrt{\omega_1\omega_2} \tag{8.165}$$

with ω_1 and ω_2 being the lower and upper frequency of the pass band, respectively. Applying this frequency transformation to the low pass filter prototype, we can map its response onto another response as illustrated in Figure 8.56, which characterizes a band-pass filter. ω' and ω, as defined in Figure 8.56, are the radian frequency in the low pass filter prototype and band-pass filter domains, respectively. It is noted that additional pass-bands at higher frequencies (harmonics) occur if distributed-element resonators are used due to the resonators' inherent multiple resonances. It is noted that the frequency mapping (8.163) may result in a negative value. This, however, does not cause any issue in the filter design, and hence the negative sign is omitted in the design. The possible negative sign is due to the mathematical mapping between the low pass and band-pass filters. The band-pass frequency response below ω_o seen in Figure 8.56(b) corresponds to the mirror of the low pass filter response in the negative frequency region in Figure 8.56(a). Since the

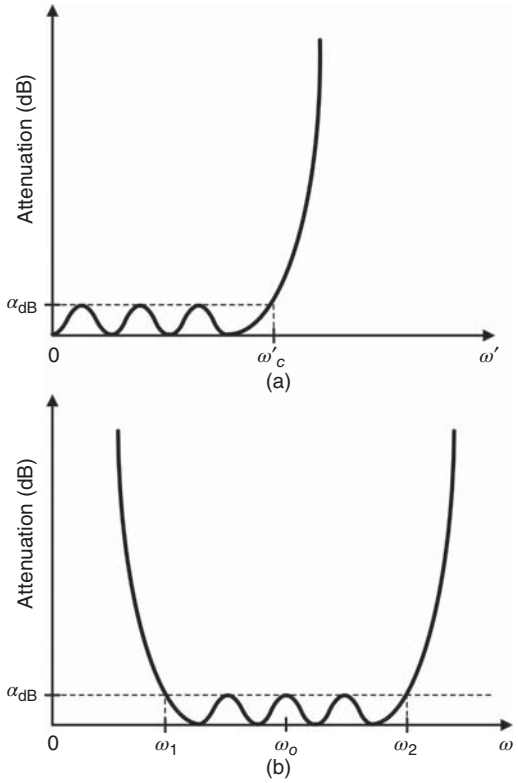


Figure 8.56. Frequency responses of the low pass filter prototype (a) and its corresponding band-pass filter (b).

responses of the low pass filter in the positive frequency region and its mirror image in the negative frequency region are identical with frequency shifting, the resultant band-pass filter response through the mapping is correct.

Applying the frequency mapping given in (8.163) to the series inductors of the low pass filter prototypes, we can derive

$$j\omega' L'_k = j\frac{\omega'_c}{\Delta\omega}\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right) L'_k$$

$$= j\omega\left(\frac{\omega'_c L'_k}{\omega_o \Delta\omega}\right) + \frac{1}{j\omega\left(\frac{\Delta\omega}{\omega'_c \omega_o L'_k}\right)}$$

$$= j\omega L_k + \frac{1}{j\omega C_k} \tag{8.165}$$

which shows that the series inductor is transferred into a series resonator consisting of an inductor and a capacitor having

$$L_k = \frac{\omega'_c L'_k}{\omega_o \Delta\omega}$$

$$C_k = \frac{\Delta\omega}{\omega'_c \omega_o L'_k} \tag{8.166}$$

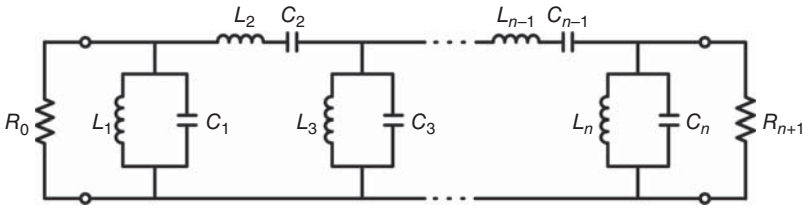


Figure 8.57. Schematic of a band-pass filter.

Similarly, we obtain for the shunt capacitors of the low pass filter prototypes, after applying the frequency mapping:

$$j\omega' C'_k = j\frac{\omega'_c}{\Delta\omega}\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right) C'_k$$

$$= j\omega\left(\frac{\omega'_c C'_k}{\omega_o \Delta\omega}\right) + \frac{1}{j\omega\left(\frac{\Delta\omega}{\omega'_c \omega_o C'_k}\right)}$$

$$= j\omega C_k + \frac{1}{j\omega L_k} \tag{8.167}$$

which shows that the shunt capacitor is transferred into a shunt resonator consisting of an inductor and a capacitor having

$$L_k = \frac{\Delta\omega}{\omega'_c \omega_o C'_k}$$

$$C_k = \frac{\omega'_c C'_k}{\omega_o \Delta\omega} \tag{8.168}$$

Again, $\omega'_c = 1$ rad/s for the low pass filter prototypes. Figure 8.57 shows the schematic of a band-pass filter transformed from the low pass filter prototype in Figure 8.51(a) for n odd. Similarly, other band-pass filters can be obtained from other low pass filter prototypes shown in Figure 8.51 through the frequency mapping (8.163) and the transformation equations (8.166) and (8.168). The impedance scaling described earlier for the low pass filters is then applied to obtain the final element values for the band-pass filters.

The design of a band-pass filter can be proceeded from a low pass filter prototype as follows. First, the number of filter elements is determined by calculating the normalized low pass filter prototype frequency ω'/ω'_c given in the frequency transformation (8.163). This value is then used to determine the number of the low pass filter elements which is the same as that of the band-pass filter elements as done for the low pass filter design. The attenuation characteristic of the band-pass filter can be determined from (8.141) and (8.143) for the maximally flat and Chebyshev topology, respectively, upon using the frequency mapping (8.163).

8.5.4 Band-Stop Filter Design

The response of a band-stop filter can be obtained from that of a low pass filter prototype by transforming ω' (of low pass filter prototype) to ω (of band-stop filter) using the following form:

$$\frac{\omega'_c}{\omega'} = \frac{1}{\Delta\omega}\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right) \tag{8.169}$$

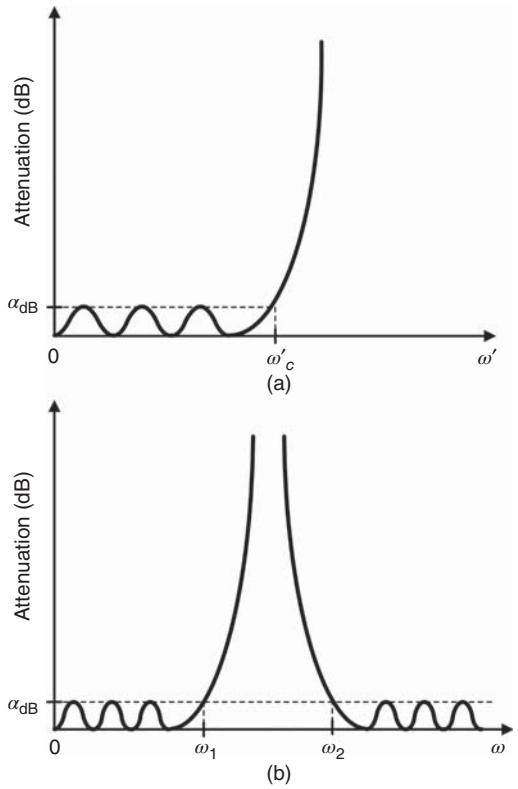


Figure 8.58. Frequency responses of the low pass filter prototype (a) and its corresponding band-stop filter (b).

where $\Delta\omega$ is the fractional bandwidth defined as

$$\Delta\omega = \frac{\omega_2 - \omega_1}{\omega_o} \tag{8.170}$$

and ω_o is the center frequency defined by

$$\omega_o = \sqrt{\omega_1 \omega_2} \tag{8.171}$$

with ω_1 and ω_2 being the lower and upper frequency of the stop band at the maximum insertion loss or pass-band ripple (α_{dB}) as seen in Figure 8.58, respectively. Applying this frequency transformation to the low pass filter prototype, we can map its response onto the response of a band-stop filter as illustrated in Figure 8.58. ω' and ω, as defined in Figure 8.58, are the radian frequency in the low pass filter prototype and band-stop filter domains, respectively.

Applying the frequency mapping (8.169) to the series inductors of the low pass filter prototypes, we obtain

$$\begin{aligned}\frac{1}{j\omega' L'_k} &= \frac{1}{j\omega'_c \Delta\omega}\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right)\frac{1}{L'_k} \\ &= j\frac{\omega_o}{\omega\omega'_c \Delta\omega L'_k} + \frac{\omega}{j\omega_o \omega'_c \Delta\omega L'_k} \\ &= j\omega C_k + \frac{1}{j\omega L_k}\end{aligned} \tag{8.172}$$

which leads to

$$\omega L_k = \frac{\omega_o \omega'_c \Delta\omega L'_k}{\omega}$$

$$\omega C_k = \frac{\omega_o}{\omega \omega'_c \Delta\omega L'_k} \tag{8.173}$$

from which, we obtain at the center frequency ω_o:

$$L_k = \frac{\omega'_c \Delta\omega L'_k}{\omega_o}$$

$$C_k = \frac{1}{\omega'_c \omega_o \Delta\omega L'_k} \tag{8.174}$$

It is now apparent that each series inductor is transferred into a parallel resonator consisting of an inductor and a capacitor whose respective inductance and capacitance are given in (8.174).

Similarly, we can obtain for the shunt capacitors of the low pass filter prototypes, after applying the frequency mapping (8.169):

$$\frac{1}{j\omega' C'_k} = \frac{1}{j\omega'_c \Delta\omega} \left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega} \right) \frac{1}{C'_k}$$

$$= j \frac{\omega_o}{\omega \omega'_c \Delta\omega C'_k} + \frac{\omega}{j\omega_o \omega'_c \Delta\omega C'_k}$$

$$= j\omega L_k + \frac{1}{j\omega C_k} \tag{8.175}$$

from which, we get

$$\omega L_k = \frac{\omega_o}{\omega \omega'_c \Delta\omega C'_k}$$

$$\omega C_k = \frac{\omega_o \omega'_c \Delta\omega C'_k}{\omega} \tag{8.176}$$

where $\omega'_c = 1$ rad/s. A shunt capacitor is thus transferred into a series resonator consisting of an inductor and a capacitor having

$$L_k = \frac{1}{\omega_o \omega'_c \Delta\omega C'_k}$$

$$C_k = \frac{\omega'_c \Delta\omega C'_k}{\omega_o} \tag{8.177}$$

Figure 8.59 shows the schematic of a band-stop filter transformed from the low pass filter prototype in Figure 8.51(a) for n odd. Similarly, other band-stop filters can be obtained from other low pass filter prototypes shown in Figure 8.51 through the frequency mapping (8.169) and the transformation equations (8.174) and (8.177). The impedance scaling described earlier for the low pass filter is finally applied to obtain the final element values for the band-stop filter.

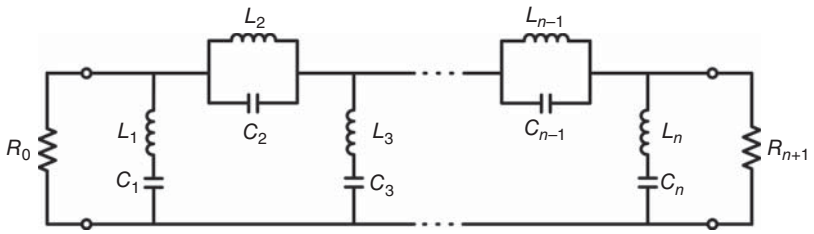


Figure 8.59. Schematic of a band-stop filter.

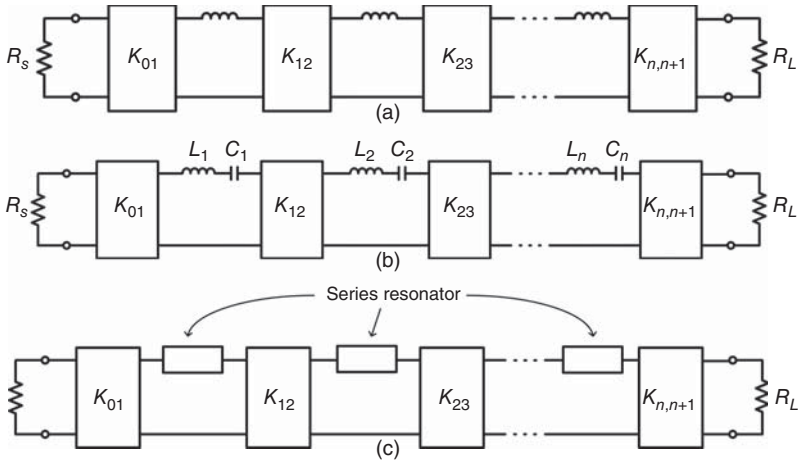


Figure 8.60. (a) Modified low pass filter prototype using impedance inverters. (b) Band-pass filter consisting of only series resonators of inductors and capacitors and impedance inverters. (c) Generalized band-pass filter using impedance inverters. $K_{ok}, k = 1, 2, \ldots, n+1$, are the impedance inverter parameters. R_s and R_L are the source and load impedance, respectively.

The design of a band-stop filter can be proceeded from a low pass filter prototype as follows. First, the number of filter elements is determined by calculating the normalized low pass filter prototype frequency ω'/ω'_c given in the frequency transformation (8.169). This value is then used to determine the number of the low pass filter elements, which is the same as that of the band-stop filter elements as done for the low pass filter design. The attenuation characteristic of the band-stop filter can be determined from (8.141) and (8.143) for the maximally flat and Chebyshev topology, respectively, upon using the frequency mapping (8.169).

8.5.5 Filter Design Using Impedance and Admittance Inverters

As discussed in Section 5.5, a series resonator can be transformed into a parallel resonator or vice versa by using an impedance or admittance inverter. Impedance and admittance inverters can therefore facilitate the design of band-pass and band-stop filters, allowing only one kind of resonators (either series or parallel resonator), instead of both types of resonators, to be used. We will illustrate this design approach using band-pass filters. Similar formulation can be employed for band-stop filter design using impedance and admittance inverters.

As with a typical filter design, we start with a low pass filter prototype such as that in Figure 8.51(a) for n odd. This low pass filter prototype can be modified to include the impedance inverters as shown in Figure 8.60. Applying the low pass to band-pass frequency transformation given in (8.163), we replace the series inductors by series resonators to obtain the band-pass filter in Figure 8.60(b). If distributed elements such as transmission lines are used for the series resonators, then the band-pass filter can generally be redrawn as in Figure 8.60(c). In addition to the impedance inverters, admittance inverters can also be used to realize band-pass filters as shown in Figure 8.61. It is noted that some band-pass filters (or band-stop filters) lend themselves well to the impedance or admittance inverter, and so either the impedance or admittance inverter

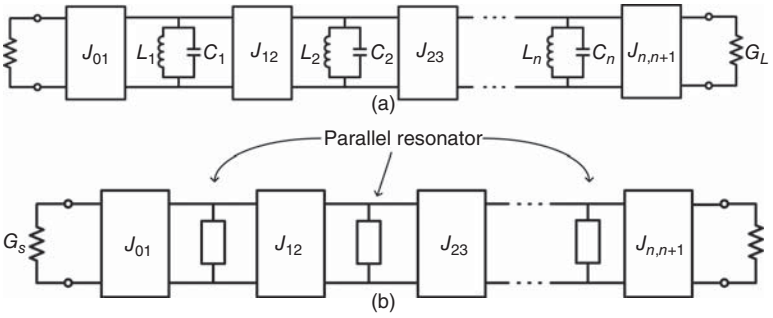


Figure 8.61. (a) Band-pass filter consisting of only shunt resonators of inductors and capacitors and admittance inverters. (b) Generalized band-pass filter using admittance inverters. J_{ok}, $k = 1, 2, \ldots, n+1$, are the admittance inverter parameters. $G_S = 1/R_s$ and $G_L = 1/R_L$ are the source and load conductance, respectively.

can be used depending on the band-pass filter (or band-stop filter) topology. For instance, the end-coupled band-pass filter to be discussed in the next section is well suited for the admittance inverter.

As examples for the band-pass filter design using distributed resonators and inverters, we consider the design of end- and parallel-coupled band-pass filters in the following sections.

8.5.5.1 End-Coupled Band-Pass Filter Design. Figure 8.62 shows the end-coupled band-pass filters, which consist of a sequence of coupling elements alternately located along transmission lines functioning as resonators. Figure 8.62(a) shows a filter implemented on the same metal layer with gaps representing the coupling elements. Figure 8.62(b) and (c) illustrate filters using different metal layers (such as the top-most and lower layers in a CMOS structure) whose coupling elements are formed using fully or partly overlapped sections (broadside coupling), with Figure 8.62(c) showing an example of using orthogonal transmission lines to reduce the overall filter size and the interaction between adjacent resonators. The end-coupled band-pass filter has the first spurious response at approximately twice the center frequency. It can be designed using the low pass filter prototype. The gap- or broadside-coupling section are represented by an equivalent circuit as shown in Figure 8.63(a) based on the discussion in Section 5.3.2.3 of Chapter 5. This equivalent circuit is equivalent to an admittance inverter as shown in Figure 8.63(b) as discussed in Section 7.5.1.2. The coupling element essentially acts as an admittance inverter.

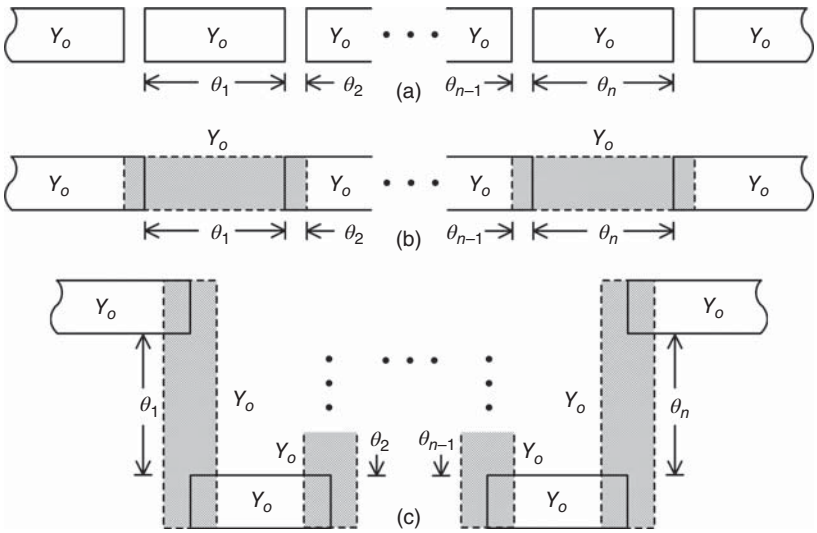


Figure 8.62. End-coupled band-pass filters with transmission-line resonators on the same metal layer with gap-coupling (a) and different metal layers with broadside-coupling (b, c). Dashed sections are on different metal layer(s).

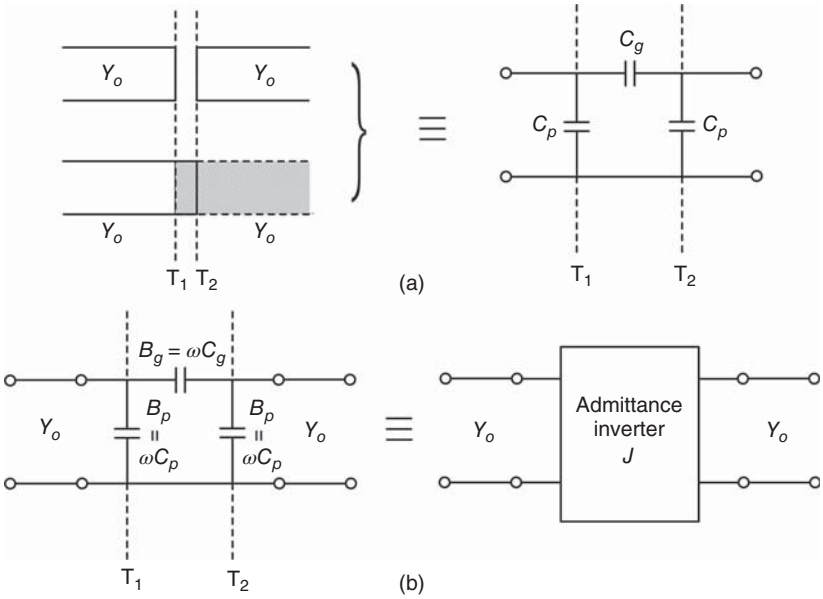


Figure 8.63. Coupling section and its equivalent circuit (a) and corresponding admittance inverter (b).

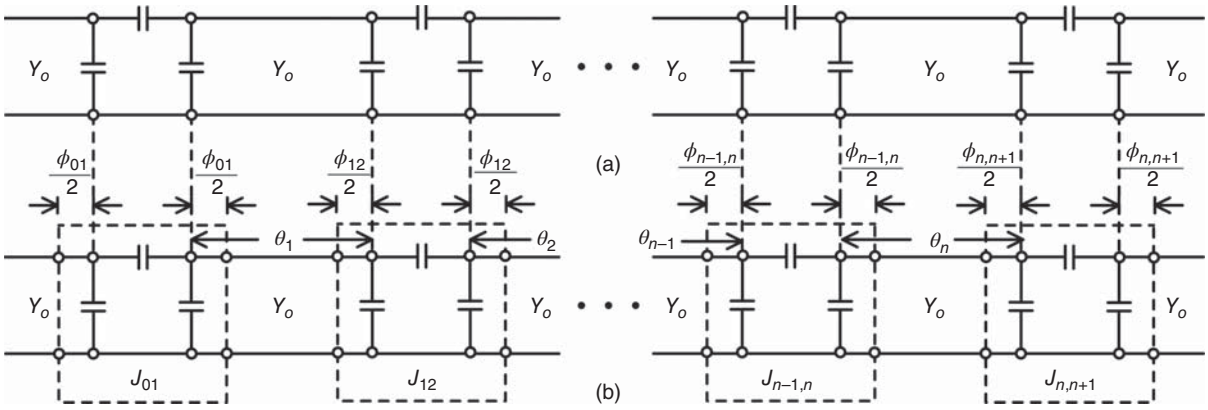


Figure 8.64. Equivalent circuits of the end-coupled band-pass filter using equivalent-circuit of the coupling elements (a) and admittance inverters (b).

We begin the design formulation by drawing an equivalent circuit of the end-coupled band-pass filter as shown in Figure 8.64(a). To draw the equivalence between the band-pass filter and the low pass filter prototype, it is necessary to redraw the band-pass filter as a cascade of admittance inverters, representing the coupling segments, separated by transmission lines of electrical length $\theta_j (j = 1, 2, \ldots, n)$, functioning as series resonators, in Figure 8.64(b). The electrical length θ_j of the transmission line between coupling elements j and $j + 1$ is thus given as

$$\theta_j = \theta_o + \frac{1}{2}(\phi_{j-1,j} + \phi_{j,j+1}) \tag{8.178}$$

where $\theta_o = \pi$ radians at the design center frequency. The electrical length ϕ and parameter J of the admittance inverter constituted by the coupling element's equivalent circuit are given by (5.211) and (5.212) in Chapter 5 and are repeated here:

$$\frac{J}{Y_o} = \left| \tan\left(\frac{\phi}{2} + \tan^{-1}\frac{B_p}{Y_o}\right)\right| \quad \text{mho} \tag{8.179}$$

$$\phi = -\tan^{-1}\left(\frac{2B_g}{Y_o} + \frac{B_p}{Y_o}\right) - \tan^{-1}\frac{B_p}{Y_o} \text{ rad} \tag{8.180}$$

Equations (8.189) and (8.180) allow the dimensions of a coupling element, such as the gap size of a coupling gap, represented by the normalized admittance inverter parameter J/Y_o to be determined for a given value of J/Y_o. This process typically involves iteratively numerical calculation and optimization.

The required admittance inverter parameters, which are needed to determine the coupling-element dimensions, can be derived, assuming a narrow bandwidth around the center frequency where the electrical length is approximately equal to π radians, as

$$\frac{J_{o1}}{Y_o} = \sqrt{\frac{\pi\Delta\omega}{2g_o g_1 \omega_c'}} \tag{8.181}$$

$$\left.\frac{J_{j,j+1}}{Y_o}\right|_{j=1 \text{ to } n-1} = \frac{\pi\Delta\omega}{2\omega_c'}\sqrt{\frac{1}{g_j g_{j+1}}} \tag{8.182}$$

$$\frac{J_{n,n+1}}{Y_o} = \sqrt{\frac{\pi\Delta\omega}{2g_n g_{n+1}\omega_c'}} \tag{8.183}$$

where $g_o, g_1, \ldots, g_{n+1}$ are the low pass filter prototype values, ω_c' is the low pass filter prototype's cutoff frequency, Y_o is the characteristic admittance of the transmission line, and

$$\Delta\omega = 2\left(\frac{\omega_2 - \omega_1}{\omega_2 + \omega_1}\right) \tag{8.184}$$

is the fractional bandwidth with ω_1 and ω_2 being the lower and upper frequencies of the pass band, respectively. The mapping from the low pass filter prototype to the corresponding band-pass filter is derived approximately as

$$\frac{\omega'}{\omega_c'} = \frac{2}{\Delta\omega}\left(\frac{\omega - \omega_o}{\omega}\right) \tag{8.185}$$

where ω' and ω, as defined in Figure 8.56, are the radian frequency in the low pass filter prototype and band-pass filter domains, respectively, and

$$\omega_o = \frac{2\omega_1\omega_2}{\omega_1 + \omega_2} \tag{8.186}$$

is the design center frequency. Note that the frequency mapping (8.163), while exact for the lumped-element band-pass filters transformed from the low pass filter prototypes such as that shown in Figure 8.57, may not give accurate mapping for other band-pass filter structures such as the end-coupled band-pass filter. For the end-coupled band-pass filter, the frequency mapping given in (8.185) is more accurate. When the coupling gap is very small, or the overlapped coupling section is not offset and its in-between dielectric layer is very thin, the coupling is very tight and the coupling element can be simply represented as a series capacitor whose susceptance, $B_{j,j+1}$, is given by

$$\frac{B_{j,j+1}}{Y_o} = \frac{J_{j,j+1}/Y_o}{1 - (J_{j,j+1}/Y_o)^2} \tag{8.187}$$

and the electrical length of the resonator j becomes

$$\theta_j = \pi - \frac{1}{2}\left[\tan^{-1}\left(\frac{2B_{j-1,j}}{Y_o}\right) + \tan^{-1}\left(\frac{2B_{j,j+1}}{Y_o}\right)\right] \text{ rad} \tag{8.188}$$

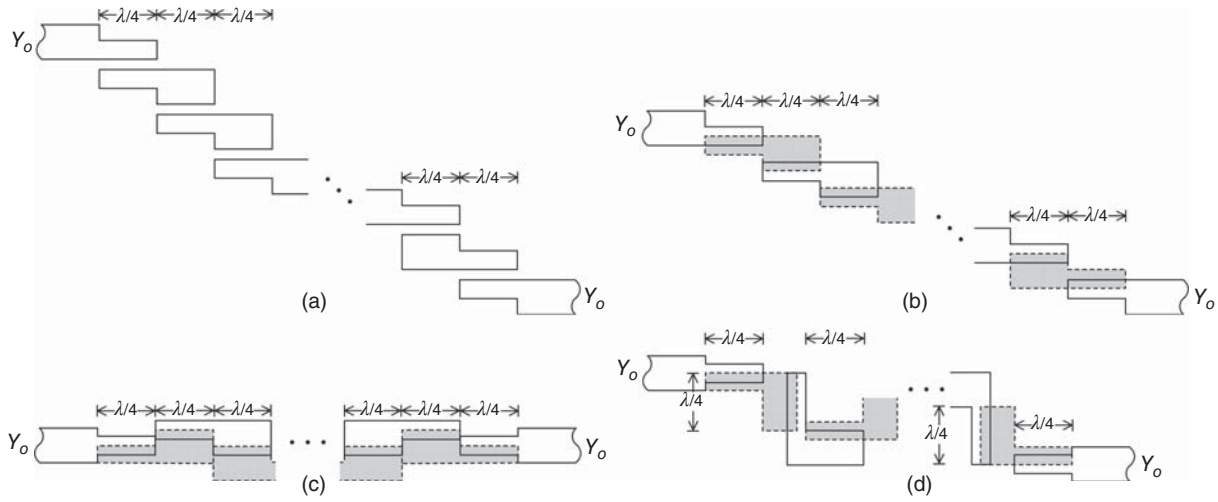


Figure 8.65. Parallel-coupled band-pass filters with resonators on the same metal layer (edge coupling) (a) and different metal layers (broadside coupling) (b, c, d). Dashed sections are on different metal layer(s).

A design procedure for the end-coupled band-pass filter can be summarized as follows. First, determine the number of filter elements using the low pass filter prototype and low pass to band-pass mapping described in (8.185). Second, calculate the normalized admittance inverter parameters from the filter specifications using (8.181) to (8.183). Third, determine the dimensions of the coupling elements so that their admittance inverter parameters given in (8.179) are equal to the admittance inverter parameters calculated from (8.181) to (8.183). Finally calculate the electrical lengths of the transmission-line resonators using (8.188) at the design center frequency ω_o.

To obtain a wide bandwidth, adjacent resonators must be tightly coupled, requiring large series coupling capacitances at the coupling segments, and hence small gaps or fully overlapped coupling sections with a thin dielectric in between, especially the first and last coupling elements. Using gaps as the coupling elements, this filter type is applicable only for a narrow and moderate bandwidth of up to about 15%. For a greater bandwidth, strong coupling mechanisms such as that based on overlapping segments (broadside coupling) need to be used.

8.5.5.2 Parallel-Coupled Band-Pass Filter Design. Figure 8.65 shows the parallel-coupled band-pass filters. Figure 8.65(a) shows a filter implemented on the same metal layer with coupling occurring along the edges of the resonators. Figure 8.65(b) and (c) illustrate filters using different metal layers (such as the top-most and lower layers in a CMOS structure) where the broadside coupling over the entire or partial width is utilized [13]. Figure 8.65(d) exemplifies the use of orthogonal transmission lines to reduce the overall filter size and the interaction between adjacent resonators. The filter consists of a series of half-wavelength transmission-line resonators with adjacent resonators parallel each other along half of their length. In this arrangement, the resonators are coupled along the edge (or over the full or partial width) along a quarter-wavelength. As a result, tight coupling can be achieved without using narrow gaps or, when broadside coupling is used as in Figure 8.65(b), much stronger coupling is obtained, as compared to the foregoing end-coupled band-pass filter. The coupling, however, is no longer purely capacitive, as in the case of the end-coupled band-pas filter, since the overlapping lengths are one quarter-wavelength long at the center frequency and the phase varies along these lengths. The resultant coupling between resonators is partly electric and partly magnetic, and hence is more complicated to model. The parallel-coupled band-pass filter offers several advantages as compared to the end-coupled band-pass filters. First, the filter length is reduced approximately by half. Second, a symmetric attenuation response versus frequency is obtained with the first spurious response occurring at three times the center frequency. Third, more relaxed dimensions (e.g., larger gaps) between adjacent resonators can be obtained.

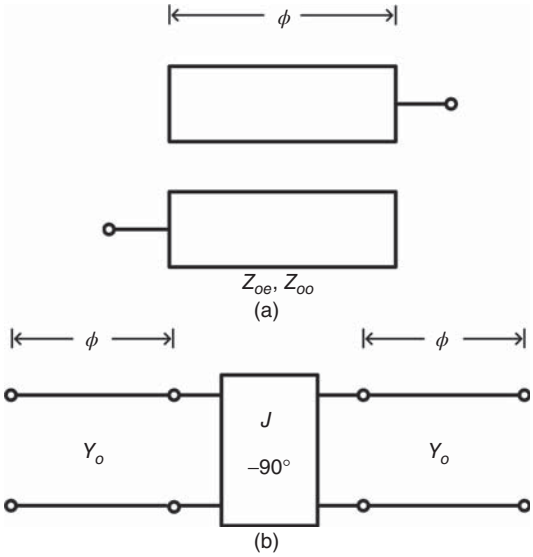


Figure 8.66. Parallel-coupled transmission line (a) and its equivalent admittance-inverter network (b).

Toward the objective of representing the parallel-coupled band-pass filter with admittance inverters, we show in Figure 8.66 a parallel-coupled transmission line in a homogeneous medium, characterized by its electrical length ϕ and even- and odd-mode characteristic impedances Z_{oe} and Z_{oo}, and its equivalent network with J representing an ideal admittance inverter having a constant image admittance J and a phase shift of −90° at all frequencies. The $ABCD$ matrix of the transmission line having length ϕ and characteristic admittance Y_o, and that of an ideal admittance inverter J, seen in Figure 8.66, can be derived as

$$\begin{bmatrix} A_{\mathrm{TL}} & B_{\mathrm{TL}} \\ C_{\mathrm{TL}} & D_{\mathrm{TL}} \end{bmatrix} = \begin{bmatrix} \cos\phi & \dfrac{j}{Y_o}\sin\phi \\ jY_o\sin\phi & \cos\phi \end{bmatrix} \tag{8.189}$$

$$\begin{bmatrix} A_J & B_J \\ C_J & D_J \end{bmatrix} = \begin{bmatrix} 0 & -\dfrac{j}{J} \\ -jJ & 0 \end{bmatrix} \tag{8.190}$$

respectively.

Utilizing the equivalent admittance-inverter network for the coupled lines as shown in Figure 8.66(b), we can draw in Figure 8.67 an equivalent circuit of the parallel-coupled band-pass filter similar to that of the end-coupled band-pass filter discussed earlier. The admittance inverter parameters can be obtained from

$$\frac{J_{o1}}{Y_o} = \sqrt{\frac{\pi\Delta\omega}{2g_o g_1}} \tag{8.191}$$

$$\left.\frac{J_{j,j+1}}{Y_o}\right|_{j=1 \text{ to } n-1} = \frac{\pi\Delta\omega}{2\omega_c'}\sqrt{\frac{1}{g_j g_{j+1}}} \tag{8.192}$$

$$\frac{J_{n,n+1}}{Y_o} = \sqrt{\frac{\pi\Delta\omega}{2g_n g_{n+1}}} \tag{8.193}$$

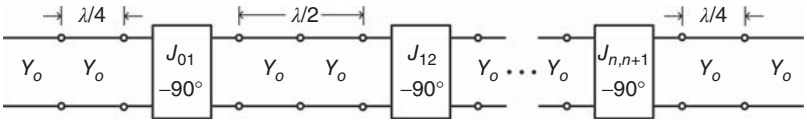


Figure 8.67. Equivalent circuit of the parallel-coupled band-pass filter.

where $g_o, g_1, \ldots, g_{n+1}$ are the low pass filter prototype values, ω_c' is the low pass filter prototype's cutoff frequency, Y_o is the characteristic admittance of the terminating transmission line, and the fractional bandwidth is given in (8.184). The even- and odd-mode characteristic impedances of the coupled lines can be derived as

$$Z_{oe}^{j,j+1}|_{j=0 \text{ to } n} = \frac{1}{Y_o}\left[1 + \frac{J_{j,j+1}}{Y_o} + \left(\frac{J_{j,j+1}}{Y_o}\right)^2\right] \tag{8.194}$$

$$Z_{oo}^{j,j+1}|_{j=0 \text{ to } n} = \frac{1}{Y_o}\left[1 - \frac{J_{j,j+1}}{Y_o} + \left(\frac{J_{j,j+1}}{Y_o}\right)^2\right] \tag{8.195}$$

As noted early, the coupled lines are assumed to be embedded in a homogeneous medium where the even- and odd-mode phase velocities in the coupled lines are equal. In this case, the length of the coupled line is a quarter-wavelength at the design center frequency f_o (i.e., $\ell = \lambda/4 = c/4f_o\sqrt{\varepsilon_r}$) where ε_r is the relative dielectric constant. In inhomogeneous transmission lines such as microstrip lines, the even and odd modes have different phase velocities, leading to an asymmetric filter response. Additionally, each of the two open ends of a coupled-line section has fringing capacitance, hence resulting in an extra length at each end. To compensate for the difference in the mode velocities and the open-end effects, the actual physical length of the coupled-line section can be determined as [14].

$$\ell = \frac{\frac{\pi}{2} \pm \delta_e - 2\Delta\theta_e}{2\pi/\lambda_e} = \frac{\frac{\pi}{2} \mp \delta_o - 2\Delta\theta_o}{2\pi/\lambda_o} \tag{8.196}$$

where λ_e and λ_o are the even- and odd-mode wavelengths at the design center frequency, respectively,

$$\delta_i = Z_{oi}\frac{\frac{\pi}{2}(\sqrt{\varepsilon_{\text{reff}}^e} - \sqrt{\varepsilon_{\text{reff}}^o}) + 2(\Delta\theta_e\sqrt{\varepsilon_{\text{reff}}^o} - \Delta\theta_o\sqrt{\varepsilon_{\text{reff}}^e})}{\pm Z_{oe}\sqrt{\varepsilon_{\text{reff}}^e} \pm Z_{oo}\sqrt{\varepsilon_{\text{reff}}^o}} \tag{8.197}$$

and

$$\Delta\theta_i = \omega W C_{fi} Z_{oi} \tag{8.198}$$

is the extra electrical length at each open end. The subscript i stands for e or o denoting the even and odd modes, respectively; $Z_{oe}(i = e)$ and $Z_{oo}(i = o)$ are the even- and odd-mode characteristic impedance of the coupled line; $\varepsilon_{\text{reff}}^e$ and $\varepsilon_{\text{reff}}^o$ are the even and odd-mode effective dielectric constants; ω is the radian frequency; W is the strip width; C_{fe} and C_{fo} are the even and odd-mode fringing capacitances per unit width at the open end. The + and − signs in front of δ_e and δ_o correspond to a couple-lime medium in which the even-mode effective dielectric constant is greater than that of the odd mode (e.g., microstrip line), and the other sign pair is used for a medium having smaller even-mode effective dielectric constant (e.g., suspended stripline).

The mapping from the low pass filter prototype to the corresponding band-pass filter is approximately given by

$$\frac{\omega'}{\omega'_c} = \frac{2}{\Delta\omega}\left(\frac{\omega - \omega_o}{\omega}\right) \tag{8.199}$$

where

$$\omega_o = \frac{\omega_1 + \omega_2}{2} \tag{8.200}$$

is the design center frequency with ω_1 and ω_2 being the passband lower and upper frequencies, respectively. As for the end-coupled band-pass filter, the frequency mapping given in (8.185) is more accurate than that in (8.163) used for the lumped-element band-pass filters.

A design procedure for the parallel-coupled band-pass filter can be summarized as follows. First, determine the number of filter elements using the low pass filter prototype and low pass to band-pass mapping described in (8.199). Second, calculate the normalized admittance inverter parameters from the filter specifications using (8.191) to (8.193). Third, calculate the even- and odd-mode characteristic impedances from (8.194) and (8.195), from which the dimensions of the coupled lines are determined. Finally, calculate the lengths of the coupled-line sections using (8.196)–(8.198) at the design center frequency ω_o.

For broadband parallel-coupled band-pass filters, the coupling at the (source and load) end sections is very tight. This may result in gaps that are physically unrealizable when parallel-coupled microstrip lines on the same metal layer are used. Under this situation, broadside coupling using different metal layers can avoid the problems and hence becomes useful.

REFERENCES

1. J. Reed and G. J. Wheeler, "A Method of Analysis of Symmetrical Four-Port Networks," *IRE Trans. Microwave Theory Tech.*, Vol. 8, pp. 116–118, Jan. 1960.
2. J. Lange, "Interdigitated Stripline Quadrature Hybrid," *IEEE Trans. Microwave Theory Tech.*, Vol. 17, pp. 1150–1151, Dec. 1969.
3. R. Waugh and D. LaCombe, "Unfolding the Lange Coupler," *IEEE Trans. Microwave Theory Tech.*, Vol. 20, pp. 777–779, Nov. 1972.
4. D. Pozar, Microwave Engineering, Third Edition, John Wiley & Sons, Inc., New York, 2005.
5. M. K. Chirala and C. Nguyen, "Multilayer Design Techniques for Extremely Miniaturized CMOS Microwave and Millimeter-Wave Distributed Passive Circuits," *IEEE Trans. Microwave Theory Tech.*, Vol. 54, pp. 4218–4224, Dec. 2006.
6. TSMC 0.25-μm CMOS Process, MOSIS Foundry, Marina del Rey, California, USA.
7. T. Wang and K. Wu, "Size-reduction and Band-broadening Design Technique of Uniplanar Hybrid Ring Coupler using Phase Inverter for M (H) MIC's," *IEEE Trans. Microwave Theory Tech.*, Vol. 27, pp. 198–206, Feb. 1999.
8. C.-Y. Chang, C.-C. Yang, "A Novel Broad-band Chebyshev-response Rat-race Ring Coupler," *IEEE Trans. Microwave Theory Tech.*, Vol. 47, pp. 455–462, Apr. 1999.
9. E. Wilkinson, "An N-Way Hybrid Power Divider," *IRE Trans. Microwave Theory Tech.*, Vol. 8, pp. 116–118, Jan. 1960.
10. G. Matthaei, L. Young, and E.M.T. Jones, Microwave Filters, Impedance-Matching Networks, and Coupling Structures, Artech House, Dedham, Mass., 1980.
11. R. Levy, "Tables of Element Values for the Distributed Low-Pass Prototype Filter," *IEEE Trans. Microwave Theory Tech.*, Vol. 13, pp. 514–536, Sept. 1965.
12. J. D. Rhodes, Theory of Electrical Filters, John Wiley & Sons, Inc., New York, New York, 1976.
13. M. Tran and C. Nguyen, "Modified Broadside-Coupled Microstrip Lines Suitable for MIC and MMIC Applications and a New Class of Broadside-Coupled Band-Pass Filters," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-41, No. 8, pp. 1336–1342, Aug. 1993.
14. C. Nguyen, "Accurate Equation for Determining Resonator Length in Half-Wavelength Parallel-Coupled Band-Pass Filter," *Electro. Lett.*, Vol. 29, No. 6, pp. 532–533, Mar. 1993.

PROBLEMS

8.1 Prove that Eq. (8.7) cannot be satisfied. This demonstrates that there exists no matched, lossless, reciprocal three-port component.

8.2 Prove that a matched three-port network must be either lossy or nonreciprocal.

8.3 Derive Eqs. (8.11) and (8.12) to prove that a matched, lossless, nonreciprocal three-port network is a circulator.

8.4 Derive the *S*-parameters for a lossless, reciprocal three-port component with perfect match at input port 1. Calculate the return loss at output ports 2 and 3, and the insertion loss between ports. Describe the operation of this component.

8.5 Applying the conditions in Eqs. (8.22)–(8.24) to derive Eq. (8.25).

8.6 Derive (8.29) and (8.30).

8.7 Consider a directional coupler using two parallel edge-coupled transmission lines in a 50-Ω system. Determine the even- and odd-mode characteristic impedances for 3-dB coupling. What do you think of possible design problem(s) in realizing such a tight coupling on parallel edge-coupled transmission lines such as microstrip lines. Demonstrate this problem using microstrip lines on any available CMOS structure.

8.8 Derive Eqs. (8.82), (8.83), and (8.84).

8.9 Derive Eq. (8.92).

8.10 The [*S*] matrix of a dual directional coupler measured at 10 GHz using a vector network analyzer is

$$\begin{bmatrix} 0.1\angle 5^o & 0.08\angle -40^o & 0.313\angle 100^o & 0.95\angle -70^o \\ 0.08\angle -40^o & 0.07\angle 17^o & 0.94\angle -60^o & 0.303\angle 110^o \\ 0.313\angle 100^o & 0.94\angle -60^o & 0.11\angle -4^o & 0.07\angle -55^o \\ 0.95\angle -70^o & 0.303\angle 110^o & 0.06\angle 10^o & 0.09\angle -25^o \end{bmatrix}$$

a) Calculate the return losses at all four ports.
b) Calculate the coupling in dB, assuming the input signal is at port 1.
c) Calculate the coupling in dB, assuming the input signal is at port 3.
d) Calculate the directivity when the input signal is at port 1.
e) Calculate the directivity when the input signal is at port 4.
f) Calculate the isolation when the input signal is at port 1.
What is your observation of the calculated Isolation with respect to the calculated Directivity and Coupling? Is this expected?

8.11 One application of directional couplers is sampling power of RF circuits. Consider the measurement setup shown in Figure P8.1, in which a 10-dB directional coupler is used to measure the output power of an RF source.

a) Assume the directional coupler and power sensor (detector) have no loss at the measurement frequency. If the power meter reads 8 dBm, what is the power entering the directional coupler at point A?
b) Assume the directional coupler has 0.4-dB insertion loss and the power sensor has an insertion loss of 0.5 dB at the measurement frequency, what is the power level leaving the directional coupler at point B, assuming 8 dBm is read on the power meter?

8.12 Consider an RF measurement setup as shown in Figure P8.2. The directional coupler is assumed to be lossless and having infinite directivity. The antenna has a VSWR of 2. The power incident to port 1 of the 20-dB directional coupler is 5 W. Calculate the output powers at ports 3 and 4.

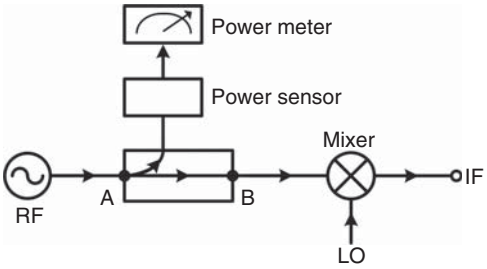


Figure P8.1.

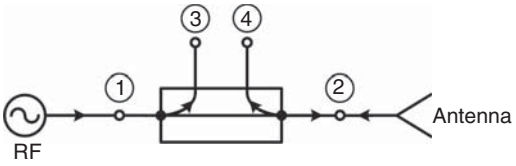


Figure P8.2.

8.13 Using any available CMOS structure or the CMOS profile shown in Figure P15.2, design a 34–36 GHz 10-dB coupler using parallel-coupled microstrip lines whose strips lie on the top-most metal layer (M6 in Figure P15.2) and ground plane on the bottom-most metal layer (M1 in Figure P15.2). The terminating impedance is 50 Ω. Perform necessary optimization to achieve the required 10-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 90° as possible, directivity as high as possible, and return loss at the input, through and coupled ports as high as possible across 34–36 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using both non-EM (i.e., circuit) and EM simulators from 30 to 40 GHz. Compare and comment on the results between non-EM and EM simulations. Now using the EM-simulated results on return losses coupling, determine an appropriate operating bandwidth (lower and upper frequencies) over which the coupler has a reasonably flat coupling around 10 dB (e.g., 10.5 ± 0.4 dB) with decent return losses. The determined frequency range may be different from 34–36 GHz.

8.14 Repeat Problem 8.13 using parallel-coupled strip lines. If the CMOS profile in Figure P15.2 is employed, then use the metal layer 5 (M5) for the two strips and the metal layers 1 (M1) and 6 (M6) for the two ground planes. It is noted that the Oxide layer thickness between M5–M6 and M5–M1 is different. If you also design the coupler in Problem 8.13, then compare the performance between that coupler and the coupler designed in this problem. Discuss the performance difference, particularly the directivity, isolation, and amplitude and phase imbalance. As the strip lines can also be conveniently realized in CMOS structures, they provide an alternative solution to RFIC design in addition to the commonly used microstrip lines and CPW.

8.15 Using any available CMOS structure or the CMOS profile shown in Figure P15.2, design a 50-Ω based 32–38 GHz 3-dB coupler using broadside-coupled transmission lines whose strips lie on two different metal layers and ground plane on the bottom-most metal layer. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 90° as possible, directivity as high as possible, and return loss at the input, through and coupled ports as high as possible across 32–38 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using an EM simulator from 20 to 50 GHz. Layout the designed coupler in the employed CMOS structure and make the size as small as possible.

8.16 Using any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15, design a 50-Ω based 30–40 GHz 3-dB unfolded or folded Lange coupler using parallel-coupled microstrip lines whose strips lie on the top-most metal layer and ground plane on the bottom-most metal layer. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 90° as possible, directivity as high as possible, and return loss at the input, through and coupled ports as high as possible across 30–40 GHz. Describe possible techniques to reduce the size of the designed Lange coupler. Describe all the design steps and present the simulations for *S*-parameters of the final design using both non-EM (circuit) and EM simulators from 20 to 50 GHz. Compare and comment on the results between non-EM and EM simulations.

8.17 Repeat Problem 8.16 using broadside-coupled transmission lines whose strips lie on different metal layers (e.g., M2–M6 in Figure P15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P15.2). Layout the designed Lange coupler in the employed CMOS/BiCMOS structure and make the size as small as possible.

8.18 Determine the *S*-parameters at the design center frequency for a 3-dB parallel-coupled-line directional coupler whose ports are shown in Figure 8.8. Compare these *S*-parameters to those in (8.105) of the branch-line coupler as shown in Figure 8.33. Are these *S*-parameters the same? Explain why they are different or same.

8.19 Using Eq. (8.97) and assume the port impedance is 50 Ω, determine the characteristic impedance of the ring hybrid with 180° phase shifter in the $3\lambda/4$ arm for the electrical length θ from 60° to 120° in 10° steps. Plot the amplitude and phase imbalance between the through and coupled ports of the ring hybrid corresponding to these electrical length θ, assuming the ring's transmission line is perfect, as a function of frequency. Compare the bandwidths between different electrical lengths based on the amplitude and phase imbalance (arbitrarily chosen) and draw a conclusion.

8.20 Using any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15, design a conventional ring hybrid operating from 33 to 37 GHz using a microstrip line with the strip on the top-most metal layer (e.g., M6 in Figure P.15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P.15.2). The port impedance is 50 Ω. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 180° as possible, and return loss at the input, through and coupled ports as high as possible across 33–37 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using both non-EM (i.e., circuit) and EM simulators from 30 to 40 GHz. Compare and comment on the results between non-EM and EM simulations. Now using the EM-simulated results on return losses as well as an amplitude and phase imbalance between the input port and two output ports, determine an appropriate operating bandwidth (lower and upper frequencies) over which the hybrid has good amplitude and phase balances with decent return losses. Provide these balances and return losses. The determined frequency range may be different from 33 to 37 GHz.

8.21 Using any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15, design a 30–40 GHz ring hybrid using two different layers for the ring (e.g., M5 and M6 in Figure P15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P15.2). The port impedance is 50 Ω. Implement a 180° phase shifter in the $3\lambda/4$ arm and meandering technique to reduce the circuit size. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 180° as possible, and return loss at the input, through and coupled ports as high as possible across 30–40 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using an EM simulator from 25 to 45 GHz.

8.22 Consider a conventional ring hybrid and assume the ring's transmission line is perfect, plot the amplitude and phase imbalance between the through and coupled ports as a function of normalized frequency f/f_o, where f_o is the design center frequency. Estimate the operating bandwidths based on different amplitude and phase imbalances (e.g., ±0.5 dB and ±6° around 180°).

8.23 Using any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15, design a 33–37 GHz branch-line coupler using a microstrip line with the strip on the top-most metal layer (e.g., M6 in Figure P15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P15.2). The port impedance is 50 Ω. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 90° as possible, directivity as high as possible, and return loss at the input, through and coupled ports as high as possible across 33–37 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using a non-EM (circuit) and an EM simulator from 30 to 40 GHz. Compare and comment on the results between non-EM and EM simulations.

8.24 Consider a branch-line coupler and assume the transmission line in each branch is perfect, plot the amplitude and phase imbalance between the through and coupled ports as a function of normalized frequency f/f_o, where f_o is the design center frequency. Estimate the operating bandwidths based on different amplitude and phase imbalances (e.g., ±0.5 dB and ±4° around 90°). If you also work on Problem 8.21, then compare the bandwidth for the (conventional) ring hybrid and branch-line coupler for the same amplitude and phase imbalance. If for a given specification, the bandwidths are different, explain why.

8.25 Calculate and plot the attenuation in dB versus $|\omega/\omega_c| - 1$ for the maximally flat low pass filter having maximum insertion loss of 3 dB and order of 11–15.

8.26 Calculate and plot the attenuation in dB versus $|\omega/\omega_c| - 1$ for the Chebyshev low pass filter having 0.01, 0.5, and 2 dB ripple for different orders from 1 to 10.

8.27 Calculate and list in tables the element values for the maximally flat low pass filter prototype having 3-dB maximum insertion loss and order from 11 to 15.

8.28 Calculate and list in tables the element values for the Chebyshev low pass filter prototype having 0.01, 0.5, and 2 dB ripple and order from 1 to 10.

8.29 Design a low pass filter having 0.1-dB ripple in the pass band, cutoff frequency of 5 GHz, and minimum rejection of 30 dB at 7 GHz. The source and load impedances are 50 Ω. Calculate the frequency response of the designed filter (both insertion loss and return loss) using your own computer program or a commercially available program. Does your filter meet the specifications that you use in the design? If not, then what should you do to meet the specifications?

8.30 Repeat Problem 8.29 for a maximally flat low pass filter with 3-dB maximum IL. Compare and comment on the responses between this filter and the filter in Problem 8.29.

8.31 Realize the low pass filter designed in Problem 8.28 using any available CMOS structure or the CMOS profile shown in Figure P15.2 with spiral inductors and metal–insulator–metal (MIM) capacitors. The spiral inductors are on the top-most metal layer (e.g., M6 in Figure P15.2). Present the simulation results for the *S*-parameters of the filter using a (non-EM) circuit simulator and an EM simulator up to 10 GHz. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment of the simulation results between the circuit and EM simulators.

8.32 Prove that a transmission line having high characteristic impedance Z_o and physical length ℓ less than 1/8th of a wavelength is approximately equivalent to a series inductor whose inductance is $L = \frac{Z_o}{v_p}\ell$ where v_p is the phase velocity.

8.33 Prove that a transmission line having small characteristic impedance Z_o and physical length ℓ less than 1/8th of a wavelength is approximately equivalent to a shunt capacitor whose capacitance is $C = \frac{\ell}{Z_o v_p}$ where v_p is the phase velocity.

8.34 Design a transmission line low pass filter having 0.1-dB ripple, 20-GHz cutoff frequency, and at least 30-dB rejection at 26 GHz based on Reference [12] as described in Section Transmission-Line Low Pass Filter Design. Use any available CMOS structure or the CMOS profile shown in Figure P15.2 and a microstrip line with the strip on the top-most metal layer (e.g., M6 in Figure P15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P15.2). The port impedance is 50 Ω. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band to 20 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using a non-EM (circuit) simulator and an EM simulator from 10 KHz to 40 GHz. Compare and comment on the results between non-EM and EM simulations.

8.35 Design a high pass filter having 0.1-dB ripple in the pass band, cutoff frequency of 10 GHz, and minimum rejection of 30 dB at 7 GHz. The source and load impedances are 50 Ω. Calculate the frequency response of the designed filter (both insertion loss and return loss) using your own computer program or a commercially available program. Does your filter meet the specifications that you use in the design? If not, then what should you do to meet the specifications?

8.36 Repeat Problem 8.35 for a maximally flat high pass filter with 3-dB maximum IL. Compare and comment on the responses between this filter and the filter in Problem 8.35.

8.37 Draw the schematics of band-pass filters corresponding to the low pass filter prototypes in Figure 8.51(b) for both *n* even and odd.

8.38 Design a lumped-element Chebyshev band-pass filter having 0.5-dB ripple, pass band from 20 to 30 GHz, and minimum rejection of 40 dB at 15 GHz. The source and load impedances are 50 Ω.

8.39 Repeat Problem 8.38 for a maximally flat band-pass filter with 3-dB maximum IL. Compare and comment on the responses between this filter and the filter in Problem 8.38.

8.40 Realize the band-pass filter designed in Problem 8.38 using any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15 with spiral inductors and MIM capacitors. The spiral inductors are on the top-most metal layer (e.g., M6 in Figure P15.2). Present the simulation results for the *S*-parameters of the filter using a circuit simulator and an EM simulator from 10 to 40 GHz. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment on the simulation results between the circuit and EM simulators.

8.41 Draw the schematics of band-stop filters corresponding to the low pass filter prototypes in Figure 8.51(b) for both *n* even and odd.

8.42 Design a lumped-element Chebyshev band-stop filter having 0.5-dB ripple corresponding to the stop-band edges at 20 and 30 GHz, and minimum rejection of 30 dB at 23 and 26 GHz. The source and load impedances are 50 Ω.

8.43 Repeat Problem 8.42 for a maximally flat band-stop filter with 3-dB insertion loss at the stop-band edges. Compare and comment on the responses between this filter and the filter in Problem 8.42.

8.44 Realize the band-stop filter designed in Problem 8.42 using any available CMOS structure or the CMOS profile shown in Figure P15.2 with spiral inductors and MIM capacitors. The spiral inductors are on the top-most metal layer (e.g., M6 in Figure P15.2). Present the simulation results for the *S*-parameters of the filter using a circuit simulator and an EM simulator from 10 to 40 GHz. Perform the necessary optimization to achieve the required specifications with return loss as high and low as

possible across the pass bands and the stop-band, respectively. Comment on the simulation results between the circuit and EM simulators.

8.45 Starting from the low pass filter prototype shown in Figure 8.51(a) for n odd, draw the schematics of corresponding band-stop filters using admittance inverters and impedance inverters.

8.46 Derive Eqs. (8.181)–(8.183).

8.47 Design an end-coupled Chebyshev band-pass filter having 0.2-dB ripple, pass band from 34 to 38 GHz, and minimum rejection of 20 dB at 41 GHz. The source and load impedances are 50 Ω. The resonators are realized using a microstrip line on any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15. The top conductor is on the top-most metal layer (e.g., M6 in Figure P15.2) and the (assumed infinitely large) bottom conductor (ground plane) is on the lowest metal layer (e.g. M1 in Figure P15.2). Present the simulation results for the S-parameters of the filter using a circuit simulator and an EM simulator from 20 to 50 GHz. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment on the simulation results between the circuit and EM simulators. Include a layout of the filter if possible.

8.48 Design a three-element end-coupled Chebyshev band-pass filter having 0.1-dB ripple and pass band from 30 to 40 GHz. The source and load impedances are 50 Ω. The resonators are realized using a microstrip line with broadside coupling on any available CMOS structure or the CMOS profile shown in Figure P15.2 in the problem section of Chapter 15. The top conductor is alternated between M6 and M4 in Figure P15.2 (or equivalent metal layers in any available CMOS/BiCMOS structure). The (assumed infinitely large) bottom conductor (ground plane) is on M1 in Figure P15.2 (or the lowest metal layer in any available CMOS structure). Present the simulation results for the S-parameters of the filter using a circuit simulator (if circuit simulation is feasible for this type of circuit structure) and an EM simulator from 20 to 50 GHz. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment on the use of broadside coupling. Include a layout of the filter if possible.

8.49 Design a parallel-coupled Chebyshev band-pass filter having 0.2-dB ripple, pass band from 34 to 38 GHz, and minimum rejection of 20 dB at 41 GHz. The source and load impedances are 50 Ω. The resonators are realized using a microstrip line on any available CMOS structure or the CMOS profile shown in Figure P15.2. The top conductor is on the top-most metal layer (e.g., M6 in Figure P15.2) and the (assumed infinitely large) bottom conductor (ground plane) is on the lowest metal layer (e.g., M1 in Figure P15.2). Present the simulation results for the S-parameters of the filter using a circuit simulator and an EM simulator from 20 to 50 GHz. Perform the necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment on the performance between this filter and that of the end-coupled band-pass filter in Problem 8.47. Include a layout of the filter if possible.

8.50 Design a three-element parallel-coupled Chebyshev band-pass filter having 0.1-dB ripple and pass band from 30 to 40 GHz. The source and load impedances are 50 Ω. The resonators are realized using a microstrip line with broadside coupling on any available CMOS structure or the CMOS profile shown in Figure P15.2. The top conductor is alternated between M6 and M4 in Figure P15.2 (or equivalent metal layers in any available CMOS structure). The (assumed infinitely large) bottom conductor (ground plane) is on M1 in Figure P15.2 (or the lowest metal layer in any available CMOS structure). Present the simulation results for the S-parameters of the filter using a circuit simulator (if circuit simulation is feasible for this type of circuit structure) and an EM simulator from 20 to 50 GHz. Perform necessary optimization to achieve the required specifications with insertion loss as low as possible and return loss as high as possible across the pass band. Comment on the performance between this filter and that of the end-coupled band-pass filter in Problem 8.48. Comment on the use of broadside coupling. Include a layout of the filter if possible.

8.51 Using any available CMOS structure or the CMOS profile shown in Figure P15.2, design a conventional ring hybrid operating from 33 to 37 GHz using a microstrip line with the strip on the top-most metal layer (e.g., M6 in Figure P15.2) and ground plane on the bottom-most metal layer (e.g., M1 in Figure P15.2). The port impedance is 50 Ω. Perform necessary optimization to achieve the required 3-dB coupling with an amplitude variation as small as possible, phase imbalance between the through and coupled ports as close to 180° as possible, and return loss at the input, through and coupled ports as high as possible across 33–37 GHz. Describe all the design steps and present the simulations for *S*-parameters of the final design using both non-EM (i.e., circuit) and EM simulators from 30 to 40 GHz. Compare and comment on the results between non-EM and EM simulations. Now using the EM-simulated results on return losses as well as the amplitude and phase imbalance between the input port and two output ports, determine an appropriate operating bandwidth (lower and upper frequencies) over which the hybrid has good amplitude and phase balances with decent return losses. Provide these balances and return losses. The determined frequency range may be different from 33 to 37 GHz.

CHAPTER 9

FUNDAMENTALS OF CMOS TRANSISTORS FOR RFIC DESIGN

CMOS (complementary metal-oxide-silicon) devices including transistors and diodes are the key elements in CMOS RFICs (radio frequency integrated circuits), performing various functions such as amplification, signal generation, mixing, and switching. Advances in CMOS technology have pushed the operation of these devices throughout the microwave region (1–30 GHz) and into the lower region of the millimeter-wave spectrum (30–300 GHz). It is expected that silicon-based CMOS (and related BiCMOS, etc.) can eventually perform at very high frequencies. Many excellent textbooks have written about CMOS devices and their detailed information can be found in these. In this chapter, we present only fundamentals of CMOS transistors, which are deemed useful for the design of CMOS RFICs.

9.1 MOSFET BASICS

MOSFETs (metal-oxide-semiconductor field-effect transistors) are junction field-effect transistors (JFETs) with a Schottky-junction gate made from silicon (Si) substrate. There are two kinds of MOSFET in CMOS technologies: n-channel or n-type (n-MOSFET, NMOSFET, or simply NMOS) and p-channel or p-type (p-MOSFET, PMOSFET, or simply PMOS). In most CMOS technologies, NMOS transistors perform better than their PMOS counterparts and are thus typically preferred in various RFICs. For example, NMOS transistors have higher current drive and transconductance due to higher mobility of electrons as compared to that of holes. For a given gate length and width, and bias currents, NMOS transistors have higher output resistance, leading to more ideal current sources and higher gain in amplifiers.

9.1.1 MOSFET Structure

Figure 9.1 shows simplified structures of typical NMOS and PMOS transistors. The NMOS transistor is fabricated on a p-type Si substrate (or bulk) with heavily doped n-type regions for the source (S) and drain (D). The PMOS transistor employs a complementary structure with an n-type Si substrate (or bulk) and heavily

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

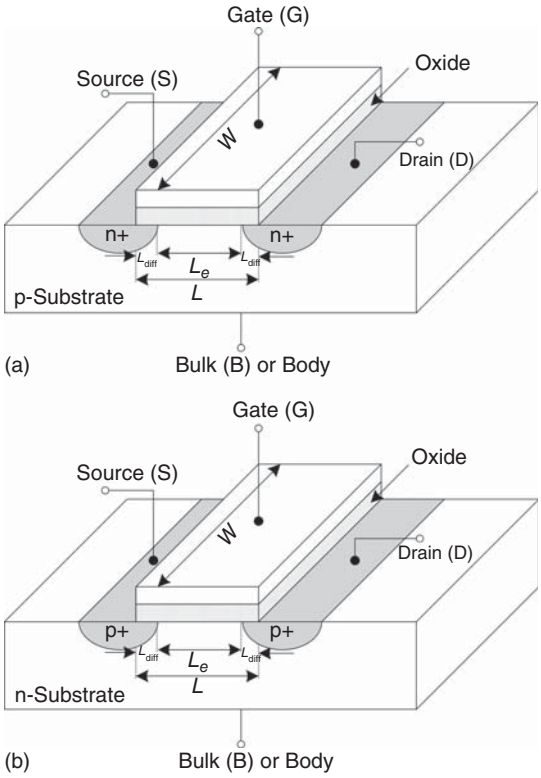


Figure 9.1. Simplified structures of typical NMOS (a) and PMOS (b) transistors.

doped p-type (p+) regions for the source and drain. The p- and n-type regions in these transistors effectively form (parasitic) p–n junction diodes, which are reversed biased in normal operations. In practical CMOS RFICs, however, both the NMOS and PMOS transistors are fabricated on the same (p-type) Si substrates. In order to separate from one another, either NMOS or PMOS transistor is fabricated within a particular area in the p-substrate called "well." In most CMOS processes, the PMOS devices are fabricated within an n-well in a p-substrate, as shown in Figure 9.2 [1]. In both NMOS and PMOS transistors, a heavily doped conductive polysilicon (or poly) constitutes the gate (G) of the transistors, and the substrate or bulk essentially forms the fourth terminal, effectively making MOSFET a four-terminal device, as seen in Figures 9.1 and 9.2. Typically, the source and bulk (B) terminals are tied with each other. In some circuits, however, the bulk is floated by connecting its terminal with a high resistance (e.g., 10 kΩ). The floating can also be achieved by using an inductor or a parallel RC resonator at a desired frequency which, while providing less noise, results in a larger size as compared to using a resistor. It is noted that it is still desired to have the DC voltage at the floating point to be zero; as such a capacitor is not suitable to be used as a floating element. Bulk-floating achieves some desired characteristics such as reduced parasitic capacitances. Bulk-floating also improves the linearity and power handling capability of MOSFETs under certain gate and drain bias voltages because their parasitic p–n junction diodes can be kept from being forward-biased under large input signals. Moreover, floating the bulk forces the bulk resistances underneath the source and drain junctions to open with respect to the ground, leading to smaller resistive loss in the conductive bulk and hence lower substrate loss, than with the bulk grounded. The floating bulk can also be negatively biased to further improve the linearity and power handling.

The drain and source electrodes form ohmic contacts to the active layer, whereas the gate electrode forms a Schottky barrier. The gate terminal sits over a thin layer of SiO_2, typically around 10 nm, which insulates the gate from the Si substrate. Instead of polysilicon, the gate can also be made from metal such as copper. The gate acts essentially as the top plate of a parallel-plate capacitor. During the fabrication process, the source and drain regions diffuse into the gate region, causing an overlap between the source and gate terminals and

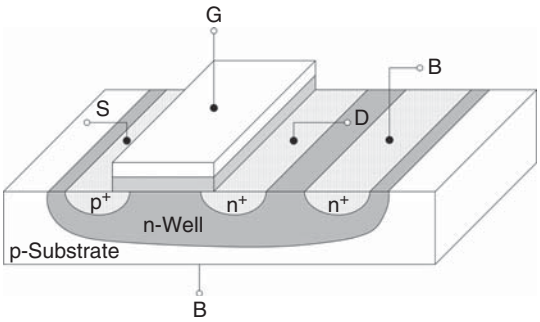


Figure 9.2. Simplified structure of a PMOS transistor fabricated in the same (p-type) substrate with NMOS transistors.

between the drain and gate terminals. This results in an effective gate length, L_e, less than the actual physical gate length, L, as $L_e = L - 2L_{\text{diff}}$, where L_{diff} is the diffused length due to lateral diffusion shown in Figure 9.1.

Figure 9.3 shows the top-view of a three-finger 0.35-μm MOSFET.

Figure 9.4 shows typical symbols used for NMOS and PMOS transistors. The bulk terminal is omitted in Figure 9.4(b) and (d) for brevity, which are typically used when the bulk is tied either to the ground or to V_{dd} (the voltage supplied to the drain), as normally used in most circuits.

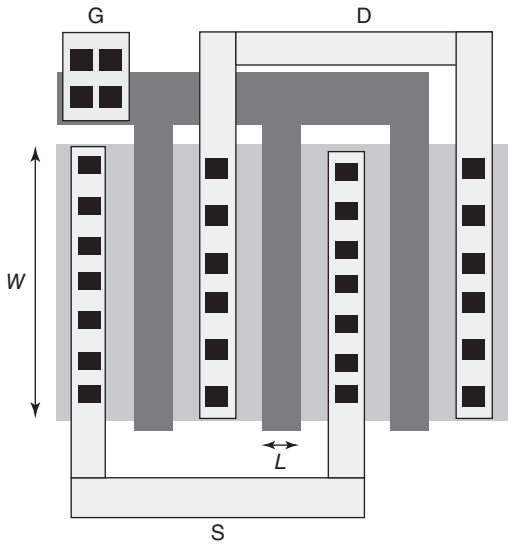


Figure 9.3. Top-view of a three-finger 0.35-μm MOSFET. The total gate width and length are $3W$ and L, respectively. The rectangular squares are contacts (poly to metal 1).

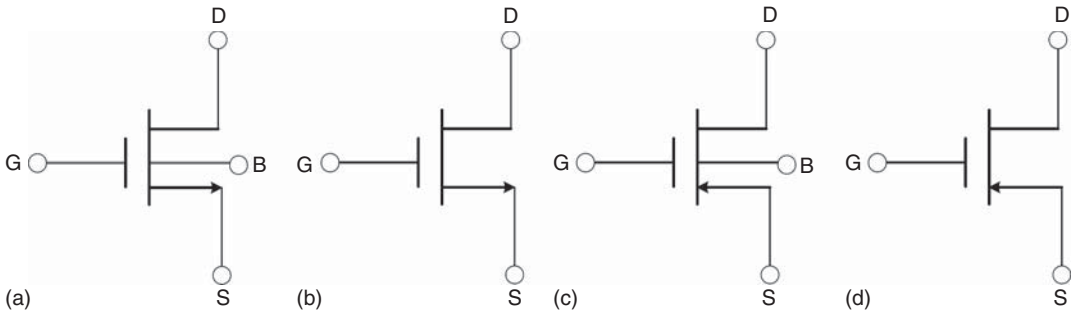


Figure 9.4. Typical symbols for NMOS (a, b) and PMOS (c, d) devices. The fourth terminal (B) is shown explicitly in (a) and (c).

The two most important physical parameters of MOSFETs for CMOS RFIC design are perhaps the transistor's gate length and gate width. A reduction in the gate length improves the gain, noise figure, and frequency of operation. On the other hand, increasing the gate width enhances the radio frequency (RF) power capability. Additionally, the gate length and width also affect the parasitic resistances and capacitances of transistors. For instance, increasing the gate width reduces the resistance accounting for the resistive loss in the bulk between the source and drain. Large devices, however, have significant parasitic capacitances, causing considerable effects to circuit matching and eventually limiting circuit's bandwidth – especially in the high frequency regions. These parasitic capacitances are more pronounced in sub-micrometer CMOS processes. For a given gate length, trade-off between different gate widths is necessary for optimized circuit performance.

9.1.2 MOSFET Operation

9.1.2.1 Biasing. Figure 9.5 shows typical bias schemes used in NMOS and PMOS devices [2]. The source and bulk terminals are assumed to be tied together. As shown, the bulk terminal of NMOS and PMOS transistors must always be connected to the respective lowest and highest voltage in a circuit, which are typically at the ground and location where V_{dd} is provided to the drain, respectively. In normal operation (i.e., active or conducting operation), the bias voltages applied to the gate and drain terminals of NMOS transistors, with respect to the source, are all positive, whereas those for PMOS devices are negative. Accordingly, positive current flows from drain to source and source to drain in NMOS and PMOS transistors, respectively. Moreover, the source–drain diode junction is always reversed-biased in normal operations. For PMOS devices fabricated within an n-well as shown in Figure 9.2, the well must therefore be connected to a potential that keeps the drain–source junction diode in reverse-biased condition, which is typically the highest positive voltage in the circuit.

9.1.2.2 Operating Principle. For the sake of simplicity, it is assumed that the source and bulk are tied together. MOSFET operation is generally based on the modulation of the channel or drain–source current, I_{ds}, by a vertical electric field existing between the gate and the substrate across the silicon dioxide layer, which is resulted from the applied voltages between the gate and source (V_{gs}) and between the drain and source (V_{ds}).

Figure 9.6 shows the I–V curves describing the drain–source current, I_{ds}, as a function of the drain–source voltage, V_{ds}, for different gate–source voltage, V_{gs}. In general, the operation mode of MOSFET transistors can be classified into two regions, based on the behavior of I_{ds} versus V_{ds} and V_{gs}, as linear (or triode) region and saturation region as shown in Figure 9.6. The MOSFET's operation also depends on the magnitude of

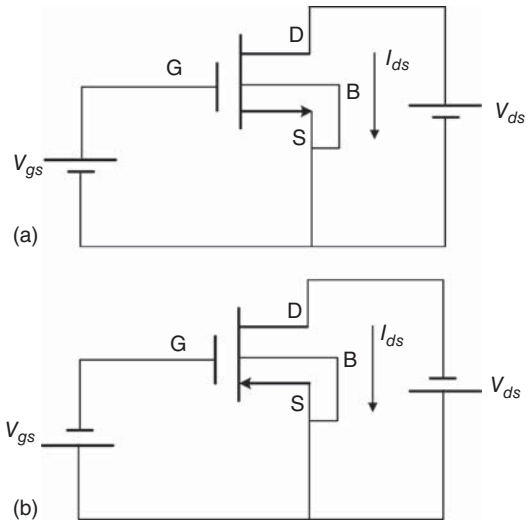


Figure 9.5. Typical bias schemes for NMOS (a) and PMOS (b) transistors.

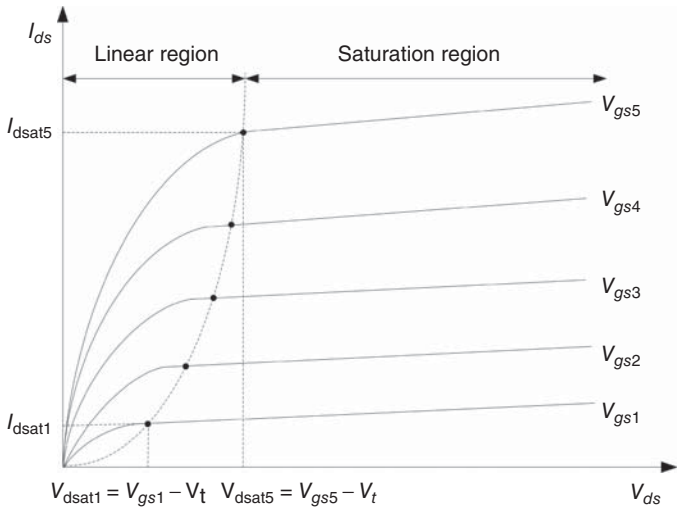


Figure 9.6. Typical $I-V$ characteristics of MOSFETs showing two different linear and saturation operating modes. V_t is $V_{tm}(V_{tp})$ for NMOS (PMOS) device. For short-channel devices, as V_{ds} is increased in the saturation region, the current increases slightly due to the effect of channel-length modulation.

applied bias voltages for a given device, upon which a MOSFET can be characterized as having long- or short-channel behavior.

When $V_{gs} = 0$, a positive (negative) V_{ds} in the NMOS (PMOS) transistor reverse-biases the p–n junction diode formed between the p (n)-type bulk and n+ (p+) drain, resulting in no free charges between the drain and source. Therefore, no conduction would occur between the drain and source terminals – the drain–source current is essentially zero. The MOSFET is said to be in the *off* or nonconducting state. Off-state MOSFETs are essentially passive devices, which cannot be used for CMOS RFICs requiring amplification and mixing (except passive mixing). However, they are useful for switching and phase shifting applications, such as transmit–receive (T-R) switches for RF transceivers and phase shifters for RF phased-array antennas, as well as for passive mixers for receivers requiring very low intermodulation.

Now assuming the source and drain are at the same potential (tied together with the bulk); that is, $V_{ds} = 0$, and a nonzero gate–source voltage is applied to the gate. Under the application of V_{gs}, positive for NMOS and negative for PMOS device, conduction occurs between the drain and source terminals and a current I_{ds} flows from the drain to source. A conducting channel now exists under the gate between the drain and source. It is noted that V_{ds} must be positive (negative) for NMOS (PMOS) transistors.

As V_{gs} applied to the gate of NMOS (PMOS) device is slightly increased positively (negatively), a vertical electric field occurs between the gate and the substrate across the oxide layer, much the same as the electric filed existing between the two plates of a parallel-plate capacitor. The holes (electrons) of the p-type (n-type) substrate near the interface between the substrate and the oxide layer then move away from the interface, effectively forming a depletion region in the substrate underneath the oxide layer. This depletion region does not contain mobile carriers and, since there are also no free carriers between the drain and source as mentioned earlier, no conduction would occur between the drain and source when V_{ds} is applied to the drain. As V_{gs} of the NMOS (PMOS) transistor is increased (decreased) toward a threshold voltage, V_t, the MOSFET is still maintained in the off state.[1] The threshold voltage is an intrinsic parameter of the transistor and is positive (negative) for the NMOS (PMOS) device. The threshold voltage is different for NMOS and PMOS devices. In general, the threshold voltage is also different from one NMOS (PMOS) device to another NMOS (PMOS) device. It is, however, typically assumed constant for all NMOS (PMOS) devices in a CMOS technology (e.g., 0.25-μm CMOS).

When V_{gs} is increased (decreased) passing V_t, the vertical electric field becomes sufficiently strong to attract minority electron (hole) in the NMOS (PMOS) device from the substrate to the gate area [2]. The

[1] Actually, a very small current flows when $V_{gs} < V_t$.

(insulating) oxide layer, however, prevents these minority carries from reaching the gate, resulting in an accumulation of the electrons (holes) at the interface between the substrate and oxide layer. These electrons (holes) then form an inversion region between the drain and source of the NMOS (PMOS) transistor. This region represents a conducting channel (or simply channel) between the drain and source. It is noted that the carrier distribution across the channel is fairly uniform since the drain and source are held at the same potential.

Under the bias condition of V_{gs} that produces a conducting channel, a positive (negative) bias voltage is applied between the drain and source of the NMOS (PMOS) transistor. The applied V_{ds} produces a longitudinal electric field from the source (drain) to drain (source), which moves the electrons (holes) toward the drain (source) of the NMOS (PMOS) transistor. These mobile electrons (holes), in turn, produce a positive (negative) current flowing from the drain (source) to source (drain). The magnitude of this current depends on V_{gs}, V_{ds}, material properties, and device geometry and dimensions. It reaches a maximum value when $V_{gs} = 0$ at a certain value of V_{ds}. When V_{ds} is limited to

$$V_{ds} < V_{gs} - V_{tm} \quad \text{(for NMOS devices)} \tag{9.1}$$

and

$$V_{ds} > V_{gs} - V_{tp} \quad \text{(for PMOS devices)} \tag{9.2}$$

where V_{tm} and V_{tp} are the threshold voltages for NMOS and PMOS devices, respectively, the inversion channel remains almost constant along the device length. Under these conditions, the device is said to be in its linear region.

When V_{ds} is increased beyond its limit stated in Eq. (9.1) or (9.2), then the inversion layer retracts from the drain and so does not extend completely from the source to the drain. Under this condition, the channel is said to be *pinched-off* or the device is *pinched-off*. At *pinch-off*, the current I_{ds} becomes constant versus V_{ds}, and hence the device essentially operates in its saturation region. The range of V_{ds}, under which a saturation region occurs, varies for different V_{gs}.

The above discussion essentially defines three operating regions for MOSFETs: off region, linear region, and saturation region, with the linear and saturation regions indicated in the I–V curves of Figure 9.6.

Long channel refers to "low" electric field operation (i.e., when applied voltages are sufficiently small). On the other hand, short channel refers to "high" electric field operation (i.e., when applied voltages are sufficiently high). The definition of "long" or "short" is relative with respect to the induced electric field (more than with respect to the applied voltages). For short-gate-length CMOS transistors (e.g., $L = 0.18$ μm), even small applied voltages can cause high electric field effects, thus making the device operate as short-channel. So in CMOS devices used for RFICs, which are typically short-gate devices, the short-channel effects are more pronounced than in those used in analog circuits operating at low frequencies.

Long-Channel Operation

Linear Region. A linear region is defined as the region in which I_{ds} is approximately proportional to V_{ds}, which is achieved when V_{gs} is sufficiently large or V_{ds} is sufficiently small. In the linear region, a conducting channel is formed across the source to drain. The drain–source current is given as

$$I_{ds} = \mu_n C_{\text{ox}} \frac{W}{L} \left[\left(V_{gs} - V_t \right) V_{ds} - \frac{1}{2} V_{ds}^2 \right] \tag{9.3}$$

where μ_n is the electron mobility ($\text{m}^2/\text{V-s}$), W is the gate width, L is the (effective) gate length, which is also equal to the channel length, and

$$C_{\text{ox}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \tag{9.4}$$

is the capacitance per unit area (F/m^2) between the gate electrode and substrate, with ε_{ox} and t_{ox} being the dielectric constant (or permittivity) and thickness of the oxide layer, respectively. Note that when $V_{gs} = V_t$, the inversion layer begins to be formed.

Equation (9.3) shows that, when V_{ds} is small, I_{ds} varies linearly versus V_{ds} and, hence, the MOSFET in linear region acts as a voltage-control resistor controlled by V_{ds}. The resistance is determined by V_{gs}, transistor size, and process parameters. By varying V_{gs}, a range of resistance values can be obtained.

The channel charge density can be determined from

$$Q_n(x) = -C_{\text{ox}}[V_{gs} - V(x) - V_t] \tag{9.5}$$

where $V(x)$ and $Q_n(x)$ are the channel potential and charge density along the channel, respectively. For a given value of V_{gs}, with V_t being held constant, the magnitude of the charge density $|Q_n(x)|$ reduces as $V(x)$ is increased. Consider a location in the channel at the drain terminal, at which $V(x) = V_{ds}$. At a particular value of V_{ds}, referred to as V_{dsat}, $Q_n(x) = 0$. This results in $V_{gs} - V_{\text{dsat}} - V_t = 0$, hence leading to

$$V_{\text{dsat}} = V_{gs} - V_t \tag{9.6}$$

This particular value of V_{ds} defines the end of the linear region or the beginning of the saturation region. That is, when $V_{ds} \le V_{\text{dsat}} = V_{gs} - V_t$, the device is said to be in linear region.

At V_{dsat}, the drain–source current I_{ds} reaches a maximum value given by

$$I_{\text{dsat}} \equiv I_{ds,\max} = \frac{1}{2}\mu_n C_{\text{ox}} \frac{W}{L}(V_{gs} - V_t)^2 \tag{9.7}$$

as derived from (9.3) after making use of (9.6). For small V_{ds}, V_t is not a function of V_{ds}. Otherwise, it depends on V_{ds}.

Saturation Region. As V_{ds} is increased, I_{ds} increases. When $V_{ds} = V_{\text{dsat}}$, the current reaches a maximum value of I_{dsat} and the transistor enters the saturation region. In the saturation region, I_{ds} remains almost constant as V_{ds} is increased. Equation (9.7) suggests that I_{dsat} in the saturation region depends on the "square" of the gate-source voltage and is not a function of V_{ds}. Actually, I_{dsat} depends "weakly" on V_{ds}.

The transconductance of MOSFET devices is given as

$$g_m = \left.\frac{\partial I_{ds}}{\partial V_{gs}}\right|_{V_{ds}=\text{constant}} \tag{9.8}$$

Taking derivative of I_{dsat} in (9.7) with respect to V_{gs} and substituting into (9.8) gives the transconductance in the saturation region as

$$g_m = \mu_n C_{\text{ox}} \frac{W}{L}(V_{gs} - V_t) \tag{9.9}$$

or

$$g_m = \sqrt{2\mu_n C_{\text{ox}} \frac{W}{L} I_{\text{dsat}}} = \frac{2I_{\text{dsat}}}{V_{gs} - V_t} \tag{9.10}$$

by making use of (9.7). As expressed explicitly in Eq. (9.10) and the fact that I_{dsat} practically depends slightly on V_{ds}, the transconductance in the saturation region depends strongly on V_{gs} and weakly on V_{ds}. The transconductance basically measures the change of I_{ds} in response to change in V_{gs}. Effectively, it measures the efficiency of MOSFETs in converting voltage to current. A large-g_m device produces a large change in I_{ds} for a small change in V_{gs} and is thus desirable. The saturation region is commonly used for amplification purposes.

It is noted that all the foregoing equations for the linear and saturation regions are derived at DC. The values of I_{ds} obtained from the current equations constitute the DC behavior of MOSFET devices. Basically, they provide the I–V curves describing relation between current and voltages at DC which are needed for device modeling. The derived equations for I_{ds} are based on a rather simple principle. They can be used to understand the behavior of I_{ds} as a function of the gate length and width and applied bias voltages, etc.

However, they are not very accurate, particularly at high frequencies and for submicron devices. Theoretical expressions for I_{ds} from the physics of submicron devices at RF frequencies are very difficult to derive accurately, complicated, and computationally intense. On the other hand, accurate equations for I_{ds} based on actual measurement of the devices used in CMOS RFICs are relatively easy to obtained and hence commonly used. These are addressed in Section 9.2.

Short-Channel Operation

MOSFETs are considered operating under "short-channel condition" when the induced electric filed is sufficiently large. At sufficiently high electric fields, the carrier velocity does not increase with increasing electric field.[2] In silicon, when the electric field reaches approximately 1 V/μm, the electron drift velocity begins to saturate to around 10^5 m/s as seen from $v_n = \mu_n E$. As the gate length is reduced, the saturated drain–source voltage V_{dsat} decreases, signifying that the velocity saturates at smaller drain-source voltage as gate length is reduced.

The drain–source current I_{ds} saturates when the carrier velocity v_n reaches saturation in short-channel MOSFETs. At saturation, the corresponding drain–source voltage can be expressed as

$$V_{\text{dsat}} \simeq \frac{2L\mu_{\text{eff}}(V_{gs} - V_t)}{V_{gs} - V_t + 2L\mu_{\text{eff}}} \tag{9.11}$$

where

$$\mu_{\text{eff}} = \frac{\mu_o}{1 + \theta(V_{gs} - V_t)} \tag{9.12}$$

with μ_o representing the low field mobility, which is independent of field strength, and θ being a fitting parameter approximately equal to $10^{-7}/t_{\text{ox}}$ (V^{-1}).

The drain–source current can be derived as

$$I_{ds} = WC_{\text{ox}}v_{\text{sat}}\frac{(V_{gs} - V_t)^2}{V_{gs} - V_t + 2\frac{v_{\text{sat}}L}{\mu_{\text{eff}}}} \tag{9.13}$$

where $v_{\text{sat}} = 10^5$ m/s is the saturated carrier velocity.

An inversion layer exits during $V_{gs} \geq V_t$. When $V_{gs} < V_t$, I_{ds} decreases exponentially versus V_{gs}. That is, the channel still conducts current but very weak. This region is called the "weak-inversion" or "subthreshold" region. In this region, MOSFETs have small transconductance and are, therefore, not normally used in RFICs.

The short-channel effects result in saturation for the velocity. Hence, I_{ds} can be approximated as

$$I_{ds} \simeq v_{\text{sat}}WC_{\text{ox}}(V_{gs} - V_t) \tag{9.14}$$

where

$$v_{\text{sat}} = \mu_n E_{\text{sat}} \tag{9.15}$$

Applying $g_m \equiv \frac{\partial I_{ds}}{\partial V_{gs}}$ then gives

$$g_m \simeq v_{\text{sat}}WC_{\text{cox}} \tag{9.16}$$

9.2 MOSFET MODELS

Linear and nonlinear models of MOSFETs are needed for the design of CMOS RFICs. Simple theoretical models can be easily derived and used at low frequencies. These models, however, are not suitable for circuits operating in the RF region, particularly those in the high RF range and for submicron devices.

[2]For small electric fields, the velocity is proportional to electric field as $v_n = \mu_n E$.

Accurate models for MOSFETs suitable for RF applications can be classified into "physics-based" models and "measurement-based" (or empirical) models. The linear and nonlinear models are used for active RFICs. For passive RFICs, such as switches, phase shifters, and so on, it is more suitable to employ passive MOSFET models – for example, models corresponding to the "on" and "off" states of MOSFETs for CMOS switches.

9.2.1 Physics-Based Models

Physics-based models are based on the device physics, which are generally governed by the following equations:

$$\begin{aligned} \nabla^2\psi &= -\frac{e}{\varepsilon}(N_d - n) \\ q\frac{\partial n}{\partial t} &= \nabla \cdot \vec{J} \\ \vec{J} &= -en\vec{v} + eD\nabla n \\ \vec{J}_T &= \vec{J} + \varepsilon\frac{\partial \vec{E}}{\partial t} \end{aligned} \tag{9.17}$$

where n is instantaneous carrier distribution; N_d is donor concentration; e represents the electronic charge; ε denotes dielectric permittivity or dielectric constant, $\vec{E}$ is electric field; $\vec{v}$ is average carrier velocity; and ψ is electric scalar potential. For a given MOSFET device, Eq. (9.17) is solved for the unknown n, ψ, and $\vec{J}_T$ subject to boundary conditions imposed by the device geometry, external applied voltages, and velocity-electric field relation for charge carriers. A physics-based model provides direct relationship between the electrical performance and the geometry and physical parameters of devices and is, thus, very useful for device design and understanding of device operation. They are indeed crucial for device designers. Determining these models requires not only understanding of the device physics, but also using of the full-wave electromagnetic (EM) field analysis. These models are, however, very difficult to be derived accurately in the RF range, particularly at millimeter-wave frequencies, mainly due to the lack of accurate analysis methods and difficulty in the formulation process. Furthermore, inaccurate process parameters in immature processes also contribute to the difficulty of developing accurate physics-based models. Moreover, the computation time in using these models for circuit design, especially complex CMOS RFICs containing many MOSFETs operating over a wide frequency range, is prohibitively extensive. They are thus mainly used for the design of devices. These models have limited use in CAD (computer-aided design) programs for RFIC design due to (i) long computation time making it impractical for circuit design; (ii) unavailability of accurate physical parameters of devices (N_d, geometry, etc.) to RFIC designers; and (iii) very difficult and time consuming to model distributed effects at RF frequencies, particularly millimeter-wave frequencies. Details of physics-based MOSFET models can be found in many papers and books.

9.2.2 Empirical Models

Empirical models can be derived using numerical results obtained from physics-based models and/or measurements of devices under DC and RF operations at different bias voltages and across interested frequencies. Practical empirical models, however, are typically measurement-based, in which measured data and simple closed-form equations are used together to develop the models. For linear models, measured scattering (S) parameters over interested frequencies under a small-signal condition are needed. The nonlinear models, however, requires both small-signal measured S-parameters and DC data of devices; for example, I_{ds} versus V_{ds} for different values of V_{gs}. For convenience in device modeling and circuit simulation, I_{ds} is normally described by a closed-form equation derived to match the measured I_{ds} data. The empirical models do not need the physics and geometries of devices, yet, if properly obtained, can describe accurately the actual device

electrical performance due to direct relation between the models and measured performance of the devices. This type of model can be used for any solid-state devices including CMOS, BiCMOS, SiGe. These models are typically represented by lumped-element equivalent circuits and are, thus, very suitable for CAD and virtually employed by all commercially available programs. For CMOS RFIC design, SPICE or SPICE-like models are the most popular ones. SPICE models are classified into various types, including Level 1, 2 and 3, BSIM2 (Berkeley Short-Channel IGFET Model 2), BSIM3 (Berkeley Short-Channel IGFET Model 3), etc. SPICE or SPICE-like models have always been under refinement, and new and improved models are released frequently. These models as well as other possible models are typically available from CMOS foundry for particular CMOS processes for certain frequency ranges. CMOS RFIC designers frequently use these foundry models directly with proper bias conditions within the provided frequency ranges, instead of developing them by themselves, to design CMOS RFICs. However, when the operating frequencies are beyond those that the models are provided, extrapolation of these models or development of new models to those frequencies are often implemented.

9.2.2.1 Nonlinear Equivalent-Circuit Models. Figure 9.7(a) shows an exaggerated MOSFET structure illustrating model parameters and its corresponding nonlinear equivalent-circuit models. It is particularly noted that the models in Figure 9.7(b) and (c) contain two currents, I_{ds} and I_{dsb} controlled by V_{gs} and V_{bs}, respectively, flowing from the drain to source as described in Figure 9.8. In typical usages, the internal source and the bulk are connected. Under this condition, there is no potential difference between the source and bulk ($V_{bs} = 0$), resulting in $I_{bs} = 0$, and the model can be approximated as shown in Figure 9.7(c). It is noted that the equivalent circuits shown in Figure. 9.7(b) and (c) can also be used for small-signal models provided that all the elements are considered to be linear. Closed-form expressions for some of the parameters of the equivalent-circuit model can be derived from simple DC analysis, elementary physics, and simple device physics.

L_g, L_d, and L_s represent the gate, drain, and source electrode inductance. R_g, R_d, and R_s are the gate, drain, and source terminal resistance. These resistances are essentially due to the conductivity of the terminal material and the contact. R_{gb}, R_{db}, and R_{sb} are the substrate resistances between the gate and bulk, drain and bulk, and source and bulk, respectively. R_{dsb} represents the substrate resistance between the drain and source accounting for the resistive coupling between these terminals, which is different from the commonly known drain–source resistance R_{ds} extracted from the drain–source current I_{ds}. R_{gb}, R_{db}, R_{sb}, and R_{dsb} are the main sources of loss in MOSFET, primarily due to the lossy Si substrate. C_{ds} is the drain–source capacitance, accounting for the coupling between drain and source due to the lossy Si substrate.

C_{gs} and C_{gd} represents the total capacitance between the gate and source and gate to drain, respectively. These capacitances are basically the overlapping capacitances between the gate and source and gate and drain, including fringing capacitances, resulting from diffusion. The fringing capacitances occur along all the edges of the electrodes and contribute only lightly to the overall capacitances. The total capacitances including fringing capacitances can be accurately determined using a full-wave EM technique or a commercially available EM simulator. These capacitances can be approximated, neglecting fringing capacitances, using the parallel-plate capacitance equation, as

$$C_{gd} \simeq C_{gs} \simeq WL_{\text{diff}}C_{\text{ox}} \tag{9.18}$$

where W is the gate width; L_{diff} is the overlapping length relating to the actual physical length L and the effective length L_e of the gate as

$$L_e = L - 2L_{\text{diff}} \tag{9.19}$$

and

$$C_{\text{ox}} = \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \tag{9.20}$$

is the capacitance per unit area. $\varepsilon_{\text{ox}} \simeq 3.9\varepsilon_o$ and t_{ox} are the dielectric constant and thickness of the SiO_2 layer underneath the gate, respectively, with ε_o being the dielectric constant of air.

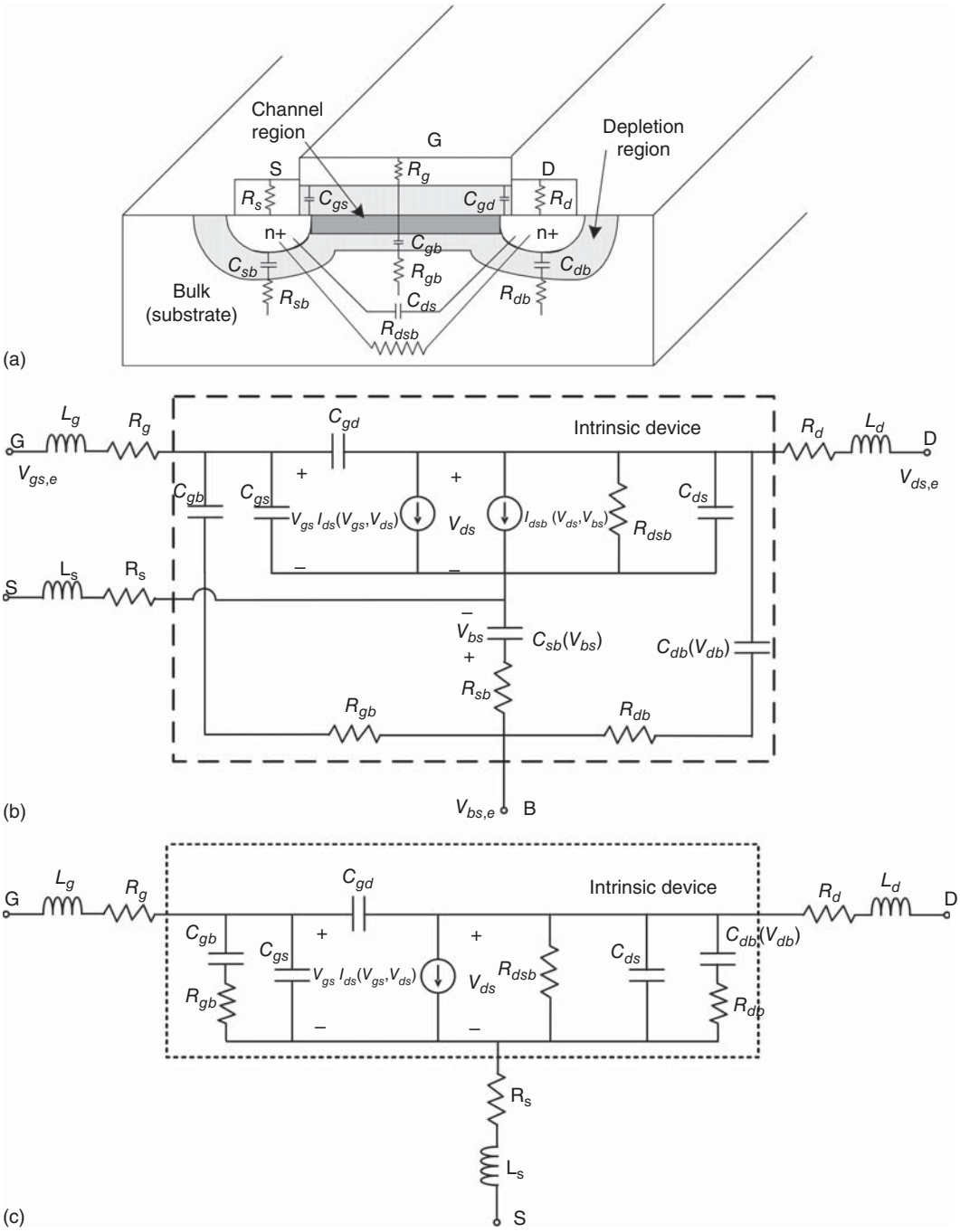


Figure 9.7. Typical MOSFET structure (a) and its equivalent-circuit models without (b) and with (c) connection between the bulk and internal source terminals. The models inside the dashed lines represent the intrinsic device. $V_{gs,e}$, $V_{ds,e}$, and $V_{bs,e}$ are the actual bias voltages applied to the device, which are different from corresponding V_{gs}, V_{ds}, and V_{bs} used for the intrinsic device.

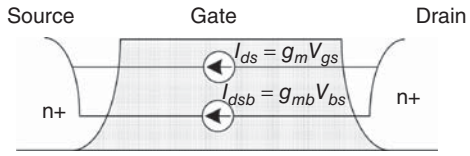


Figure 9.8. Currents flow between drain and source.

C_{gb} represents the total capacitance between the gate and bulk (substrate). It consists of the gate-to-channel capacitance C_{gs} and the channel-to-bulk capacitance C_{cb}, and can be derived as

$$C_{gb} = \frac{C_{gc}C_{cb}}{C_{gc} + C_{cb}} \tag{9.21}$$

where

$$C_{gc} \simeq C_{\text{ox}}WL \tag{9.22}$$

and

$$C_{cb} \simeq WL\frac{\varepsilon_{\text{si}}}{d} \tag{9.23}$$

with $\varepsilon_{\text{si}} \simeq 11.7\varepsilon_o$ being the dielectric constant of the silicon substrate and

$$d = \left(\frac{2\varepsilon_{\text{si}}}{qN_{\text{sub}}}\,|\phi_s - \phi_F|\right)^{1/2} \tag{9.24}$$

with ϕ_s being the surface potential, ϕ_F being the Fermi level, $q = 1.602 \times 10^{-9}$ C being the electron charge; and N_{sub} being the doping concentration of the silicon substrate. In the linear and saturation regions, $|\phi_s - \phi_F| \simeq 2\phi_F$.

C_{sb} and C_{db} are the junction capacitances between the source region and bulk and between the drain and bulk, respectively. They are function of the drain and source voltage with respect to the bulk, respectively, and are given by

$$C_{sb} = \frac{C_{sbo}}{\left(1 - \frac{V_{sb}}{\phi_o}\right)^{1/2}} \tag{9.25}$$

and

$$C_{db} = \frac{C_{dbo}}{\left(1 - \frac{V_{db}}{\phi_o}\right)^{1/2}} \tag{9.26}$$

where V_{sb} and V_{db} are the corresponding reverse-biased voltages across these junctions. C_{sbo} and C_{dbo} are the respective junction capacitances at 0 V; and ϕ_o is the junction built-in potential. It is noted that these capacitances depend on the transistor's bias conditions; that is, on the device's operating region. C_{sb} and C_{db} have low quality factor and hence can cause significant losses to the MOSFET especially at high frequencies.

Table 9.1 summarizes the capacitances C_{gs}, C_{gd}, C_{gb}, C_{sb}, and C_{db} in three different regions: off, linear, and saturation. The capacitance C_{ov} included in the table is defined as

$$C_{ov} = WL_{\text{diff}}C_{ox} \tag{9.27}$$

TABLE 9.1. Capacitances for the Off, Linear, and Saturation Regions

Capacitance	Off	Linear	Saturation
C_{gs}	C_{ov}	$C_{gc}/2 + C_{ov}$	$2C_{gc}/3 + C_{ov}$
C_{gd}	C_{ov}	$C_{gc}/2 + C_{ov}$	C_{ov}
C_{gb}	$C_{gc}C_{cb}/(C_{gc} + C_{cb})$	0	0
C_{sb}	C_{jsb}	$C_{jsb} + C_{cb}/2$	$C_{jsb} + 2C_{cb}/3$
C_{db}	C_{jdb}	$C_{jdb} + C_{cb}/2$	C_{jdb}

In the linear and saturation regions, the gate-bulk capacitance C_{gb} can be neglected because the channel (inversion layer) acts as a shield between the gate region and the bulk. However, this assumption is only approximately accurate at DC or low frequencies. At high frequencies, C_{gb} should not be neglected.

$I_{ds}(V_{gs}, V_{ds})$ is the voltage-dependent current source upon V_{gs} and V_{ds}. The transconductance g_m from the gate can be derived from I_{ds} as

$$g_m = g_{mo} e^{-j\omega\tau_g} \tag{9.28}$$

where

$$g_{mo} = \left.\frac{\partial I_{ds}}{\partial V_{gs}}\right|_{V_{ds}=\text{constant}} \tag{9.29}$$

is the corresponding transconductance at DC and τ_g is the transit time under the gate accounting for the phase delay. τ_g represents the internal time delay between the drain current and gate voltage. g_{mo} describes the dependence of I_{ds} on the gate voltage V_{gs}. It is noted that g_{mo} from Eq. (9.29), although is only strictly valid at DC, its value for good submicron devices is relatively constant with frequency, perhaps up to close to the device's cut-off frequency.

$I_{dsb}(V_{ds}, V_{bs})$ is the bulk-source/drain–source voltage-dependent current source. The transconductance g_{mb} from the bulk can be derived from I_{dsb} as

$$g_{mb} = g_{mbo} e^{-j\omega\tau_{gb}} \tag{9.30}$$

where

$$g_{mbo} = \left.\frac{\partial I_{dsb}}{\partial V_{bs}}\right|_{V_{ds}=\text{constant}} \tag{9.31}$$

is the corresponding transconductance at DC and τ_{gb} is the transit time under the bulk. τ_{gb} represents the phase delay, which is the internal time delay between the drain current and bulk voltage. It is noted that the bulk acts as a second gate of MOSFETs.

The drain–source current I_{ds} can also be used to determine the resistance R_{ds} between the drain and source, also known as the output resistance. This resistance plays an important role in the performance of circuits. For example, R_{ds} limits the maximum voltage gain in amplifiers. It can be obtained as

$$R_{ds} = \left.\frac{\partial V_{ds}}{\partial I_{ds}}\right|_{V_{gs}=\text{constant}} \tag{9.32}$$

which dictates the dependence of I_{ds} on the drain voltage V_{ds}.

The drain–source current can be expressed as

$$I_{ds} \simeq \frac{1}{2}\mu_n C_{\text{ox}} \frac{W}{L}(V_{gs} - V_t)^2(1 + \lambda V_{ds}) \tag{9.33}$$

or

$$I_{ds} \simeq I_{\text{dsat}}(1 + \lambda V_{ds}) \tag{9.34}$$

using Eq. (9.7), where λ is the channel-length modulation coefficient, which dictates the dependence of I_{ds} upon V_{ds} in the saturation region. This dependence causes the current to slightly increase in the saturation region as V_{ds} is increased for short-channel devices. λ can be determined from measured I–V curves of MOSFETs.

Employing Eq. (9.33), we can derive the drain–source resistance as

$$R_{ds} = \frac{\partial V_{ds}}{\partial I_{ds}} = \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)^{-1}$$

$$= \left[\frac{1}{2}\mu_n C_{\text{ox}}\lambda\frac{W}{L}\left(V_{gs}-V_t\right)^2\right]^{-1} \tag{9.35}$$

which can then be approximated as

$$R_{ds} \simeq \frac{1}{\lambda I_{ds}} \tag{9.36}$$

The gate transconductance at DC can be derived, using (9.33), as

$$g_{mo} = \frac{\partial I_{ds}}{\partial V_{gs}} = \mu_n C_{\text{ox}}\frac{W}{L}(V_{gs}-V_t)(1+\lambda V_{ds})$$
$$= \sqrt{2\mu_n C_{\text{ox}}\frac{W}{L}(1+\lambda V_{ds})I_{ds}} \tag{9.37}$$

The bulk transconductance at DC can be derived as

$$g_{mbo} = \frac{\partial I_{ds}}{\partial V_{bc}} = g_{mo}\frac{\gamma}{2\sqrt{2\phi_F + V_{sb}}} \tag{9.38}$$

where

$$\gamma = \frac{\sqrt{2q\epsilon_{\text{si}}N_{\text{sub}}}}{C_{\text{ox}}} \tag{9.39}$$

is called the body-effect coefficient, typically having a value between 0.3–0.4 $\text{V}^{1/2}$ and V_{sb} is the source-bulk potential difference.

Other parameters can be obtained from DC measurements and fitting the model's *S*-parameters to measured *S*-parameters of specific devices. They can also be determined from theoretical analysis, which is complex and is thus not practically realistic for circuit design. It is noted that all the foregoing equations assume "long channel" operation or "long channel" devices.

A nonlinear model consists of both linear and nonlinear elements, which are indicated by their bias dependence. Table 9.2 summarizes the linear and nonlinear elements in the nonlinear model of Figure 9.7(b), among which I_{ds} and I_{dsb} are considered the main contribution to the nonlinear behavior of MOSFETs.

It should be noted that the bias voltages used in determining the device characteristics such as the gate–source voltage V_{gs} and drain–source voltage V_{ds} for the drain–source current I_{ds} are the voltages applied to the intrinsic device. They should be distinguished from the actual bias voltages applied to the device. Consider the intrinsic model shown within the dashed lines of Figure 9.5(b). The drain–source, gate–source, and bulk-source voltages for the corresponding intrinsic device can be derived as

$$V_{ds} = V_{ds,e} - I_{ds}(R_s + R_d) \tag{9.40}$$
$$V_{gs} = V_{gs,e} - I_{ds}R_s \tag{9.41}$$

and

$$V_{bs} \simeq V_{bs,e} \tag{9.42}$$

where the subscript "*e*" is used to indicate the (external) actual bias voltages applied to devices.

Taking the derivative of both sides of Eq. (9.41) with respect to V_{gs} gives

$$1 = \frac{\partial V_{gs,e}}{\partial V_{gs}} - R_s\frac{\partial I_{ds}}{\partial V_{gs}} \tag{9.43}$$

TABLE 9.2. Linear and Nonlinear Elements in the Large-Signal Model Shown in Figure 9.7(b)

Linear elements	Nonlinear elements
L_g, L_d, L_s R_g, R_d, R_s $R_{gb}, R_{db}, R_{sb}, R_{dsb}$ C_{gs}, C_{gd}, C_{ds}	$I_{ds}(V_{ds}, V_{gs}), I_{dsb}(V_{ds}, V_{bs})$ $C_{sb}(V_{sb}), C_{db}(V_{db})$

which can be rewritten as

$$1 = \frac{\partial V_{gs,e}}{\partial V_{gs}}(1 - R_s g_{mo,e}) = \frac{\partial V_{gs,e}}{\partial I_{ds}} \frac{\partial I_{ds}}{\partial V_{gs}}(1 - R_s g_{mo,e}) \tag{9.44}$$

or

$$\frac{g_{mo}}{g_{mo,e}}(1 - R_s g_{mo,e}) = 1 \tag{9.45}$$

upon using (9.29), where g_{mo} and $g_{mo,e}$ are the DC transconductance for the intrinsic and actual device, respectively. The transconductance $g_{mo,e}$ of the device can now be expressed in term of the intrinsic transconductance g_{mo}, from (9.45), as

$$g_{mo,e} = \frac{g_{mo}}{1 + R_s g_{mo}} \tag{9.46}$$

and the intrinsic transconductance g_{mo} can be obtained from $g_{mo,e}$ as

$$g_{mo} = \frac{g_{mo,e}}{1 - R_s g_{mo,e}} \tag{9.47}$$

9.2.2.2 Linear Equivalent-Circuit Models. The equivalent circuits, as shown in Figure 9.7(b) and (c), consist of linear and nonlinear elements. These circuits can be used as linear models, provided that all of their nonlinear elements are described as linear elements – for instance, the junction capacitance C_{db} independent of the drain-bulk bias voltage. In the linear models, the current sources I_{ds} and I_{dsb} in Figure 9.7(b) and (c) are replaced by the linear transconductance parameters $g_m = g_{mo}e^{-j\varpi\tau_g}$ and $g_{mb} = g_{mbo}e^{-j\varpi\tau_{gb}}$, respectively, as illustrated in Figure 9.9. That is, g_{mo}, τ_g, and g_{mbo}, τ_{gb} are used in the linear model in lieu of I_{ds} and I_{dsb}, which are nonlinear function of V_{gs}, V_{ds} and V_{gs}, V_{bs}, respectively. Additionally, resistances representing the dependence of I_{ds} on V_{ds} and I_{dsb} on V_{ds} should also be included in the linear model. Figure 9.10 shows explicitly the linear equivalent-circuit models extracted from Figure 9.7(b) and (c). In these models, R_{dsb} actually includes the (resistive-coupling) substrate drain–source resistance R_{dsb} and the drain–source output resistance R_{ds} (dependent on V_{gs} and V_{ds}) in parallel, hence making the composite R_{dsb} dependent on V_{gs} and V_{ds} as well.

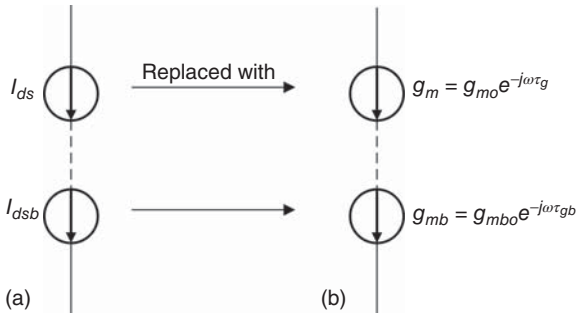


Figure 9.9. Transformation of currents from nonlinear (a) to linear (b) model.

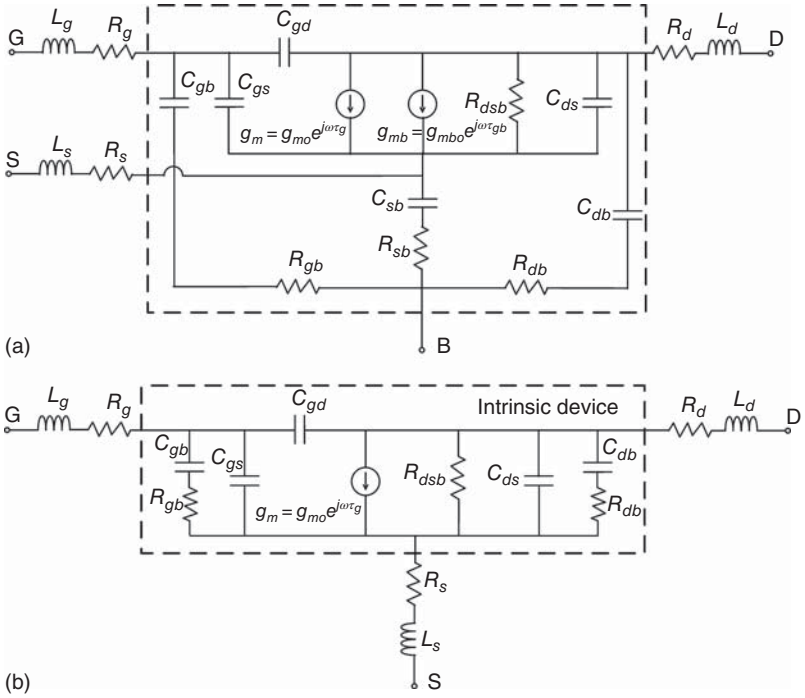


Figure 9.10. Linear equivalent circuit models when the bulk and internal source terminals are separated (a) and tied together (b). The values for R_{dsb}, C_{db}, C_{sb}, g_m, and g_{mb}, are typically at particular bias voltages at which the small-signal model is extracted. The current sources are $I_{ds} = g_m V_{gs}$ and $I_{dsb} = g_{mb} V_{bs}$ which are linear at applied V_{gs} and V_{bs}, respectively.

In practice, all the element values of the linear model are obtained by fitting the *S*-parameters measured at particular bias voltages over interested frequencies to those calculated from the equivalent-circuit model. Initial values used in the fitting process can be calculated using the equations provided in this chapter and/or obtained from DC measurements. The resulting linear model, as can be expected, is valid only at the values of the bias voltages used, which provide the small-signal behavior for the particularly measured device at these voltages over the frequency range measured and used in the fitting process. The difference between small-signal models at different bias voltages is mainly in the values of C_{sb}, C_{db}, R_{dsb}, g_m, and g_{mb}. A linear model with fixed element values can be developed for each set of bias voltages. It can also be extracted for different bias voltages by expressing $C_{sb}(V_{sb})$, $C_{db}(V_{db})$, $R_{dsb}(V_{ds}, V_{gs})$, $g_m(V_{ds}, V_{gs})$, and $g_{mb}(V_{ds}, V_{bs})$ as a function of corresponding bias voltages. The expressions for these functions can be obtained by fitting results obtained at different bias voltages into proper expressions. In general, in order to obtain an accurate model taking into account the process variation, devices at different locations on each wafer are measured and results are used to develop corresponding model. This is needed particularly at high RF frequencies and/or when a CMOS process is not yet stable. For stable CMOS processes, however, the model developed for a particular device may also be used for other devices on different locations of chip and wafer or on different chips and wafers.

9.2.2.3 Extraction of Equivalent-Circuit Models. To obtain the nonlinear model of Figure 9.7(b) or (c), all the linear as well as nonlinear elements need to be evaluated. This process, as outlined in this section, can also produce the linear models given in Figure 9.10.

Estimate of Linear Elements

Approximate values for the resistance R_s, R_d, and R_g can be determined following the procedure in [3]. Consider a device shown in Figure 9.11(a) with the source and bulk terminals tied together to ground, which is operated in the linear (or strong-inversion) region where a conducting channel exists under the bias voltages $V_{gs} > V_t$ and $V_{ds} = 0$; the small signal–signal equivalent circuit of the device can be simplified as shown in

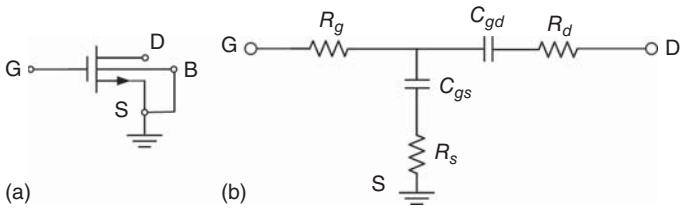


Figure 9.11. (a, b) A simplified device model in the linear region with $V_{gs} > V_t$ and $V_{ds} = 0$. Source and bulk are tied together.

Figure 9.11(b). The device's impedance parameters can be derived from Figure 9.11(b) as

$$Z_{11} = R_g + R_s - j\omega C_{gs} \tag{9.48}$$

$$Z_{22} = R_d + R_s - j\omega(C_{gs} + C_{gd}) \tag{9.49}$$

$$Z_{12} = Z_{21} = R_s - j\omega C_{gs} \tag{9.50}$$

from which we can extract

$$R_s = \mathrm{Re}(Z_{12}) \tag{9.51}$$

$$R_g = \mathrm{Re}(Z_{11} - Z_{12}) \tag{9.52}$$

$$R_d = \mathrm{Re}(Z_{22} - Z_{12}) \tag{9.53}$$

Figure 9.12 shows the calculated results for R_s, R_d, and R_g versus frequency using (9.51)–(9.53) for a 0.18-μm NMOS biased with $V_{gs} = 2\,\mathrm{V}(> V_t)$ and $V_{ds} = 0$. As can be seen, the resistances are almost constant with respect to frequency as expected. The results slightly change when V_{gs} is varied from 1.8 to 3 V.

The DC transconductance g_{mo}, g_{mbo} and the output resistance R_{ds} can be estimated from the device's measured I–V curves based on Eqs. (9.29), (9.31), and (9.32), respectively. It is noted that the measured g_{mo} represents the (external) transconductance $g_{mo,e}$, not the intrinsic transconductance g_{mo}, from the gate. These transconductances are related by Eq. (9.46) or (9.47). For instance, the (external) transconductance $g_{mo,e}$ can be determined directly from the measured I–V curve by taking the ratio between the incremental change in the measured current ΔI_{ds} and the corresponding gate–source voltage difference ΔV_{gs} at a particular V_{ds} as

$$g_{mo,e} = \left.\frac{\partial I_{ds}}{\partial V_{gs}}\right|_{V_{ds}=\text{constant}} \simeq \left.\frac{\Delta I_{ds}}{\Delta V_{gs}}\right|_{V_{ds}=\text{constant}} \tag{9.54}$$

from which the intrinsic transconductance g_{mo} can be determined. As noted earlier, the measured g_{mo} at DC may be used as a good approximation for the transconductance at RF frequencies below the device's

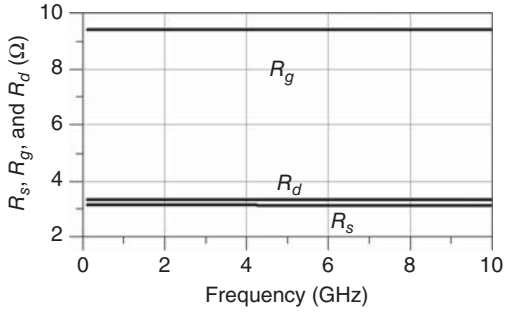


Figure 9.12. Calculated R_s, R_d, and R_g versus frequency.

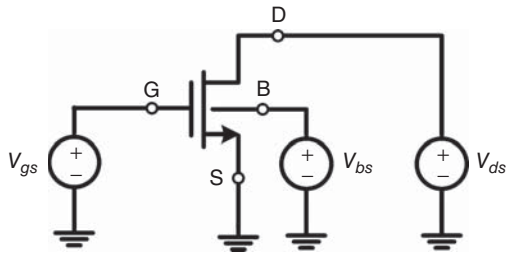


Figure 9.13. Circuit set-up used for extracting $I-V$ characteristics and device parameters.

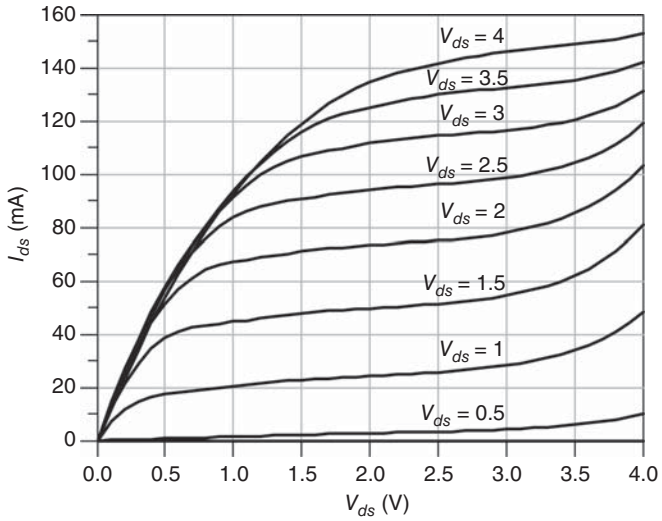


Figure 9.14. Extracted $I-V$ characteristics. $V_{bs} = 0$ V.

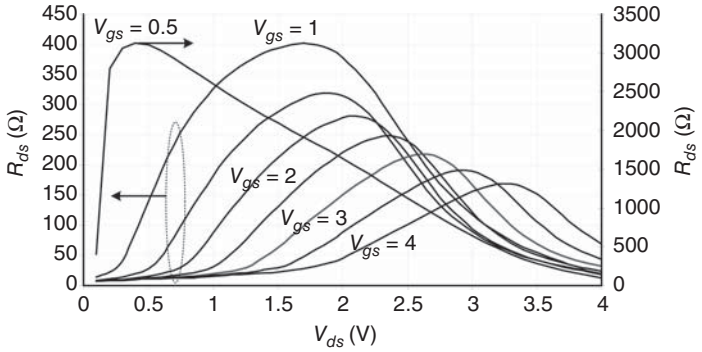


Figure 9.15. Calculated $R_{ds} \cdot V_{bs} = 0$ V.

cut-off frequency. Figures 9.13–9.17 show the circuit setup for parameter extraction, extracted $I-V$ curves, and calculated results for R_{ds} versus V_{ds} for different values of V_{gs} (at $V_{bs} = 0$ V), g_{mo} versus V_{gs} for different V_{ds} (at $V_{bs} = 0$ V), and g_{mbo} versus V_{bs} for different V_{ds} (at $V_{gs} = 1$ V) using Eqs. (9.32), (9.29), and (9.31), respectively, for a 0.18-μm NMOS having 10 fingers of 10 μm each. It is noted that R_{ds} is large when $V_{gs} = 0.5$ V since device is operated in weak inversion.

Parameters of the small-signal model can also be approximately determined based on the model's admittance matrix. We consider a small-signal model for intrinsic MOSFETs, with the internal source and bulk

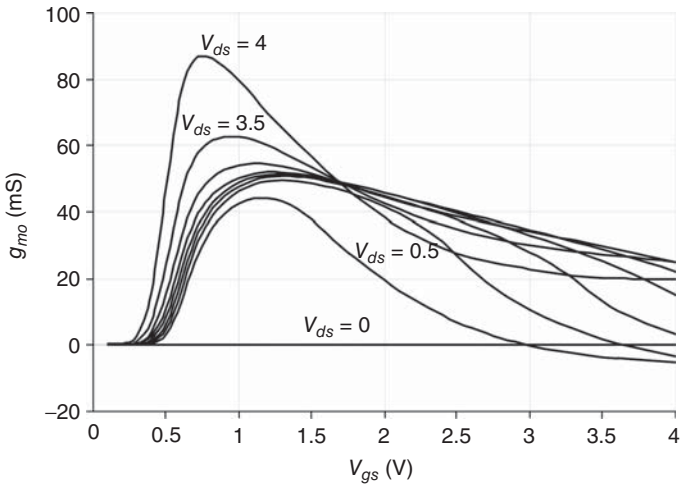


Figure 9.16. Calculated $g_{mo} \cdot V_{bs} = 0$ V.

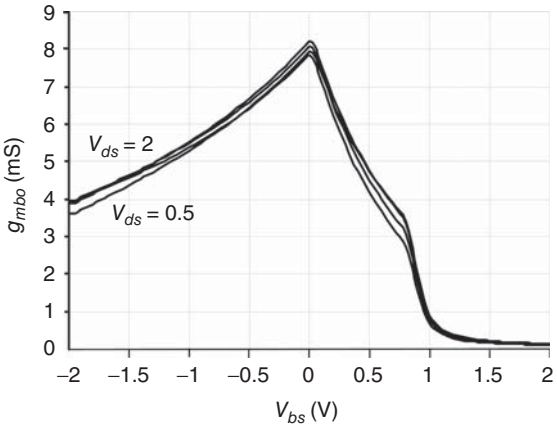


Figure 9.17. Calculated $g_{mbo} \cdot V_{gs} = 1$ V.

tied together to ground, as shown in Figure 9.18, which can be extracted from Figure 9.10(b). The parameters of the corresponding admittance matrix of the intrinsic model with ports 1 and 2 at the gate and drain, respectively, can be derived as

$$Y_{11} = \left.\frac{I_1}{V_1}\right|_{V_2=0} = j\omega(C_{gd} + C_{gs}) + \frac{j\omega C_{gb}}{1 + j\omega R_{gb} C_{gb}} \tag{9.55}$$

$$Y_{21} = \left.\frac{I_2}{V_1}\right|_{V_2=0} = g_m - j\omega C_{gd} \tag{9.56}$$

$$Y_{12} = \left.\frac{I_1}{V_2}\right|_{V_1=0} = -j\omega C_{gd} \tag{9.57}$$

$$Y_{22} = \left.\frac{I_2}{V_2}\right|_{V_1=0} = j\omega(C_{gd} + C_{ds}) + g_{dsb} + \frac{j\omega C_{db}}{1 + j\omega R_{db} C_{db}} \tag{9.58}$$

where $g_{dsb} = 1/R_{dsb}$. We can obtain g_m and C_{gd} from (9.56) and (9.57), respectively, as

$$g_m = \mathrm{Re}(Y_{21}) \tag{9.59}$$

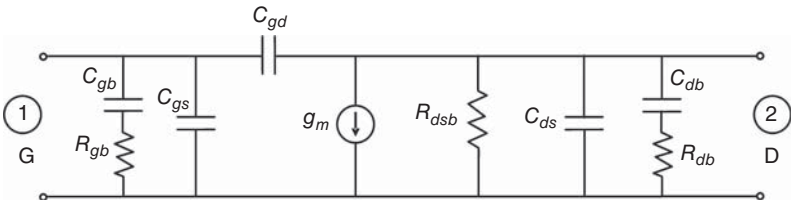


Figure 9.18. Intrinsic small-signal signal model of MOSFETs with the source and bulk tied together.

$$C_{gd} = -\frac{\text{Im}(Y_{12})}{\omega} \tag{9.60}$$

where Re (.) and Im (.) stand for the real and imaginary parts, respectively. Adding (9.55) to (9.57) and neglecting C_{gb} due to the fact that it is typically small, we get

$$C_{gs} = \frac{\text{Im}(Y_{11} + Y_{12})}{\omega} \tag{9.61}$$

Equation (9.55) can be rewritten as

$$R_{gb} - \frac{j}{\omega C_{gb}} = \frac{1}{Y_{11} - j\omega(C_{gd} + C_{gs})} \tag{9.62}$$

from which we can obtain

$$R_{gb} = \text{Re}\left[\frac{1}{Y_{11} - j\omega\left(C_{gd} + C_{gs}\right)}\right] \tag{9.63}$$

and

$$C_{gb} = -\frac{1}{\omega \text{Im}\left(\frac{1}{Y_{11} - j\omega(C_{gd} + C_{gs})}\right)} \tag{9.64}$$

Taking the real part of Y_{22} from (9.58), we get

$$\text{Re}(Y_{22}) = g_{dsb} + \frac{\omega^2 C_{db}^2 R_{db}}{1 + \omega^2 C_{db}^2 R_{db}^2} \tag{9.65}$$

which, upon taking the limit as ω approaches zero, gives

$$R_{dsb} = g_{dsb}^{-1} = \left.\frac{1}{\text{Re}\,(Y_{22})}\right|_{\omega \to 0} \tag{9.66}$$

at sufficiently low frequencies. Equation (9.65) can be rewritten as

$$R_{db} + \frac{1}{\omega^2 C_{db}^2 R_{db}} = \frac{1}{\text{Re}(Y_{22}) - g_{dsb}} \tag{9.67}$$

which yields, after taking the limit as ω approaches infinity,

$$R_{db} = \left.\frac{1}{\text{Re}\,(Y_{22}) - g_{dsb}}\right|_{\omega \to \infty} \tag{9.68}$$

which approximates R_{db} at very high frequencies. g_{dsb} from (9.66) can be used assuming that its low and high frequency values are approximately equal. C_{db} can now be estimated from (9.67) as

$$C_{db} = \frac{1}{\sqrt{\omega^2 R_{db}\left[\frac{1}{\mathrm{Re}(Y_{22})-g_{dsb}} - R_{db}\right]}} \tag{9.69}$$

upon substituting g_{dsb} and R_{db} from (9.66) and (9.68), respectively. The imaginary part of Y_{22} is given from (9.58) as

$$\mathrm{Im}(Y_{22}) = \omega\left(C_{gd} + C_{ds} + \frac{C_{db}}{1+\omega^2 C_{db}^2 R_{db}^2}\right) \tag{9.70}$$

from which, we can derive C_{ds} as

$$C_{ds} = \frac{\mathrm{Im}(Y_{22})}{\omega} - \frac{C_{db}}{1+\omega^2 C_{db}^2 R_{db}^2} - C_{gd} \tag{9.71}$$

A particularly useful instrument for measuring the DC parameters of devices is the "Semiconductor Parameter Analyzer." As for any measurement-based modeling or parameter-extraction techniques, it is essential to calibrate the complete measurement system including test fixture, onto which the device is mounted or on-wafer probes along with cables so that all external effects are removed from the device. External resistance coming from on-wafer probes, for instance, may introduce intolerable error to the device's measured resistances, particularly for those having small resistances. Low frequency oscillation due the device's instability may also occur and needs to be carefully suppressed using some means such as connecting large capacitors and/or resistors.

Extraction of Linear Model

The following steps are typical in determining a linear model:

1. Measure (or obtain from available sources) the small-signal *S*-parameters of actual MOSFET device over interested frequency band at desired operating bias condition (V_{ds}, V_{gs}, V_{bs}). For nonlinear model, multiple bias points are needed. The measurement can be carried out in a test fixture, on which the device is mounted, for discrete devices or on-wafer for on-chip devices. A de-embedding technique, involving calibration such as measurements of short, open, and through-line along with a de-embedding program as described in Chapter 15, are used to extract the device's *S*-parameters from those measured. The de-embedding must be done carefully to extract accurate *S*-parameters of the device. Otherwise, the fitting described in the Step 3 will not produce the best results. This is especially important in CMOS RFIC design since active devices are extremely small as compared to interconnects (such as on-wafer pads) needed for making measurements; any miscalibration will lead to significant error, particularly phase. It is especially noted that some *S*-parameters, such as S_{12}, are very sensitive to errors introduced by the test fixture or the on-wafer probe. When test fixtures are used, the inductance arises from the required bonding wires need to be kept as small as possible. Sufficiently low frequencies should be used to minimize the effect of external parasitic inductance, such as that from bonding wires, on those linear parameters which are computed from low frequency data. More accurate results for these parameters, which change only slightly with frequency, can thus be obtained. The de-embedding technique is discussed in Chapter 15 (CMOS RFIC Simulation, Layout and Test).
2. Measure (or estimate from available data such as device's IV curves and/or *S*-parameters and/or equations given in the previous sections) values of some elements of the model (R_g, R_s, R_d, R_{dsb}, R_{db}, R_{gb}, g_{mo}, g_{mbo}, C_{gd}, C_{gs}, C_{ds}, C_{gb}, C_{db}) at DC or interested frequency for the device interested and use them as initial guesses in the subsequent fitting process.

3. Use a commercially available or another available program to fit the measured *S*-parameters to those calculated from the linear equivalent-circuit model in Figure 9.10 and vary the model's parameters until a close match between the *S*-parameters is achieved. This process will result in all elements of the linear model including: L_g, L_d, L_s, R_g, R_d, R_s, C_{gs}, C_{gd}, C_{ds}, R_{gb}, R_{db}, R_{sb} (independent of bias voltages); C_{gb}, C_{sb}, C_{db}, R_{dsb} (at applied bias conditions); g_{mo}, g_{mbo} (transconductance at DC); and transit time τ_g, τ_{gb}.

The foregoing process essentially produces a linear model for MOSFETs. As an example, a 0.18-μm CMOS transistor is modeled using the linear equivalent-circuit model shown in Figure 9.10(b) and the following measured *S*-parameters from 0.1 to 20 GHz:

Frequency (GHz)	S_{11} (Mag/Deg)	S_{12} (Mag/Deg)	S_{21} (Mag/Deg)	S_{22} (Mag/Deg)
0.100	1.000/−0.561	0.001/89.530	3.451/179.399	0.897/−0.796
0.500	1.000/−2.803	0.007/87.657	3.443/177.011	0.893/−3.950
1.000	0.999/−5.599	0.015/85.363	3.419/174.113	0.881/−7.714
2.000	0.994/−11.144	0.029/81.030	3.342/168.800	0.844/−14.236
3.000	0.988/−16.604	0.042/77.044	3.252/164.134	0.803/−19.357
4.000	0.980/−21.960	0.055/73.307	3.164/159.909	0.768/−23.489
5.000	0.971/−27.194	0.068/69.741	3.081/155.948	0.738/−27.055
6.000	0.960/−32.290	0.079/66.310	3.001/152.154	0.714/−30.312
7.000	0.948/−37.231	0.090/63.002	2.922/148.492	0.693/−33.389
8.000	0.936/−42.005	0.101/59.819	2.843/144.949	0.675/−36.338
9.000	0.923/−46.604	0.110/56.763	2.764/141.525	0.658/−39.182
10.00	0.909/−51.021	0.119/53.836	2.685/138.222	0.644/−41.924
11.00	0.896/−55.254	0.127/51.040	2.607/135.039	0.630/−44.565
12.00	0.883/−59.304	0.134/48.373	2.529/131.979	0.617/−47.103
13.00	0.870/−63.171	0.141/45.833	2.453/129.039	0.605/−49.540
14.00	0.857/−66.861	0.147/43.415	2.378/126.215	0.594/−51.873
15.00	0.845/−70.377	0.153/41.116	2.306/123.505	0.583/−54.107
16.00	0.833/−73.726	0.158/38.929	2.235/120.904	0.574/−56.243
17.00	0.822/−76.915	0.163/36.849	2.167/118.406	0.565/−58.285
18.00	0.812/−79.950	0.167/34.869	2.101/116.006	0.557/−60.236
19.00	0.802/−82.838	0.171/32.984	2.037/113.699	0.549/−62.101
20.00	0.793/−85.586	0.174/31.187	1.977/111.479	0.542/−63.885

Figure 9.19 shows the resultant linear model including all the element values for 0.1–20 GHz. Figure 9.20 displays the measured and modeled *S*-parameters from 0.1 to 20 GHz on the Smith charts. Figure 9.21 shows the relative errors between the measured and modeled magnitudes of *S*-parameters, which are calculated as

$$\Delta|S_{ij}| = \frac{|S_{ij}^m| - |S_{ij}^{\text{meas}}|}{|S_{ij}^{\text{meas}}|} \tag{9.72}$$

where S_{ij}^m and S_{ij}^{meas} are the modeled and measured $S_{ij}(i, j = 1, 2)$ parameters. As can be seen, the modeled results match well with those measured with errors between *S*-parameters within 1–1.5% across 0.1–20 GHz. The frequency f_T and $f_{\max}$ of the device can be estimated from (9.79) and (9.84) as 52 and 126 GHz, respectively. The obtained f_T and $f_{\max}$ are reasonable for the considered device. However, it should be noted that these frequencies may not be very accurate because the values of the parameters used in the expressions are obtained through fitting and thus may not represent accurately the device's physics.

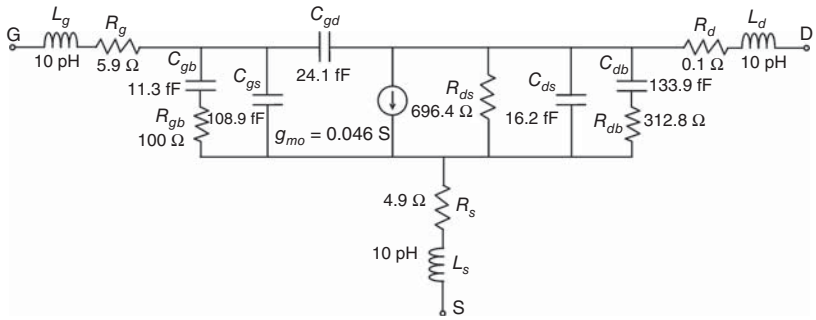


Figure 9.19. A linear model of a 0.18-μm CMOS transistor from 0.1 to 20 GHz.

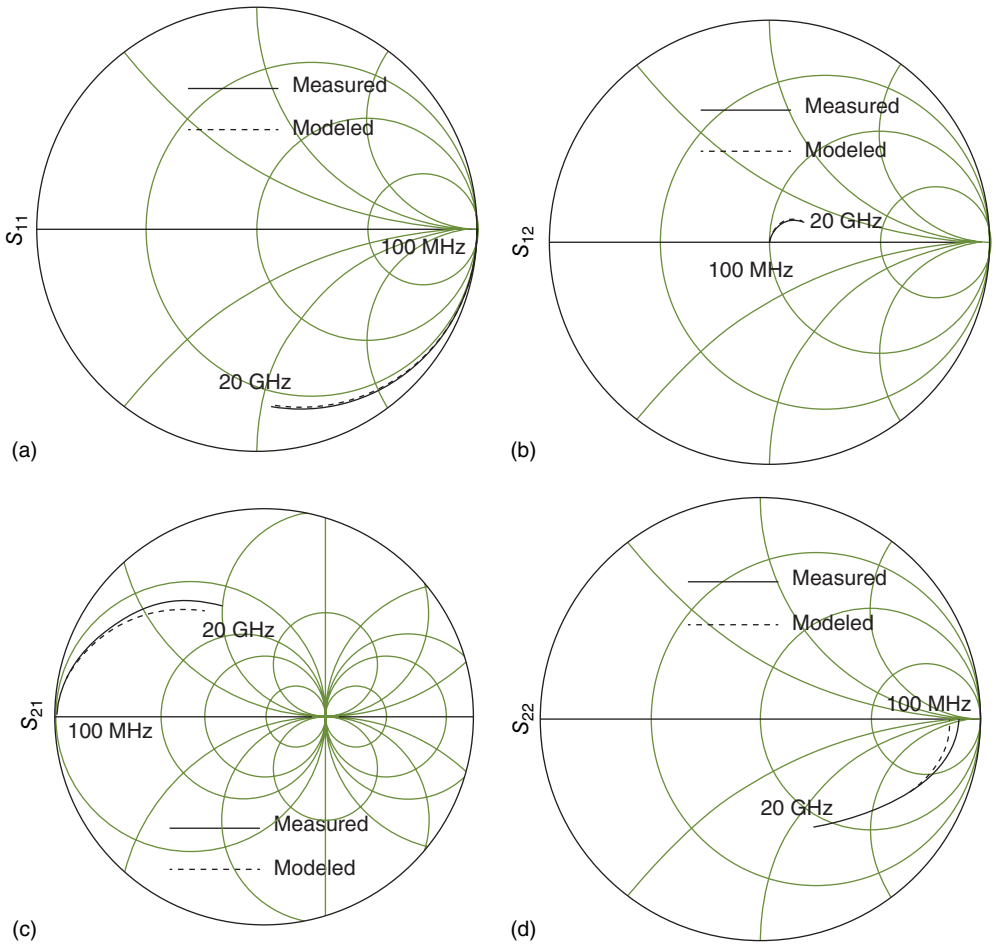


Figure 9.20. Measured and modeled *S*-parameters of a 0.18-μm CMOS transistor from 0.1 to 20 GHz: (a) S_{11}, (b) S_{12}, (c) S_{21}, and (d) S_{22}.

Determination of Nonlinear Elements

The nonlinear elements in the equivalent-circuit model are $C_{sb}(V_{sb})$, $C_{db}(V_{db})$, $I_{ds}(V_{ds}, V_{gs})$, and $I_{dsb}(V_{ds}, V_{bs})$. Note that the drain–source current I_{ds} comprises both g_{mo} and R_{ds}. Besides measured *S*-parameters at different bias conditions, nonlinear modeling typically requires measurement of the device's I–V curves.

The capacitance C_{sb} and C_{db} and output resistance R_{ds} can be obtained from various equations such as (9.25), (9.26), and (9.35), respectively, or other equations as described earlier. The current $I_{ds}(V_{ds}, V_{gs})$

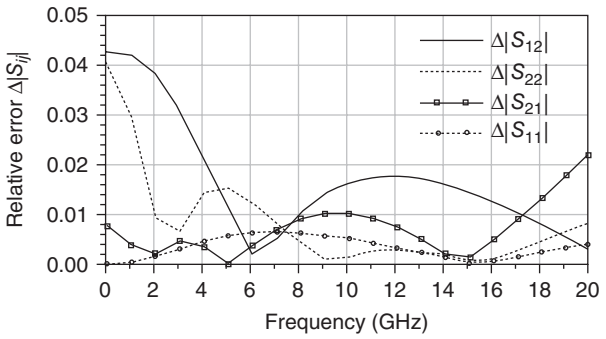


Figure 9.21. Relative errors between the measured and modeled *S*-parameters.

and $I_{dsb}(V_{ds}, V_{bs})$ can be determined using equations derived earlier. While those equations are useful in understanding the capacitance, resistance, and current behaviors as a function of bias voltages, they may not be accurate for circuit design, particularly for submicron MOSFET devices. More accurate equations may therefore need to be derived. Equations describing accurately the behavior of C_{sb}, C_{db}, and R_{ds} (hence R_{dsb}) may be derived by examining the capacitance and resistance data obtained from the *S*-parameters fitting process at different bias conditions, as outlined in Step 3 of the foregoing linear element determination procedure. The principle behind the derivation of accurate equations for the current I_{ds} or I_{dsb} is that any equation for them can be used as long as it describes closely the behavior of the actual current of the device versus bias voltages. Such an equation may be derived by fitting the DC I–V curves measured for a particular device to those obtained from that equation. Some of these equations may be adopted from those developed for GaAs field effect transistors (FETs) used in microwave circuits such as the Curtice and Ettenberg model [4].

Following the Curtice and Ettenberg model, the current I_{ds} can be expressed as

$$I_{ds} = (A_o + A_1 V + A_2 V^2 + A_3 V^3) \tanh[r V_{ds}(t)] \tag{9.73}$$

where

$$V = V_{gs}(t - \tau_g)\{1 + u[V_{dso} - V_{ds}(t)]\} \tag{9.74}$$

with A_n $(n = 0, 1, 2, 3)$ being the coefficients of the third-order polynomial, r being the hyperbolic tangent function parameter, u being the coefficient describing the pinch-off voltage change with V_{ds}, τ_g being the time delay between I_{ds} and V_{gs}, V_{dso} being the drain–source voltage at which the constants A_n are evaluated. A_n, r, and u are obtained by fitting the DC I–V curves measured to those obtained from Eqs. (9.73) and (9.74) until a close agreement between them at various bias conditions is achieved.

Similarly, I_{dsb} may also be expressed by an expression that describes closely the actual behavior of I_{dsb} versus V_{ds} and V_{bs}, whose parameters can be obtained by fitting the measured I_{dsb} to I_{dsb} calculated from its expression.

Figure 9.22 shows a diagram describing steps used in extracting the linear and nonlinear equivalent-circuit models for MOSFETs operating in the RF region.

9.2.3 SPICE Models

The most popular models used for CMOS RFICs, particular in the low RF frequency region, are perhaps the models used in the SPICE computer program developed at University of California, Berkeley, USA. There are various variations of SPICE models for MOSFETs, including Level 1, 2 and 3, BSIM3 and BSIM4 (Berkeley Short-Channel IGFET Model 4). The principle difference in these models is the modeled I–V characteristics, which result in different levels of accuracy and computational efficiency. Details of these models can be found in various sources, for example, [5–7]. It is noted that these models evolve with time and so different BSIM models could exist.

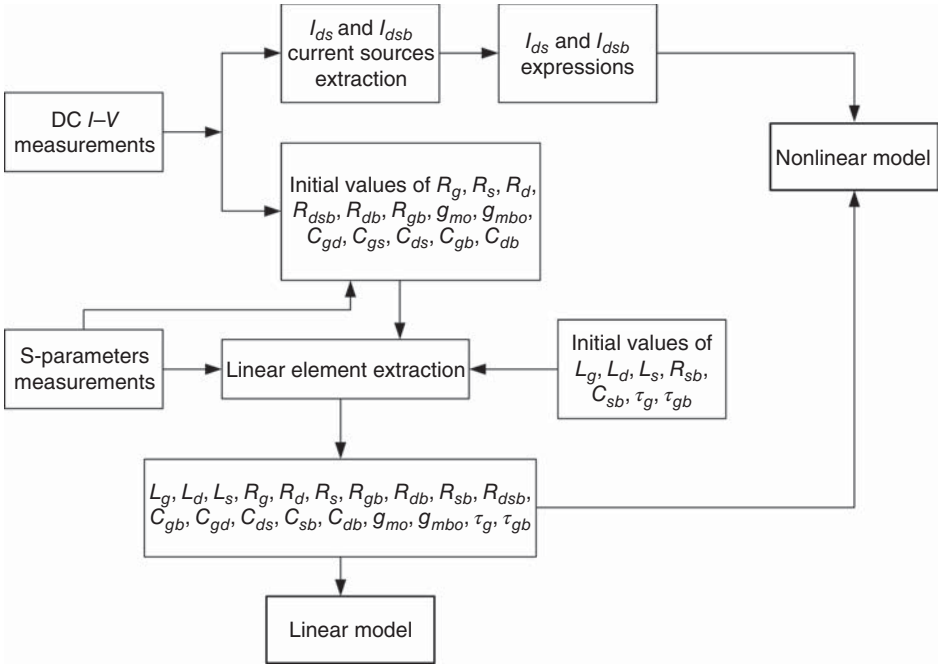


Figure 9.22. A procedure for the extraction of MOSFET linear and nonlinear equivalent-circuit models.

Level-1 Model

Perhaps the simplest model for MOSFETs is the Level-1 model, proposed by Shichman and Hodges [5]. This model is based on the square-law current expansion for I_{ds} and is suitable for long-channel devices. It does not consider short-channel effects, such as velocity saturation and mobility degradation, seen in typical MOSFETs used in RFICs, and is thus not sufficiently accurate. Level-1 model is therefore not very useful and hence seldom used for CMOS RFIC simulations. However, it can describe relatively accurate the MOSFET's switching behavior at different gate bias voltages for various drain bias voltages, through its modeled gate–drain capacitance, and is thus can be useful for simple and quick analysis of CMOS RFIC switches. Furthermore, although the Level-1 model does not provide accurate results, it is simple and is thus very useful for understanding the operation of MOSFETs. Figure 9.23 shows the Level-1 model of NMOS FETs. The drain–source current I_{ds} flowing in the conducting channel between the drain and source is modulated by the longitudinal electric field resulting from the voltages applied to the gate and drain. This current is modeled and implemented in SPICE as

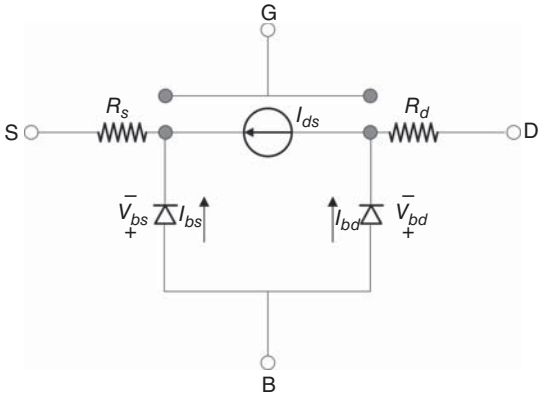


Figure 9.23. Level-1 model of NMOS FETs.

$$I_{ds} = \begin{cases} 0, & V_{gs} \le V_t \quad \text{(cutoff region)} \\ K\frac{W}{L-2L_{\text{diff}}}\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right)V_{ds}(1+pV_{ds}), & V_{gs} \ge V_t \quad \text{and } V_{ds} \le V_{gs} - V_t \quad \text{(linear region)} \\ \frac{K}{2}\frac{W}{L-2L_{\text{diff}}}(V_{gs} - V_t)^2(1+pV_{ds}), & V_{gs} \ge V_t \quad \text{and } V_{ds} \ge V_{gs} - V_t \quad \text{(saturation region)} \end{cases} \tag{9.75}$$

where

$$V_t = V_{to} + g(\sqrt{\phi - V_{bs}} - \sqrt{\phi}) \tag{9.76}$$

L and W are the gate length and width (m), respectively. $(V_{gs} - V)_t$ is equal to the drain–source saturation voltage V_{dsat} discussed earlier. Table 9.3 describes the other parameters appearing in Eqs. (9.75) and (9.76) and the model, shown in Figure 9.21, along with their example values for a 0.5-μm NMOSFET. Additional parameters are required to simulate the SPICE Level-1 model and can be found in various sources such as the SPICE program and Agilent ADS (Advanced Design System) program [8].

BSIM3 Model

BSIM3 model is a physics-based MOSFET model that was developed by the BSIM Research Group of the Department of Electrical and Computer Science Engineering at the University of California, Berkeley, USA. The third version of the BSIM3 model, BSIM3 Version 3, commonly abbreviated as BSIM3v3, was established by SEMATECH as the first industry-wide standard of its kind in December 1996. Since then, BSIM3v3 has been widely used in industry for CMOS device modeling and IC design [7]. The BSIM3 model takes into account the effects of device geometry and process parameters in MOSFET devices and thus can predict accurately the device performance, particularly in the low RF frequency range. In the high RF end such as high microwave and millimeter-wave regimes, measurement results of MOSFETs need to be incorporated into the BSIM3 model to obtain more accurate models. The latest version of BSIM3 was released on July 29, 2005. Details of BSIM3 model can be found in [7]. BSIM3 model parameters for certain MOSFETs can be found in the websites of various CMOS foundries.

It should be noted that MOSFET modeling is a continuous activity and different models are expected to be developed and made available for CMOS RFIC design as time progresses.

9.2.4 Passive MOSFET Models

Most MOSFET models, such as those foregoing ones, are for MOSFETs operating in active modes like amplification and mixing, They are thus can be referred to as "active MOSFET models." In applications involving non-amplification functions such as switching or phase shifting, the MOSFETs employed in RFICs are operated as passive devices. In these circuits, "passive MOSFET models" are accordingly needed. Passive models

TABLE 9.3. Some Level-1 Model Parameters and Example Values for a 0.5-μm NMOSFET

Parameter symbol	Parameter name	Unit	Example value
V_{to}	Threshold voltage at zero substrate bias ($V_{bs} = 0$)	V	0.7
K	Transconductance parameter	A/V^2	2×10^{-5}
p	Channel-length modulation parameter	V^{-1}	0.1
Q	Body-effect parameter	$V^{1/2}$	0.4
R_d	Drain ohmic resistance	Ω	1.5
R_s	Source ohmic resistance	Ω	1.5
ϕ	Surface potential	V	0.9
L_{diff}	Lateral diffusion length (overlapping length at source and drain)	m	0.08×10^{-6}

can be obtained directly using the models provided by a foundry with proper bias conditions. They can also be developed by CMOS RFIC designers, particularly when the desired operating frequencies are beyond those of the available foundry models.

Passive MOSFETs are operated under two conditions: *on* and *off*. When the gate–source voltage V_{gs} is higher than the threshold voltage V_t, the transistor is (theoretically) turned *on*, under which a small resistance R_{on} appears between the drain and source, assuming negligible parasitics. V_{gs} is normally set to V_{dd} in practical circuits for on-state. R_{on} is basically the (output) drain–source resistance R_{ds}, which approximately constitutes a simple "on-model" between the drain and source as shown in Figure 9.24(a). On the other hand, when V_{gs} is lower than V_t, the transistor is (theoretically) switched *off*. Typically, V_{gs} is set to 0 V in practical circuits for off-state. It should be noted that the off-state occurs only well at low frequencies. At high frequencies, the parasitic drain–source capacitance, C_{ds}, provides a path for signals from drain to source, hence degrading the off-state condition of the device. A simple off-model between the drain and source can be approximately represented by equivalent-circuit models shown in Figure 9.24(b) and (c), which are dominated by C_{ds}. The off-capacitance C_{off} and off-resistance R_{off} in Figure 9.24(c) can be derived from Figure 9.24(b) as

$$C_{\text{off}} \simeq C_{ds} + \frac{1}{2}C_{gs} \tag{9.77}$$

assuming $C_{gs} \simeq C_{gd}$ and

$$R_{\text{off}} \simeq R_{ds,\text{off}} \tag{9.78}$$

where $C_{ds,\text{off}}$ is C_{ds} under off-state condition. Passive MOSFETs are operated between the drain (input or output) and source (output or input) with a control voltage applied to the gate terminal through a typically large resistor, thus acting essentially as a two-terminal device. There is no DC power consumption in passive MOSFETs since there is no current flow at the gate.

More accurate on- and off-model for passive MOSFETs can be described in an equivalent circuit as shown in Figure 9.25. These on- and off-model can be obtained by fitting the *S*-parameters of actual devices measured under on and off bias conditions, respectively, over interested frequencies to the *S*-parameters calculated from the equivalent-circuit model. The *S*-parameters for passive MOSFETs operating under on and off conditions can also be obtained directly from a large-signal model by setting proper values for V_{gs} – for example, 1.8 and 0 V for on and off, respectively, from which on and off equivalent-circuit models can be extracted.

MOSFETs operating in passive mode with the bulk open or floated is especially attractive for switching and phase shifting involving high power, such as transmit–receive (T–R) switches for RF transceivers, as bulk floating can improve the linearity and power handling of the circuits, as mentioned earlier in Section 9.1.1. Figure 9.26 shows small-signal equivalent-circuit models for the on and off states of floating-bulk MOSFETs

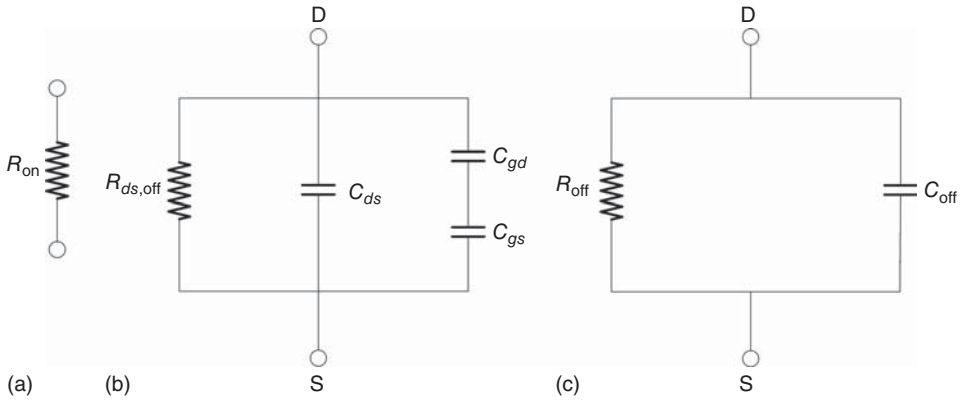


Figure 9.24. Simple small-signal *on* (a) and *off* (b, c) models for MOSFETs operating in passive mode. R_{on}, C_{off}, and R_{off} should be as small and large as possible, respectively.

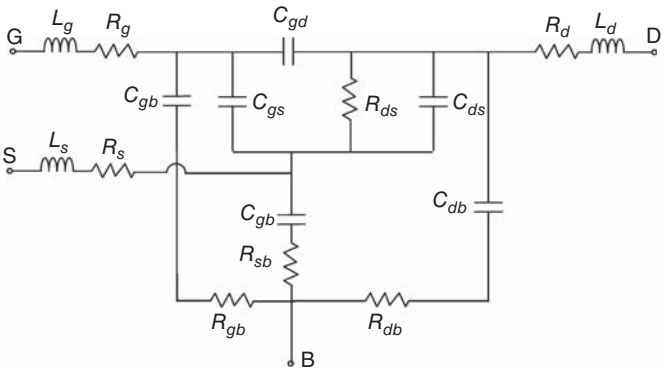


Figure 9.25. An equivalent-circuit for the on- and off-model of passive MOSFETs.

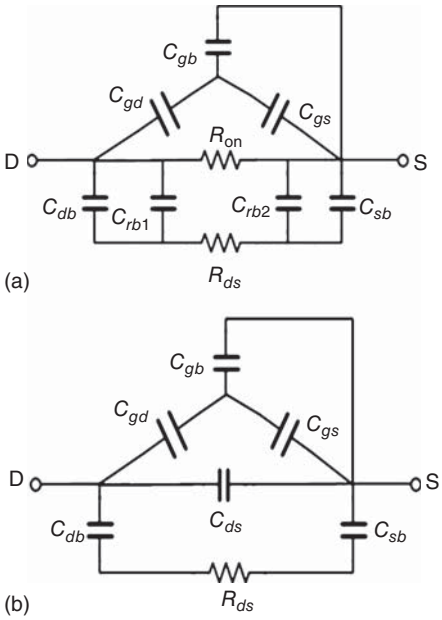


Figure 9.26. Small-signal equivalent-circuit models for MOSFETs with both the bulk and gate floated under on-state (a) and off-state (b).

when the gate is floated using a large resistor.[3] R_{ds} represents the resistive loss in the bulk between the source and drain as mentioned before. Using large gate widths for advanced CMOS devices can produce R_{ds} within several ohms, thereby resulting in low loss in the bulk. C_{gd} and C_{gs} again represent the gate–drain and gate–source capacitances due to the overlapping between the gate and diffusion areas. C_{gb} represents the gate-bulk capacitance. C_{db} and C_{sb} are the junction capacitances between the drain-bulk and source-bulk, respectively. C_{ds} is the drain–source capacitance. These capacitances are described earlier. C_{rb1} and C_{rb2} in Figure 9.26(a) represent the distributed capacitances between the inversion layer and bulk. All these parasitic capacitances are on the order of tens of femtofarad for advanced CMOS devices and increase with the device's gate width.

Figure 9.27(a) and (b) show simplified small-signal equivalent-circuit on- and off-model of MOSFETs deduced from Figure 9.26(a) and (b), respectively. The on-model consists of the on-resistance R_{on} in parallel with the on-capacitance C_{on}, which represents the total capacitance C_g seen at the gate consisting of C_{gb}, C_{gd}, and C_{gs} in Figure 9.26(a), the on-state bulk capacitance $C_{b\text{-on}}$ consisting of C_{db}, C_{sb}, C_{rb1}, and C_{rb2} in

[3]In typical passive-mode applications such as switching, the gates of MOSFETs are connected with a large resistor to allow gate bias to be applied without disturbing the RF performance.

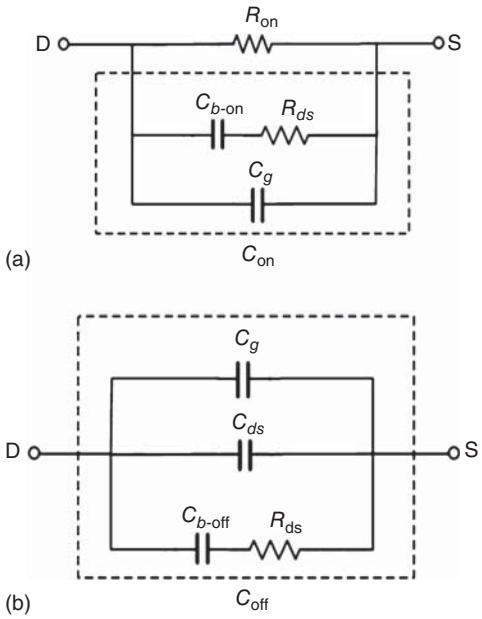


Figure 9.27. Simplified small-signal equivalent-circuit models of the bulk-floated MOSFETs under on (a) and off (b) conditions.

Figure 9.26(a), and the drain–source resistance R_{ds}. The off-model is represented by the off-capacitor C_{off}, which consists of the total capacitance C_g seen at the gate consisting of C_{gb}, C_{gd}, and C_{gs} in Figure 9.26(b), the off-state bulk capacitance $C_{b\text{-off}}$ consisting of C_{db} and C_{sb} in Figure 9.26(b), the drain–source capacitance C_{ds}, and the drain–source resistance R_{ds}.

An important remark needs to be made at this point. In three-terminal solid-state devices such as the metal–semiconductor field-effect transistor (MESFET) used in microwave circuits, the supplied bias voltage V_{gs} used to turn on and off a device is applied across the gate and source terminals of the device. The MOSFET used in CMOS RFIC, however, has four terminals with the source and bulk terminals separate, and the supply of its bias voltages may sometimes be confused. In most circuits, the bulk is grounded and, in typical design, the source is also connected to the bulk. Under this configuration, the bias voltage V_{gs} provided by a voltage supply is applied across the gate and source terminals and the MOSFET is operated properly. However, for MOSFET in certain circuits where the bulk is floating (for instance, with a large resistor connecting to ground) and the source is not tied to the bulk, such as a series MOSFET in switches designed for improved linearity and power handling, the supplied V_{gs} does not appear across the gate and source terminals as expected since the source is not DC-grounded, causing the device and, hence, the circuit not working properly. For instance, assuming the device needs 2-V V_{gs} to turn it on, when 2 V is applied to the gate, the actual voltage across the gate and source terminals is not 2 V; instead, it varies between 0 and 2 V, thus causing the device not being turned on properly. In fact, the device may be on, off or in between when $V_{gs} = 2$ V is applied. It is noted, however, that when 0-V V_{gs} is applied to the gate, the voltage appearing across the gate and source terminals, although undetermined, is always less than or equal to 0 V, thus turning the device off as desired. Under the bulk-floating operating condition with the source being separate from the bulk, the source needs to be DC-grounded to allow correct bias voltage V_{gs} across the gate and source terminals for proper device and circuit operation – for instance, using a large resistor between the source and ground.

9.3 IMPORTANT MOSFET FRQUENCIES

The two most important frequencies characterizing the performance of MOSFETs are f_T, the cut-off frequency or frequency of unity current-gain, and $f_{\max}$, the maximum frequency of oscillation or frequency of unity power-gain.

9.3.1 f_T

The frequency of unity current-gain f_T is approximately given by

$$f_T \simeq \frac{g_{mo}}{2\pi(C_{gs} + C_{gd} + C_{gb})} \tag{9.79}$$

f_T can be derived for different operating regions of MOSFETs as follows.

9.3.1.1 Saturation Region. Making use of the capacitances corresponding to the saturation region given in Table 9.1 and Eqs. (9.9), (9.20), (9.22), and (9.27) yields

$$f_T = \frac{1}{2\pi} \frac{\mu_n C_{\text{ox}} \frac{W}{L}(V_{gs} - V_t)}{2C_{\text{ox}} W \left(L_{\text{diff}} + \frac{1}{3}L\right)} \tag{9.80}$$

which becomes, assuming $L_{\text{diff}} \ll L$,

$$f_T \simeq \frac{3}{4\pi} \frac{\mu_n (V_{gs} - V_t)}{L^2} \tag{9.81}$$

Equation (9.81) shows that f_T is proportional to $1/L^2$ and increases as V_{gs} is increased (in the saturation region). It should be noted that this is valid only in the long-channel regime and the saturation region is typically used for amplification applications.

9.3.1.2 Velocity-Saturation Region. The velocity-saturation region results from the device's short-channel effect. Assuming that the short-channel effect does not change much the total capacitance ($C_{gs} + C_{gd} + C_{gb}$) that existed for the long channel, we can derive, using the capacitances from Table 9.1 and Eqs. (9.22) and (9.27) neglecting the overlapping gate length L_{diff},

$$C_{gs} + C_{gd} + C_{gb} \simeq \frac{2}{3} C_{\text{ox}} WL \tag{9.82}$$

Substituting Eqs. (9.16) and (9.82) into (9.79) and using (9.15) then yields

$$f_T \simeq \frac{3}{4\pi} \frac{v_{\text{sat}}}{L} = \frac{3}{4\pi} \frac{\mu_n E_{\text{sat}}}{L} \tag{9.83}$$

which indicates that f_T is inversely proportional to the actual physical length L of the gate when the velocity is saturated. It is noted that f_T (corresponding to saturated velocity) does not depend on the bias conditions. μ_n depends on the doping impurity concentration, material, and temperature – for example, μ_n for intrinsic silicon at 300 K is 1350 $\text{cm}^2/\text{V-s}$, while equals to 700 $\text{cm}^2/\text{V-s}$ for silicon with a doping concentration of 10^{17} cm^{-3}. $v_{\text{sat}} \simeq 10^5 \text{m/s}$.

The forgoing equations for f_T are derived assuming a common source MOSFET with its drain terminal short-circuited and gate driven by an ideal current source. Hence, the drain-bulk capacitance C_{db} is not included and the (series) gate resistance R_g is not considered. Also, the contribution of the gate–drain capacitance C_{gd} to the output current is neglected. Therefore, the values of f_T calculated from Eqs. (9.81) and (9.83) are only approximate. The actual value for f_T is about 1/3 to 1/2 of these calculated values.

9.3.2 f_{max}

The frequency of unity power-gain f_{max} is the frequency at which the maximum available gain is 1 or 0 dB and can be approximately derived as

$$f_{\text{max}} \simeq \frac{1}{2\sqrt{2\pi}} \sqrt{\frac{f_T}{R_g C_{gd}}} \tag{9.84}$$

The gate resistance R_g can be made small with proper device layout. Hence, f_{max} can be substantially greater than f_T for many MOSFETs. Both f_T and f_{max} depend on bias conditions and device geometry – for example, number of gate fingers, gate width, and gate length. As can be expected from various assumptions used in deriving Eq. (9.84) for f_{max}, values of f_{max} calculated from this equation are only approximate, which imply, for example, that a circuit using MOSFETs having a calculated f_{max} of 50 GHz can unlikely operate at 50 GHz.

9.4 OTHER IMPORTANT MOSFET PARAMETERS

Other important MOSFET parameters include maximum available gain at frequencies of interest, unilateral gain (power gain with no reverse transmission or practically very little reverse transmission), minimum noise figure, power added efficiency measured as the ratio between the difference of the output and input RF power and the DC power consumption, output power at 1-dB compression point normally considered the maximum output power, gain corresponding to minimum noise figure, and stability. It is noted the stability of MOSFETs depends not only on the device itself, but also on the circuit environment in which the device is embedded (i.e., the impedances presented to the device's terminals.)

9.5 VARACTOR DIODES

A varactor (VARiable reACTOR) diode, or simply varactor, is a variable-reactance semiconductor device, which functions primarily as a voltage-controlled capacitor. The variable reactance is provided by the junction capacitance that varies nonlinearly as a function of the applied reverse-biased voltage. This nonlinear feature is useful for various RF devices such as voltage-controlled oscillators (VCOs), continuous (analog) phase shifters, modulators, frequency multipliers, harmonic generators, parametric amplifiers, parametric oscillators, active filters, and up-converters.

9.5.1 Varactor Structure and Operation

Varactors in CMOS processes may be formed by placing a heavily positively doped region (p+) and negatively doped region (n+) within a lightly negatively doped region (n-well) in a p-substrate, essentially creating a p–n junction, as shown in Figure 9.28.

Varactor diodes are usually operated under reverse-biased conditions, under which they behave as a voltage-controlled capacitor. As the reverse-biased voltage is increased, the width of the depletion layer increases toward the semiconducting substrate, resulting in a reduction of the capacitance from a maximum value at 0 V to a minimum value at the pinch-off (or punch-through) voltage. Beyond the pinch-off level, the capacitance is essentially independent of the applied voltage. To obtain a large capacitance variation, which is needed for applications such as wide-band VCO's, the rate of change of capacitance with bias voltage should be high.

Another type of CMOS varactors is MOS varactor formed by connecting the drain, source, and bulk of a MOSFET to each other. Its capacitance depends on the voltage between the gate and bulk. The layout of MOS varactors is essentially the same as that of a typical MOSFET.

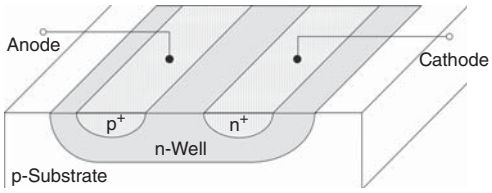


Figure 9.28. CMOS varactor diode structure.

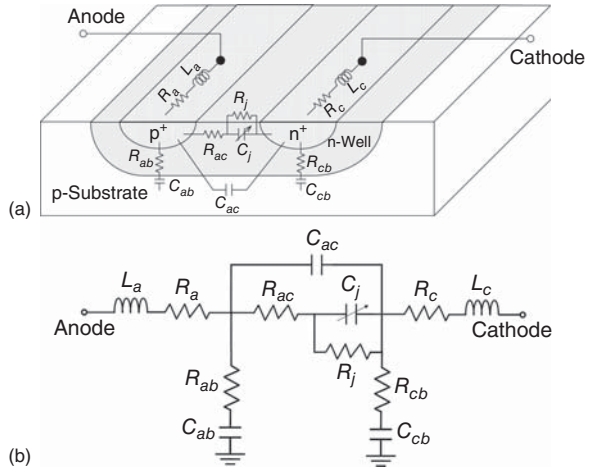


Figure 9.29. A model of varactors embedded in the varactor structure (a) and in equivalent circuit (b).

9.5.2 Varactor Model and Characteristics

Figure 9.29(a) illustrates the circuit parameters of varactors corresponding to the different regions of varactors and Figure 9.29(b) shows the corresponding equivalent-circuit model under reverse-bias operations.

L_a and L_c represent the anode and cathode electrode inductance, respectively. R_a and R_c are the anode and cathode terminal resistance. These resistances are essentially due to the conductivity of the terminal material and the contact. R_{ac} is the resistance accounting for the loss due primarily to the n-well material, which is the dominant resistance as compared to R_a and R_c. C_{ac} is the anode–cathode capacitance, accounting for the coupling between the anode and cathode due to the lossy Si substrate. R_j is the junction resistance, which is normally neglected due to its large value compared with the junction capacitive reactance at RF frequencies. R_{ab} and R_{cb} are the resistances due to the n-well and substrate between the anode and bulk and between the cathode and bulk, respectively.

C_j is the diode junction capacitance given approximately as

$$C_j(V_r) \simeq C_j(0)\left(1 - \frac{V_r}{\phi_b}\right)^{-m} \tag{9.85}$$

where $C_j(0)$ is the zero-voltage junction capacitance, which can be determined from the device physics or measurement; V_r is the reverse-biased voltage applied across the junction; ϕ_b is the junction's built-in or contact potential, typically around several tenths of a volt; and m is a parameter depending primarily on the doping profile and equal to 1/3 and 1/2 for the respective linearly graded and abrupt junctions, respectively, and greater than 1/2 for hyperabrupt junctions. m can also be obtained as the negative of the slope of $\log[C_j(V_r)/C_j(0)]$ versus $\log(1 - V_r/\phi_b)$.

A simplified equivalent-circuit model for varactors can be devised from Figure 9.29 as shown in Figure 9.30. In this model, R_s represents the total resistance contributed from R_a, R_{ac}, and R_c; L_s accounts for L_a and L_c; and R_{sub} and C_{sub} combine R_{ab}, R_{cb} and C_{ab}, C_{cb}, respectively. The diode series resistance, R_s, is normally determined by extrapolating the measured I–V characteristics. This measurement, however, is not accurate due to thermal effects in the diode junction. These effects can be accounted for by performing measurements at 10 MHz or higher using the technique described in [9]. At high RF frequencies, particularly millimeter-wave frequencies, R_s increases substantially due to the skin effect. This effect can be approximately taken in account by using the following equations [9]:

$$R_s = \begin{cases} R_{\text{dc}} + \dfrac{R_{\text{dc}}\sqrt{f/f_s}}{2\left(2-\sqrt{f/f_s}\right)}(1+2j), & f \ll f_s \\ R_{\text{dc}}\sqrt{\dfrac{f}{f_s}}(1+j), & f \gg f_s \end{cases} \tag{9.86}$$

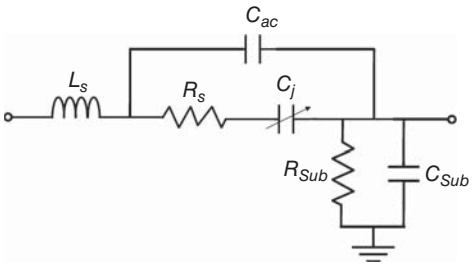


Figure 9.30. A simplified equivalent-circuit model of varactors.

where $R_{\rm dc}$ is the low frequency series resistance. f_s is a function of the metal resistivity and geometry, and can be determined from known values of $R_{\rm dc}$ and R_s at any frequency.

As for the lumped-element equivalent-circuit models of MOSFETs discussed earlier, all the values of the varactor model's parameters can be obtained by fitting measured *S*-parameter data of actual devices to those calculated from the equivalent-circuit model until a close match between them is achieved. The model obtained in this way would produce closely the actual performance of the measured devices.

As can be seen in Eq. 9.85, various capacitance ranges can be optimized for specific applications by using the different doping profiles, provided that these profiles can be implemented in CMOS technology. For a given voltage range, varactors of increasing capacitance range are the linearly graded, abrupt, and hyperabrupt junctions. When the applied reverse-biased voltage is much greater than the built-in potential, the junction capacitance $C_j \propto V_r^{-m}$. Under this bias condition and when the doping profile is designed such that m is equal to 2 (a hyperabrupt junction), $C_j \propto 1/V_r^2$. The resonant frequency of LC resonators with C simulated by such a varactor is

$$\omega_r = \frac{1}{\sqrt{LC_j}} \propto V_r \tag{9.87}$$

which varies linearly with the applied reverse-biased voltage. This result is useful for obtaining increased resonant frequency for LC resonators.

It is noted that important device parameters for varactors are their series resistance, R_s (as shown in Figure 9.30), zero-bias junction capacitance, $C_j(0)$, and capacitance at the pinch-off voltage, C_{jp}. R_s can be reduced by increasing the junction area and doping concentration while reducing the epitaxial layer thickness. $C_j(0)$ and C_{jp} are related to the semiconductor dielectric constant, the diode area and, most importantly, the depletion region thickness. To achieve a minimum value for $C_j(0)$, the diode area needs to be increased. C_{jp}, together with $C_j(0)$, determines the range of the capacitance variation, which is the most important characteristic of varactor diodes, besides the quality factor. In general, a small value for R_s and C_{jp} and a large value for $C_{j0}(0)/C_{jp}$ are desired.

Other important parameters of varactor diodes [10, 11] for circuit design are the cutoff frequency $f_c(V_r)$ at the junction reverse-biased voltage V_r

$$f_c(V_r) = \frac{1}{2\pi R_s C_j(V_r)} \tag{9.88}$$

the quality factor $Q(V_r)$ at V_r and frequency f

$$Q(V_r) = \frac{f_c(V_r)}{f} \tag{9.89}$$

the dynamic cutoff frequency f_{cd}

$$f_{cd} = \frac{1}{2\pi R_s}\left(\frac{1}{C_{j\min}} - \frac{1}{C_{j\max}}\right) \tag{9.90}$$

and the dynamic quality factor Q_d at frequency f

$$Q_d \simeq n\frac{C_j(0)}{C_j(V_r)}\frac{f_c(0)}{f} \tag{9.91}$$

where

$$n = \frac{C_1}{C_0} = \begin{cases} 0.25 & \text{for abrupt junction} \\ 0.17 & \text{for graded junction} \end{cases} \tag{9.92}$$

and $C_j(V_r)$ is the junction capacitance at V_r; $C_{j\min}$ and $C_{j\max}$ are usually the junction capacitances at the pinch-off and zero voltages, respectively; and C_0 and C_1 are the Fourier coefficients of the junction capacitance.

REFERENCES

1. B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, New York, New York, 2001.
2. J. Segura and C. F. Hawkins, *CMOS Electronics – How It Works, How It Fails*, IEEE Press and Wiley-Interscience, New York, New York, 2004.
3. A. F. Tong, K. S. Yeo, L. Jia, C. Q. Geng, J. G. Ma and M. A. Do, "Simple and Accurate Extraction Methodology for RF MOSFET Valid up to 20 GHz," *IEE Proc.-Circuits Devices Syst.*, Vol. 151, No. 6, pp. 587–592, Dec. 2004.
4. W. R. Curtice and M. Ettenberg, "A Nonlinear GaAs FET Model for Use in the Design of Output Circuits for Power Amplifiers," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 12, pp. 1383–1394, Dec. 1985.
5. H. Shichman and D. A. Hodges, "Modeling and Simulation of Insulated-Gate Field-Effect Transistor Switching Circuits," *IEEE J. Solid-St. Circ.*, Vol. SC-3, pp. 285–289, Sept. 1968.
6. G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, New York, 1993.
7. http://www-device.eecs.berkeley.edu/bsim/
8. Advanced Design System (ADS), Agilent, Palo Alto, California, USA.
9. D. N. Held, "Analysis of Room Temperature Millimeter-Wave Mixers Using Gallium-Arsenide Schottky Barrier Diodes," *Ph.D. Dissertation*, Columbia University, New York, 1977.
10. J. C. Irvin et al., in *Microwave Semiconductor Devices and Their Circuit Applications* (H. A. Watson, ed.), pp. 149–193, McGraw-Hill, New York, 1969.
11. J. W. Archer and R. A. Batchelor, "Multipliers and Parametric Devices," in *Handbook of Microwave and Optical Components*, Vol. 2, Microwave Solid-State Components (K. Chang, ed.), pp. 142–191, John Wiley & Sons, Inc., New York, 1990.

PROBLEMS

9.1 Consider any submicron MOSFET (e.g., 130 nm) available from a CMOS foundry with measured *S*-parameters at particular bias voltages V_{gs} and V_{ds} over some frequency ranges (e.g., 0.1–20 GHz). Assume both the internal source and the bulk are tied together and connected to ground. The device can be modeled as the equivalent circuit shown in Figure 9.10(b) under small-signal conditions. Use ports 1 and 2 for the gate and drain terminals, respectively, with the source and bulk connected to ground. *S*-parameters calculated from a nonlinear or small-signal model provided by the foundry can be used if measured *S*-parameters are not provided. If a device with *S*-parameters is not available then you can use the following measured *S*-parameters from 0.1 to 20 GHz for a 0.18-μm NMOS FET having total width of 10 μm and eight fingers (biased at $V_{ds} = 1.8$ V, $I_{ds} = 7.5$ mA, $V_{gs} = 0.7$ V).

Frequency (GHz)	S_{11} (Mag/Deg)	S_{12} (Mag/Deg)	S_{21} (Mag/Deg)	S_{22} (Mag/Deg)
0.1	1.000/−0.709	0.002/89.355	3.720/179.185	0.796/−1.068
0.6	0.999/−4.250	0.010/86.604	3.685/176.177	0.780/−3.794
1.1	0.997/−7.779	0.018/83.913	3.670/173.294	0.776/−6.218
1.6	0.993/−11.287	0.025/81.219	3.651/170.380	0.772/−8.746
2.1	0.989/−14.765	0.033/78.543	3.627/167.475	0.769/−11.295
2.6	0.983/−18.205	0.041/75.898	3.597/164.594	0.765/−13.833
3.1	0.976/−21.598	0.048/73.292	3.562/161.750	0.760/−16.343
3.6	0.968/−24.938	0.055/70.731	3.522/158.951	0.754/−18.813
4.1	0.960/−28.218	0.062/68.222	3.478/156.202	0.748/−21.237
4.6	0.950/−31.434	0.069/65.770	3.430/153.509	0.741/−23.608
5.1	0.940/−34.580	0.075/63.377	3.379/150.875	0.734/−25.922
5.6	0.930/−37.654	0.081/61.047	3.325/148.303	0.727/−28.177
6.1	0.919/−40.653	0.087/58.782	3.269/145.797	0.720/−30.370
6.6	0.908/−43.574	0.092/56.583	3.212/143.356	0.712/−32.499
7.1	0.897/−46.417	0.097/54.451	3.152/140.982	0.704/−34.564
7.6	0.886/−49.180	0.102/52.385	3.093/138.674	0.697/−36.565
8.1	0.874/−51.865	0.107/50.385	3.032/136.433	0.689/−38.503
8.6	0.863/−54.471	0.111/48.451	2.971/134.257	0.681/−40.377
9.1	0.852/−56.999	0.115/46.581	2.911/132.145	0.674/−42.190
9.6	0.841/−59.450	0.119/44.773	2.851/130.096	0.666/−43.942
10.1	0.830/−61.826	0.122/43.027	2.791/128.107	0.659/−45.637
10.6	0.820/−64.128	0.126/41.339	2.732/126.178	0.652/−47.275
11.1	0.810/−66.358	0.129/39.708	2.675/124.306	0.646/−48.858
11.6	0.800/−68.518	0.132/38.132	2.618/122.489	0.639/−50.390
12.1	0.790/−70.611	0.134/36.608	2.562/120.725	0.633/−51.871
12.6	0.781/−72.637	0.137/35.135	2.507/119.011	0.627/−53.304
13.1	0.772/−74.599	0.139/33.711	2.454/117.347	0.622/−54.691
13.6	0.763/−76.500	0.141/32.333	2.402/115.728	0.616/−56.034
14.1	0.754/−78.340	0.143/30.999	2.352/114.154	0.611/−57.335
14.6	0.746/−80.123	0.145/29.708	2.302/112.623	0.607/−58.597
15.1	0.738/−81.851	0.147/28.457	2.254/111.132	0.602/−59.820
15.6	0.731/−83.525	0.149/27.244	2.207/109.680	0.598/−61.007
16.1	0.724/−85.147	0.150/26.068	2.162/108.265	0.594/−62.160
16.6	0.717/−86.719	0.152/24.927	2.118/106.885	0.591/−63.280
17.1	0.710/−88.243	0.153/23.819	2.075/105.539	0.587/−64.369
17.6	0.704/−89.720	0.154/22.743	2.034/104.225	0.584/−65.429
18.1	0.698/−91.153	0.155/21.697	1.993/102.941	0.581/−66.460
18.6	0.692/−92.544	0.156/20.680	1.954/101.687	0.578/−67.464
19.1	0.687/−93.893	0.157/19.691	1.916/100.460	0.576/−68.443
19.6	0.681/−95.202	0.158/18.728	1.880/99.260	0.574/−69.397
20	0.677/−96.221	0.159/17.975	1.851/98.318	0.572/−70.144

a) Use a commercially available program to numerically determine the model's element values by fitting the measured *S*-parameters to the *S*-parameters calculated from the model. Initial values used for the fitting process can be estimated from equations provided in this chapter or from any other sources. Some of these initial values can also be estimated from the information available from the CMOS foundry of the considered device.

b) Plot the measured and modeled *S*-parameters versus frequency on Smith charts. Calculate and plot the errors between the measured *S*-parameters and those calculated from the obtained model versus frequency. Comment on the results.

c) Calculate f_T and $f_{\max}$ of the considered NMOS FET from the obtained model and comment on the results.

9.2 Various SPICE models have been developed and used for modeling NMOS and PMOS transistors. Consider the SPICE Level-2 model.

a) Describe the model in details including its equivalent-circuit representation.

b) Consider any 0.25-μm NMOS and PMOS FETs (or any other technology that is available in public domain) from any CMOS foundry. List values of the model's parameters for these devices. You need to provide definitions for the listed parameters. List the element values of the equivalent-circuit model for these devices. Include the devices' layouts if available.

9.3 Repeat Problem 9.2 for SPICE Level-3 model.

9.4 Repeat Problem 9.2 for BSIM3 model.

9.5 Repeat Problem 9.2 for BSIM4 model.

9.6 Consider a small-signal model for intrinsic MOSFETs, with the internal Source and Bulk tied together to the ground, as shown within the dashed lines of Figure 9.10(b). All the elements are considered linear.

a) Derive the admittance matrix $[Y]$ of the intrinsic model with ports 1 and 2 at the gate and drain, respectively.

b) Describe how approximate values of the intrinsic model's elements can be determined from the model's admittance matrix. Describe any assumptions that you use in arriving at the answers.

9.7 Consider a 0.18-μm MOSFET or any submicron MOSFET available from a CMOS foundry. You can also use the 0.18-μm CMOS transistor whose *S*-parameters are listed in Section 9.2.2.3. Estimate values for R_{dsb}, R_{db}, R_{gb}, g_m, C_{gd}, C_{gs}, C_{ds}, C_{gb}, C_{db} using Eqs. (9.59)–(9.61), (9.63), (9.64), (9.66), (9.68), (9.69), and (9.71) from its available *S*-parameters. State the frequencies at which you calculate these values.

9.8 Consider a 0.18-μm MOSFET or any submicron MOSFET available from a CMOS foundry with a nonlinear model. Estimate and plot $g_{mo,e}$ versus V_{gs} from 0 to 5 V for $V_{ds} = 0–4$ V in 0.5-V steps using Eq. (9.54) along with a commercially available computer program. You may need to increase the upper values for V_{gs} and/or V_{ds} suitable for the particularly used device to obtain results sufficient for a more complete examination of data. Examine and discuss the results.

9.9 Consider the same device in Problem 9.8. Extract R_s, R_d, and R_g using Eqs. (9.51)–(9.53) from its available *S*-parameters versus frequency for different values of V_{gs} greater (e.g., 2 V depending on the device) and smaller (e.g., 0.5 V depending on the device) than V_t and $V_{ds} = 0$ (i.e., considering both cases when the transistor is operated in linear and weak-inversion region, respectively.) Examine and discuss the results with respect to the bias condition and frequency.

9.10 Consider a 0.25-μm MOSFET or any submicron MOSFET available from a CMOS foundry with a nonlinear model. The output resistance R_{ds} and DC transconductances g_{mo} and g_{mbo} are given in Eqs. (9.32), (9.29), and (9.31) as

$$R_{ds}^{-1} = \left.\frac{\partial I_{ds}}{\partial V_{ds}}\right|_{V_{gs}=\text{constant}}$$

$$g_{mo} = \left.\frac{\partial I_{ds}}{\partial V_{gs}}\right|_{V_{ds}=\text{constant}}$$

and

$$g_{mbo} = \left.\frac{\partial I_{ds}}{\partial V_{bs}}\right|_{V_{ds}=\text{constant}}$$

a) Estimate and plot R_{ds} versus V_{ds} from 0 to 5 V for $V_{gs} = 0.5{-}4$ V in 0.5-V steps, g_{mo} versus V_{gs} from 0 to 5 V for $V_{ds} = 0$ to 5 V in 0.5-V steps, and g_{mbo} versus V_{bs} from −3 to 3 V for V_{ds} from 1 to 5 V in 0.5-V steps (at $V_{gs} = 2$ V) using the device's proper current–voltage characteristics or curves along with a commercially available computer program. Examine and discuss the results. Estimate and plot $g_{mbo} = \left.\frac{\partial I_{ds}}{\partial V_{bs}}\right|_{V_{gs}=\text{constant}}$ versus V_{bs} from −3 to 3 V for V_{gs} from 1 to 5 V in 0.5-V steps (at $V_{ds} = 2$ V). Compare and discuss the results with those previously obtained for g_{mbo} when V_{ds} was varied. You may need to change the values for the bias voltages V_{gs}, V_{ds}, and/or V_{bs} suitable for the particularly used device to obtain results sufficient for a more complete examination of data.

b) Estimate and plot R_{ds} (i.e., R_{dsb}) versus V_{ds} from 0 to 5 V for $V_{gs} = 0.5$, 2 and 4 V and g_{mo} versus V_{gs} from 0 to 5 V for $V_{ds} = 0$, 2, and 5 V using Eqs. (9.66) and (9.59), respectively, and compare and discuss results to those obtained in Part a.

9.11 Consider a 0.25-μm MOSFET or any submicron MOSFET available from a CMOS foundry.

a) Compute C_{gs} and C_{gd} using Eq. (9.18) neglecting the fringing capacitances. Use values for W, L_{diff}, ε_{ox}, and t_{ox} provided by the foundry.

b) Compute C_{gs} and C_{gd} taking into account the fringing capacitances using a commercially available EM simulator. Compare to those obtained in (a) and comment on the results.

9.12 Consider a 0.18-μm MOSFET or any submicron MOSFET available from a CMOS foundry to be used as a passive device. The bulk is connected to the ground.

a) Obtain the MOSFET's two-port small-signal *S*-parameters under on- and off bias conditions from 0.1 to 20 GHz. These *S*-parameters can be obtained either through measurements using a vector network analyzer or through calculations using an available linear or nonlinear model from the CMOS foundry or a commercially available program's library. Ports 1 and 2 are used for the source and drain terminal, respectively, and the gate is used for the control bias voltage through a large resistor (e.g., 10 kΩ).

b) Use a commercially available program to numerically determine the on- and off model's element values by fitting the *S*-parameters obtained in (a) to the *S*-parameters calculated from the equivalent circuit in Figure 9.25. Use proper initial values for the fitting process.

c) Calculate the errors between the *S*-parameters used for the fitting and those calculated from the resultant model versus frequency from 0.1 to 10 GHz. Plot these *S*-parameters on Smith charts for comparison and comment on the results.

9.13 Consider the same MOSFET in Problem 9.12 used as a passive device with the bulk floated using a resistor of 15 kΩ. Repeat Problem 9.12 using the equivalent-circuit models in Figures 9.25 and 9.26. Comment on the obtained on- and off-models from Figure 9.25 and between those of Figures 9.25 and 9.26. You can also use the following *S*-parameters between the Source (port 1) and the Drain (port 2) from 1 to 20 GHz for a 0.18-μm MOSFET with its bulk floated using a resistor of 15 kΩ. The device is biased with 1.8 V (on-state) and 0 V (off-state) at the gate through a 500-Ω resistor.

On-state

Frequency (GHz)	Mag(S_{11})	Deg(S_{11})	Mag(S_{12})	Deg(S_{12})	Mag(S_{21})	Deg(S_{21})	Mag(S_{22})	Deg(S_{22})
1	0.0847	−12.33	0.8984	−1.11	0.8982	−1.12	0.0855	−12.10
2	0.0767	−20.95	0.8875	−1.68	0.8873	−1.70	0.0773	−20.50
3	0.0694	−26.36	0.8781	−1.87	0.8780	−1.90	0.0699	−25.66
4	0.0640	−29.91	0.8715	−1.90	0.8713	−1.94	0.0643	−28.93
5	0.0601	−32.60	0.8667	−1.89	0.8665	−1.95	0.0604	−31.33
6	0.0573	−34.94	0.8631	−1.88	0.8628	−1.94	0.0575	−33.38
7	0.0551	−37.15	0.8603	−1.87	0.8598	−1.94	0.0553	−35.31
8	0.0534	−39.32	0.8578	−1.87	0.8573	−1.95	0.0536	−37.20
9	0.0520	−41.45	0.8557	−1.87	0.8550	−1.95	0.0521	−39.06
10	0.0509	−43.55	0.8537	−1.87	0.8529	−1.96	0.0510	−40.91
11	0.0499	−45.61	0.8519	−1.87	0.8510	−1.96	0.0499	−42.73
12	0.0490	−47.63	0.8503	−1.87	0.8492	−1.97	0.0491	−44.52
13	0.0483	−49.59	0.8487	−1.87	0.8475	−1.97	0.0483	−46.27
14	0.0476	−51.50	0.8472	−1.87	0.8459	−1.97	0.0476	−47.98
15	0.0470	−53.35	0.8458	−1.86	0.8444	−1.96	0.0470	−49.65
16	0.0465	−55.15	0.8445	−1.86	0.8430	−1.96	0.0465	−51.27
17	0.0461	−56.88	0.8433	−1.85	0.8417	−1.95	0.0460	−52.84
18	0.0457	−58.55	0.8421	−1.84	0.8404	−1.94	0.0456	−54.36
19	0.0454	−60.16	0.8410	−1.84	0.8392	−1.93	0.0453	−55.84
20	0.0451	−61.72	0.8400	−1.83	0.8381	−1.92	0.0450	−57.27

Off-state

	Mag(S_{11})	Deg(S_{11})	Mag(S_{12})	Deg(S_{12})	Mag(S_{21})	Deg(S_{21})	Mag(S_{22})	Deg(S_{22})
1	0.9889	−2.39	0.0170	98.89	0.0146	104.40	0.9966	−2.19
2	0.9824	−4.61	0.0365	98.30	0.0324	103.80	0.9914	−4.25
3	0.9742	−6.65	0.0583	97.25	0.0533	102.50	0.9848	−6.18
4	0.9656	−8.56	0.0817	95.38	0.0762	100.20	0.9777	−7.99
5	0.9572	−10.36	0.1058	93.13	0.1000	97.54	0.9707	−9.73
6	0.9491	−12.11	0.1300	90.77	0.1241	94.77	0.9637	−11.43
7	0.9411	−13.81	0.1541	88.43	0.1481	92.09	0.9567	−13.09
8	0.9331	−15.49	0.1779	86.18	0.1718	89.54	0.9496	−14.73
9	0.9250	−17.14	0.2014	84.01	0.1952	87.12	0.9422	−16.36
10	0.9167	−18.77	0.2243	81.93	0.2180	84.84	0.9346	−17.96
11	0.9082	−20.38	0.2467	79.94	0.2404	82.67	0.9267	−19.54
12	0.8994	−21.96	0.2686	78.03	0.2623	80.60	0.9184	−21.09
13	0.8904	−23.52	0.2900	76.19	0.2836	78.62	0.9099	−22.63
14	0.8812	−25.05	0.3107	74.41	0.3044	76.73	0.9011	−24.14
15	0.8718	−26.56	0.3309	72.70	0.3245	74.91	0.8921	−25.63
16	0.8622	−28.04	0.3504	71.04	0.3441	73.15	0.8828	−27.09
17	0.8524	−29.49	0.3693	69.44	0.3630	71.46	0.8734	−28.52
18	0.8425	−30.91	0.3877	67.89	0.3814	69.83	0.8637	−29.93
19	0.8325	−32.31	0.4053	66.38	0.3991	68.25	0.8539	−31.31
20	0.8225	−33.68	0.4224	64.93	0.4162	66.73	0.8440	−32.66

9.14 Repeat Parts b and c of Problem 9.13 using the equivalent-circuit models in Figure 9.27. Compare the accuracy of the obtained models with those obtained in Part b of Problem 9.13 with respect to the *S*-parameters obtained or provided in Part b of Problem 9.13. (Which models match better to the *S*-parameters?)

9.15 Describe an available varactor for RF applications from a CMOS foundry or any source including its layout, equivalent-circuit model along with its junction capacitance versus reverse-biased voltage and other element values, physical dimensions such as length, width, and number of fingers, operating frequencies, and bias conditions.

9.16 Use the available *S*-parameters (either measured or calculated from the equivalent-circuit model) of the CMOS varactor in Problem 9.15 to determine the lumped-element equivalent-circuit model shown in Figure 9.29(b).

9.17 Repeat Problem 9.16 for the equivalent-circuit model shown in Figure 9.30. Compare the obtained junction capacitance values with those in Problems 9.15 and 9.16.

9.18 Consider the CMOS varactor in Problem 9.15, determine the coefficient m of the capacitance in Eq. 9.85 from $\log[C_j(V_r)/C_j(0)]$ versus $\log(1 - V_r/\phi_b)$. Assume $\phi_b = 0.5$ V.

9.19 Consider a LC resonator made up by a spiral inductor having $L = 0.8$ nH and a varactor having $C_j(0) = 0.7$ pF and $\phi_b = 0.3$ V. Calculate and plot the resonant frequency versus reverse-biased voltage from -5 to 0 V for linearly graded, abrupt, and hyperabrupt ($m = 1$ and 2) junctions. Comment on the resultant resonant frequencies.

9.20 Consider the varactor in Problem 9.15.

a) Calculate and plot the cutoff frequency $f_c(V_r)$ versus V_r from -4 to 0 V.
b) Calculate and plot the quality factor $Q(V_r)$ versus f for V_r from 4 to 0 V in 0.5-V steps.
c) Calculate the dynamic cutoff frequency f_{cd}.
d) Calculate and plot the dynamic quality factor Q_d as a function of frequency from 0.5 to 20 GHz for the abrupt and graded junctions.

9.21 In circuit simulations using SPICE or SPICE-like computer programs, a diode is typically used to model the voltage-dependent capacitance C_j in the lumped-element equivalent circuit models shown in Figures 9.29 and 9.30. Use the varactor model (having a diode representing the junction capacitance) provided by the CMOS foundry for the varactor in Problem 9.15 (if available) to calculate and plot the junction capacitance versus reverse-biased voltage. Compare these capacitance values with those obtained in Problems 9.16 and 9.17.

CHAPTER 10

STABILITY

Stability is a very important issue in all circuits employing active devices such as transistors, in general, or RFICs in particular. Depending on stability, an active device can act as a power-generation or nonpower-generation device, and hence an RFIC employing it can behave as an oscillator or non-oscillation circuit such as amplifier, mixer, or switch. From the usage of individual components themselves or of subsystems and/or systems employing these components, stability plays an even more important role than other typical performance-measured parameters such as gain or noise figure because, without stability, a nonpower-generation component such as mixer, and hence systems containing it, becomes useless no matter how high its conversion gain and/or how low its noise figure is. Therefore, stability should always be considered in the initial and final design of all RFICs as well subsystems and systems – not only at the center frequency, but also over a frequency range sufficiently wider than the desired operating bandwidth. In this chapter, we will present an analysis of stability for RFICs employing transistors. The result of this analysis will be used directly in the amplifier design presented in Chapter 11. It should be noted that, although typical emphasis on stability is for amplifiers, it is equally important to consider stability for other RFICs such as mixers as well.

10.1 FUNDAMENTALS OF STABILITY

The stability of a circuit or an active device such as transistor is typically classified as either unconditionally stable or potentially unstable. Unconditional stability implies that the device or circuit can be used only as a nonpower-generation device or circuit such as amplifier or mixer. On the other hand, a potentially unstable device or circuit may be used for generating RF power such as oscillators or for generating no power such as amplifiers or mixers.

Consider a general two-port circuit in a Z_o-impedance system as shown in Figure 10.1. The two-port circuit is shown including its input and output matching networks, and may represent a physical two-port circuit such as an amplifier or a two-port sub-circuit in a multiport circuit such as the circuit between the RF and LO ports in a three-port mixer with its IF port terminated. At any frequency and under any bias condition, the two-port circuit is unconditionally stable when the magnitudes of the input reflection coefficient, $|\Gamma_{in}|$, and

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

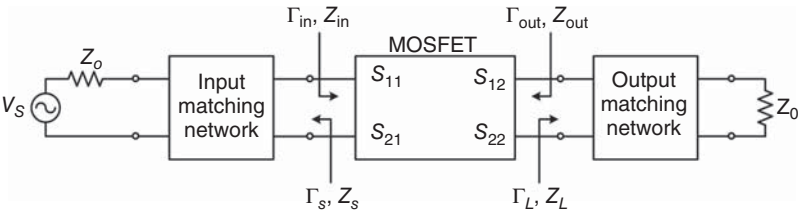


Figure 10.1. A general two-port circuit with input and output matching networks. The *S*-parameter block represents a MOSFET. Herein, we consider Z_S and Z_L as representing an input load and output load for the two-port circuit, respectively.

output reflection coefficient, $|\Gamma_{\text{out}}|$, are both less than 1, regardless any input (Z_L) and output (Z_S) terminating load impedance, provided that these loads are passive. It is noted that $|\Gamma_{\text{in}}| < 1$ and $|\Gamma_{\text{out}}| < 1$ correspond to the real part of the input impedance,$\text{Re}(Z_{\text{in}})$, and output impedance, $\text{Re}(Z_{\text{out}})$, being positive, respectively. Under the unconditional stability condition, the circuit can be used only as a nonpower-generation circuit such as amplifier or mixer. On the other hand, the two-port circuit is potentially unstable at any frequency and under any bias condition if either $|\Gamma_{\text{in}}|$ or $|\Gamma_{\text{out}}|$ is greater than 1, or, in turn, $\text{Re}(Z_{\text{in}})$ or $\text{Re}(Z_{\text{out}})$ is negative, respectively, for some passive output (Z_L) or input (Z_S) terminating impedance, respectively. Under this condition, the two-port circuit may be used for generating RF power such as oscillator or for other circuits such as amplifier. In order to be used as nonpower-generation device, the circuit, however, needs to be terminated with proper output and input impedance to make $|\Gamma_{\text{in}}| < 1$ and $|\Gamma_{\text{out}}| < 1$, respectively, or to make the total output loop resistance $\text{Re}(Z_L + Z_{\text{out}}) > 0$ and input loop resistance $\text{Re}(Z_S + Z_{\text{in}}) > 0$.

Using Eqs. (A11.14) and (A11.15) for $|\Gamma_{\text{in}}|$ and $|\Gamma_{\text{out}}|$ derived in the Appendix (Signal Flow Graph) of Chapter 11 (Amplifiers), respectively, we can write the "necessary and sufficient conditions" for the two-port circuit's unconditional stability under any passive terminating impedance Z_S and Z_L as

$$\begin{aligned} |\Gamma_S| &< 1 \\ |\Gamma_L| &< 1 \\ |\Gamma_{\text{in}}| &= \left| S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \right| < 1 \\ |\Gamma_{\text{out}}| &= \left| S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \right| < 1 \end{aligned} \tag{10.1}$$

where S_{ij} $(i,\ j = 1,\ 2)$ are the *S*-parameters of the employed transistor. Note that the first two conditions simply imply that Z_S and Z_L are passive and hence are redundant for passive terminations. Solving the last two equations of (10.1), we can obtain the necessary and sufficient conditions to achieve the unconditional stability for the transistor as [1]–[5]:

$$\begin{aligned} K &= \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} > 1 \\ 1 - |S_{11}|^2 &> |S_{12}S_{21}| \\ 1 - |S_{22}|^2 &> |S_{12}S_{21}| \end{aligned} \tag{10.2}$$

where

$$|\Delta| = S_{11}S_{22} - S_{12}S_{21} \tag{10.3}$$

or

$$\begin{aligned} K &= \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} > 1 \\ B_1 &= 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2 > 0 \end{aligned} \tag{10.4}$$

or

$$K > 1$$
$$|\Delta| < 1 \tag{10.5}$$

The last condition (10.5) is the simplest one among the foregoing conditions and normally used in practice to determine the stability. For a given MOSFET with S-parameters, the stability can then be assessed depending on the values of $|S_{11}|$ and $|S_{22}|$ as follows:

1. $|S_{11}| < 1$ and $|S_{22}| < 1$: If $K > 1$ and $|\Delta| < 1$ then the transistor is unconditionally stable. If either $K < 1$ or $|\Delta| > 1$ then the device is potentially unstable.
2. $|S_{11}| > 1$ or $|S_{22}| > 1$: The transistor is potentially unstable. Possible oscillation can occur. The circuit employing the device is stable or unstable depending upon the values of the input (Z_S) and output (Z_L) terminating impedances.

For unilateral device ($S_{12} = 0$), we obtain from (10.2):

$$K \to \infty$$
$$1 - |S_{11}|^2 > 0$$
$$1 - |S_{22}|^2 > 0 \tag{10.6}$$

from which, the necessary and sufficient conditions for unconditional stability of unilateral devices are

$$|S_{11}| < 1$$
$$|S_{22}| < 1 \tag{10.7}$$

In practice, most transistors used in RFICs are almost unilateral and, hence, the possibility of stability can be assessed by first checking $|S_{11}|$ and $|S_{22}|$. If both are less than 1, then the device is likely unconditionally stable. If either one is greater than 1 then there is a potential of instability. As can be seen in Figure 10.1, the stability of a two-port circuit depends not only on the employed MOSFET, but also on the input and output matching networks and the terminations at the circuit's input and output ports. For unconditionally stable devices, any passive terminations and passive matching networks – or, in turn, any passive input (Z_S) and output (Z_L) terminating impedances can be used– to achieve stability. For potentially unstable devices, however, the input and output terminating impedances must be chosen properly to obtain stability for circuits other than oscillators.

Assessing the stability based on conditions (10.2)–(10.5) requires evaluation of several parameters. A single stability parameter was proposed in [6] to derive another necessary and sufficient condition for achieving absolute stability. A device is unconditionally stable if and only if

$$\mu = \frac{1 - |S_{11}|^2}{|S_{22} - \Delta S_{11}^*| + |S_{12}S_{21}|} > 1 \tag{10.8}$$

or, equivalently,

$$\mu' = \frac{1 - |S_{22}|^2}{|S_{11} - \Delta S_{22}^*| + |S_{12}S_{21}|} > 1 \tag{10.9}$$

Use of such single stability parameter facilitates the assessment of stability and instability versus changes in the device's parameters. For unilateral devices, the unconditional stability condition in (10.8) becomes

$$\mu = \frac{1 - |S_{11}|^2}{|S_{22} - \Delta S_{11}^*|} > 1 \tag{10.10}$$

which only occurs if and only if $|S_{11}| < 1$ and $|S_{22}| < 1$, which is the necessary and sufficient condition for absolute stability of unilateral devices as stated in (10.7).

While the stability and required terminating impedances can be determined analytically – for instance, for a given device, Z_L or Z_S can be determined to make $\mathrm{Re}(Z_{\mathrm{in}})$ or $\mathrm{Re}(Z_{\mathrm{out}})$ positive or negative, respectively – it is more convenient and comprehensible to use a graphical technique as typically employed in actual RFIC design.

10.2 DETERMINATION OF STABLE AND UNSTABLE REGIONS

We consider again the two-port circuit shown in Figure 10.1. The condition for unconditional stability described in (10.1) indicates that, for passive terminations Z_S and Z_L, the boundary between unconditional stability and potential instability corresponds to $|\Gamma_{\mathrm{in}}| = 1$ and $|\Gamma_{\mathrm{out}}| = 1$. Therefore, to determine the input and output terminations that will lead to stability or instability, we let

$$|\Gamma_{\mathrm{in}}| = \left| S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \right| = 1 \tag{10.11}$$

$$|\Gamma_{\mathrm{out}}| = \left| S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \right| = 1 \tag{10.12}$$

which, after some manipulations, give

$$\left| \Gamma_L - \frac{\left(S_{22} - \Delta S_{11}^*\right)^*}{|S_{22}|^2 - |\Delta|^2} \right| = \left| \frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2} \right| \tag{10.13}$$

$$\left| \Gamma_S - \frac{\left(S_{11} - \Delta S_{22}^*\right)^*}{|S_{11}|^2 - |\Delta|^2} \right| = \left| \frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2} \right| \tag{10.14}$$

respectively. Equations (10.13) and (10.14) establish the boundary between stability and instability for two-port circuits.

Letting $\Gamma_{L(S)} = \Gamma_{LR(SR)} + j\Gamma_{LI(SI)}$ and substituting them into (10.13) and (10.14) leads to

$$(\Gamma_{LR} - C_{LR})^2 + (\Gamma_{LI} - C_{LI})^2 = R_L^2 \tag{10.15}$$

and

$$(\Gamma_{SR} - C_{SR})^2 + (\Gamma_{SI} - C_{SI})^2 = R_S^2 \tag{10.16}$$

respectively, after some manipulations. Equations (10.15) and (10.16) represent two sets of circles in the Γ_L- and Γ_S-plan, which hereafter are referred to as the "output stability circles" and "input stability circles," respectively. The output stability circle is centered at

$$C_L = (C_{LR}, C_{LI}) = \frac{(S_{22} - \Delta S_{11}^*)^*}{|S_{22}|^2 - |\Delta|^2} \tag{10.17}$$

where (C_{LR}, C_{LI}) is the coordinates of the center with respect to the center of the Smith chart, and has a radius of

$$R_L = \left| \frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2} \right| \tag{10.18}$$

The input stability circle has its center located at

$$C_S = (C_{SR}, C_{SI}) = \frac{(S_{11} - \Delta S_{22}^*)^*}{|S_{11}|^2 - |\Delta|^2} \tag{10.19}$$

where (C_{SR}, C_{SI}) represents the center's coordinates, and a radius of

$$R_S = \left| \frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2} \right| \tag{10.20}$$

The output stability circle contains the values of Γ_L corresponding to $|\Gamma_{\text{in}}| = 1$. For values of Γ_L not on the circle, $|\Gamma_{\text{in}}| < 1$ or $|\Gamma_{\text{in}}| > 1$. Figure 10.2 shows an output stability circle superimposed on the Smith chart. The distance from the center of the Smith chart to that of the output stability circle is $|C_L|$. Similarly, the input stability circle contains the values of Γ_S that produce $|\Gamma_{\text{out}}| = 1$. For values of Γ_S not on the circle, $|\Gamma_{\text{out}}| < 1$ or $|\Gamma_{\text{out}}| > 1$. Figure 10.3 shows an input stability circle on the Smith chart. The distance between the centers of the Smith chart and the output stability circle is $|C_S|$. As can be seen from (10.17)–(10.20), the stability circles of a device completely depend on its *S*-parameters. Therefore, for any potentially unstable MOSFET with known *S*-parameters, stability circles can be plotted on top of the Smith chart and used to determine values of Γ_S (or Z_S) and Γ_L (or Z_L) needed to achieve stability or instability for circuits employing such MOSFET device. These values can lie either inside or outside of the stability circles, but must be inside of the Smith chart in order to maintain $|\Gamma_{S(L)}| < 1$ or $\text{Re}(Z_{S(L)}) > 0$. Letting $\Gamma_L = 0$, which corresponds to the center of the

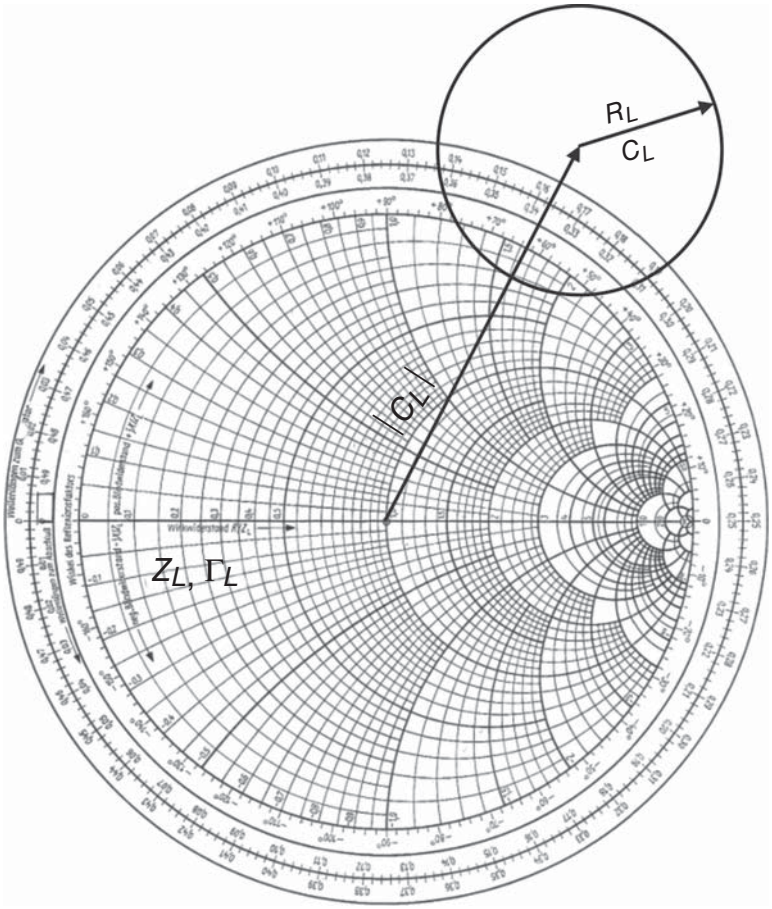


Figure 10.2. Output stability circle embedded on the Smith chart in the Γ_L-plane. Γ_L or Z_L must lie within the Smith chart, either inside or outside the stability circle. The radius of the Smith chart is 1.

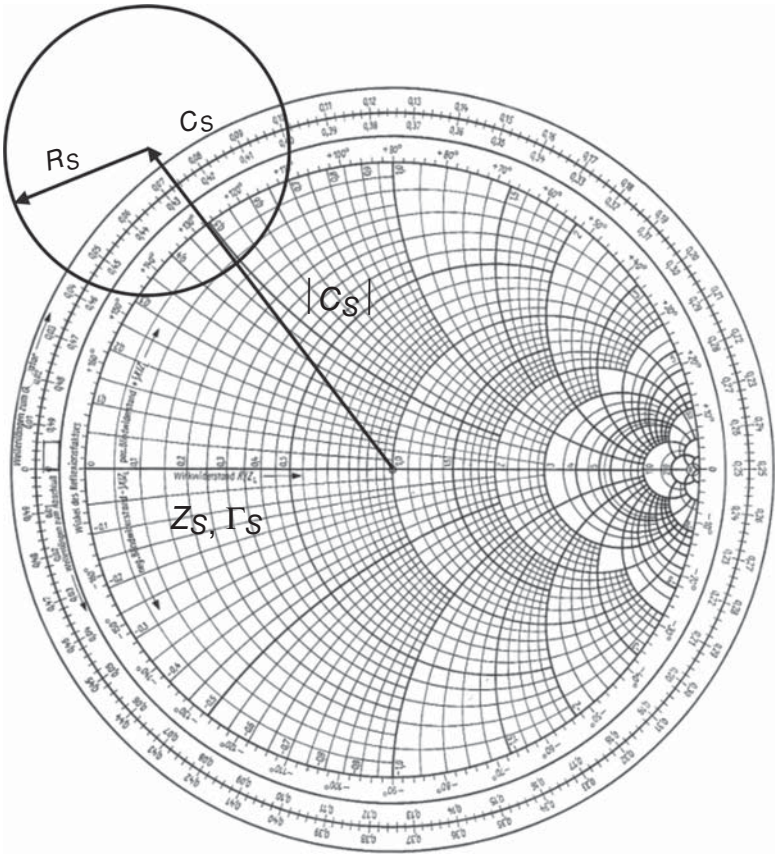


Figure 10.3. Input stability circle embedded on the Smith chart in the Γ_S-plane. Γ_S or Z_S must lie within the Smith chart, either inside or outside the stability circle.

Smith chart in the Γ_L-plan, we get $|\Gamma_{\text{in}}| = |S_{11}|$ from (10.11), which indicates that the center of the Smith chart represents stability when $|S_{11}| < 1$ and instability when $|S_{11}| > 1$. Similarly, the center of the Smith chart in the Γ_S-plan represents stability or instability when $|S_{22}|$ is less or greater than 1, respectively. These results indicate that we can determine stable and unstable regions by plotting the input and output stability circles on the Smith chart and determining if the region inside or outside the stability circles is stable or unstable by checking the stability condition at the center of the Smith chart. Figures 10.4 and 10.5 illustrate stable and unstable regions in these planes.

For unilateral devices, when either $|S_{11}| > 1$ or $|S_{22}| > 1$, the devices are potentially unstable due the negative resistance associated with S_{11} or S_{22}. We consider a normalized impedance $z = -r + jx$, where $r > 0$ and determine the corresponding reflection coefficient as

$$\Gamma = \frac{-(1+r)+jx}{1-r+jx} \tag{10.21}$$

from which, we get

$$\frac{1}{\Gamma^*} = \frac{r+jx-1}{r+jx+1} = \frac{z'-1}{z'+1} \tag{10.22}$$

where $z' = r + jx$ and hence $\text{Re}(z) = -\text{Re}(z')$ and $\text{Im}(z) = \text{Im}(z')$. We can thus see that the real part of the input or output impedance, $\text{Re}(Z_{\text{in}})$ or $\text{Re}(Z_{\text{out}})$, associated with S_{11} or S_{22} is equal to the "negative" of the real part of the impedance corresponding to $1/S_{11}^*$ or $1/S_{22}^*$, respectively. The stable regions can thus be

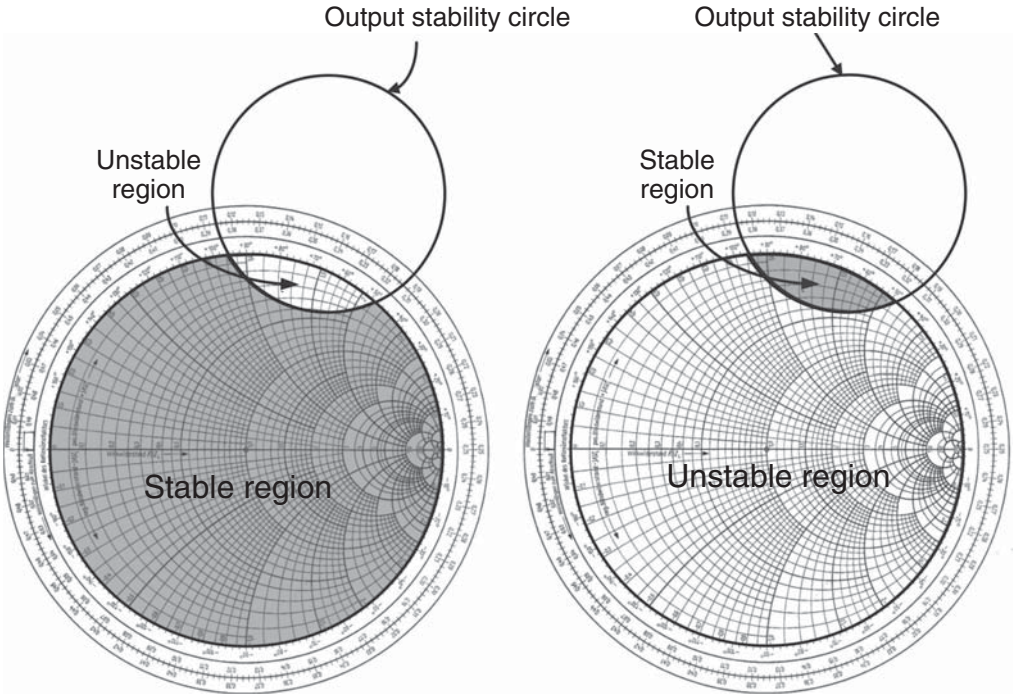


Figure 10.4. Stable and unstable regions in the Γ_L-plane for $|S_{11}| < 1$ (a) and $|S_{11}| > 1$ (b). The shaded and un-shaded regions inside the Smith chart in (a) and (b) define the regions containing Γ_L or Z_L that the output port must see in order to produce stability ($|\Gamma_{in}| < 1$) or instability ($|\Gamma_{in}| > 1$) at the input, respectively.

determined graphically by first locating $1/S_{11}^*$ and $1/S_{22}^*$ on a Smith chart and consider values of the corresponding resistance as being "negative" [i.e., considering $\mathrm{Re}(Z_{in})$ and $\mathrm{Re}(Z_{out})$], and second determine the stable regions as the regions containing Z_S and Z_L such as $\mathrm{Re}(Z_S) > |\mathrm{Re}(Z_{in})|$ and $\mathrm{Re}(Z_L) > |\mathrm{Re}(Z_{out})|$.

For unconditionally stable transistors, the devices and hence the two-port circuits employing them are always stable for any passive input and output terminating impedances. Since these impedances must be located inside the Smith chart, the input and output stability circles would fall completely outside or enclose completely the Smith chart to produce stable operation at all times, as illustrated in Figure 10.6. Accordingly, from the geometries in Figure 10.6, the device is always unconditionally stable when

$$||C_L| - R_L| > 1 \tag{10.23}$$

for $|S_{11}| < 1$, and

$$||C_S| - R_S| > 1 \tag{10.24}$$

for $|S_{22}| < 1$. When neither equation is satisfied, the device is potentially unstable and proper input and output terminating impedances need to be chosen to produce desired stability or instability for the two-port circuit.

Stability can also be achieved for a potentially unstable transistor by adding additional elements to terminals or between terminals of the device to stabilize it. This can essentially be viewed as modifying or transforming the original device into another configuration such that the unconditional stability is met. For instance, adding resistor, inductor or capacitor, or combination of these. Adding such elements to the device's terminals – for instance, at the gate of a MOSFET – will effectively transfer the input impedance of the device, and hence that of the two-port circuit, from a negative real part to a positive value or to that having less negative real part, which might lead to stability when combined with the input impedance. This can also be considered adding "impedance-transformation" networks to the device and, from that point of view, the

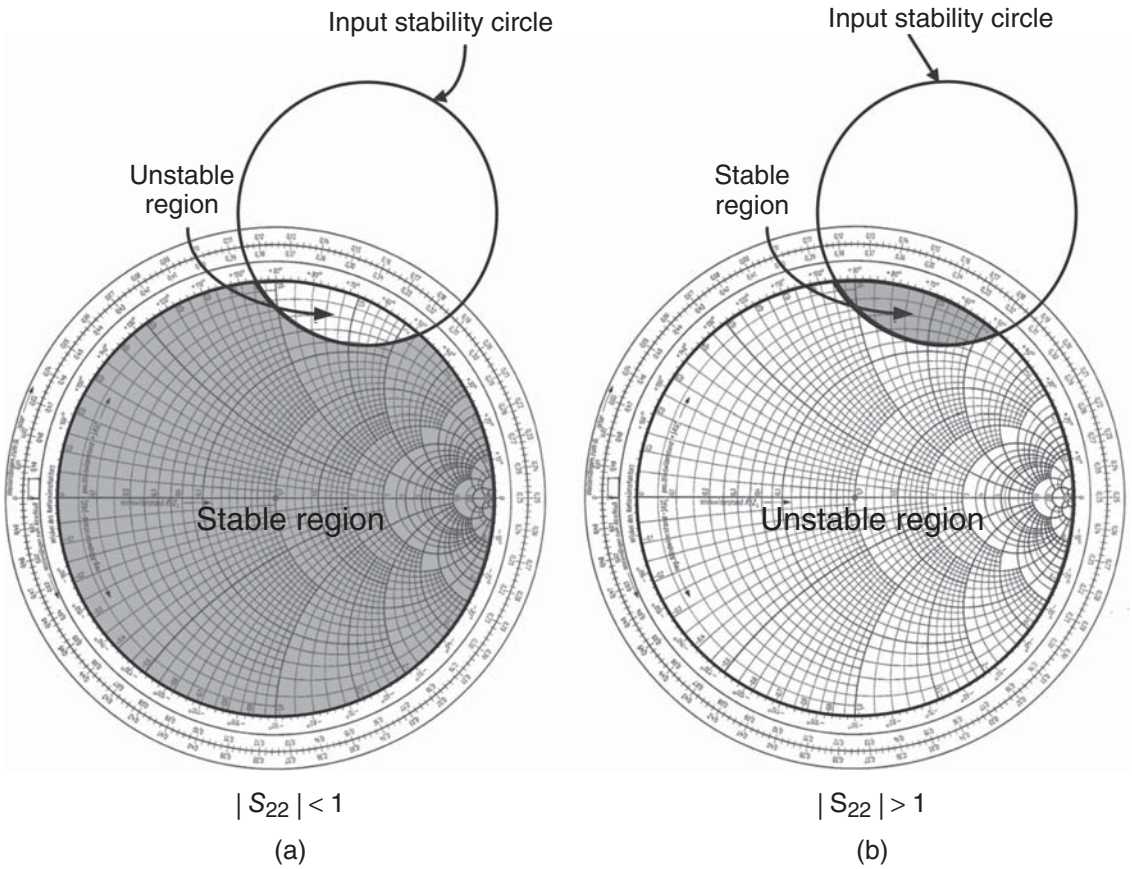


Figure 10.5. Stable and unstable regions in the Γ_S-plane for $|S_{22}| < 1$ (a) and $|S_{22}| > 1$ (b). The shaded and un-shaded regions inside the Smith chart in (a) and (b) define the regions containing Γ_S or Z_S that the input port must see in order to produce stability ($|\Gamma_{out}| < 1$) or instability ($|\Gamma_{out}| > 1$) at the output of circuits, respectively.

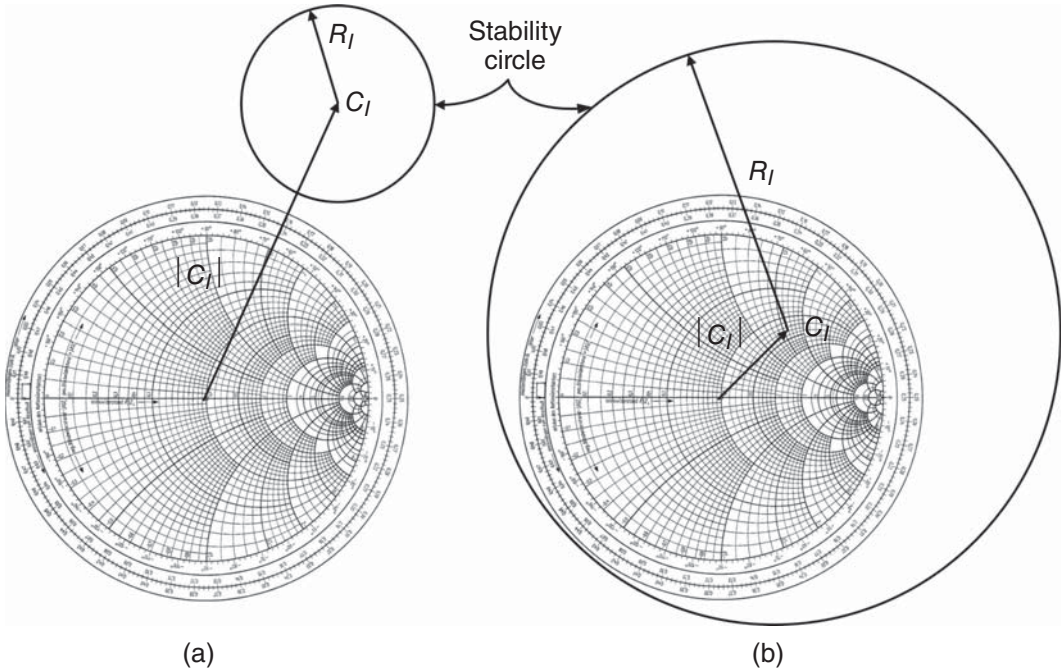


Figure 10.6. Input ($I = S$) and output ($I = L$) stability circles outside (a) and enclosing (b) the Smith chart for unconditionally stable transistors with $|S_{22}| < 1$ and $|S_{11}| < 1$, respectively.

technique - if implemented as broadband matching – can be very useful for stabilizing otherwise unstable broadband circuits. It should be noted, however, that adding resistors, particularly, produces possibly adverse effects such as increased noise figure, reduced gain and increased power consumption. Similarly, adding passive reactive elements, particularly inductors, may degrade circuit performance such as noise figure if the quality factor of the added elements is not sufficiently high, especially when the location of the added elements is highly sensitive for certain performance – for instance, adding an inductor at the gate of the device. Therefore, care should always be exercised to assess resultant effects due to added elements.

As an example to illustrate the determination of stability, we consider a 0.18-μm MOSFET with the following S-parameters at 10 GHz: $S_{11} = 0.9702\angle -27.45^{\circ}$, $S_{12} = 0.0681\angle 69.56^{\circ}$, $S_{21} = 3.0771\angle 155.75^{\circ}$, and $S_{22} = 0.7366\angle -7.22^{\circ}$, where ports 1 and 2 are at the gate and drain, respectively, and the source and bulk are tied together to the ground. The stability parameters are determined using (10.2) and (10.3) as

$$|\Delta| = S_{11}S_{22} - S_{12}S_{21} = 0.711$$

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} = 0.043$$

Since $K < 1$, the device is potentially unstable. To determine the stable and unstable regions, we plot the input and output stability circles on the Smith chart. The centers and radii of the input and output stability circles can be calculated from (10.17)–(10.20) as

$$C_S = \frac{(S_{11} - \Delta S_{22}^*)^*}{|S_{11}|^2 - |\Delta|^2} = 1.126\angle 45.085^{\circ}$$

$$R_S = \left|\frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2}\right| = 0.476$$

and

$$C_L = \frac{(S_{22} - \Delta S_{11}^*)^*}{|S_{22}|^2 - |\Delta|^2} = 5.402\angle 95.814^{\circ}$$

$$R_L = \left|\frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2}\right| = 5.266$$

Figure 10.7 shows these circles on the Smith chart. Since both $|S_{11}|$ and $|S_{22}|$ are less than 1, the center of the Smith chart falls into the stability region for both the output and input stability circles, respectively. Therefore, stable regions are located outside of the stability circles and inside of the Smith chart as noted in Figure 10.7. We then conclude that any impedance inside the Smith chart and outside the input and output stability circles can be chosen for the input and output terminating impedances, respectively, to produce stability for a two-port circuit, such as amplifier, employing such transistor at 10 GHz.

It should be noted that a final assessment of the circuit stability is always needed, which can be carried out by calculating various parameters – for instance, the input (Z_{in}) and output (Z_{out}) impedances of the circuit, making sure that their real parts are positive at the frequencies of interested. Unconditionally stable devices always produce stability regardless of any input- and output-terminating impedances, provided these terminations are passive elements. In integrated circuits, particularly those highly dense, there may not be "true" passive terminations at any port of a circuit since, away from each port, there are always active devices that may be transformed through intermediate lumped elements or interconnects into terminations, which act as "active" terminations for the considered circuit. These active terminations may produce positive $|\Gamma_S|$ or $|\Gamma_L|$ and, hence, altering the stability criterion of the entire circuit. To illustrate this undesired effect, we consider an unconditionally stable amplifier connected to a network consisting of a MOSFET and an inductor as shown in Figure 10.7. The amplifier, standing alone, is always unconditionally stable regardless of any

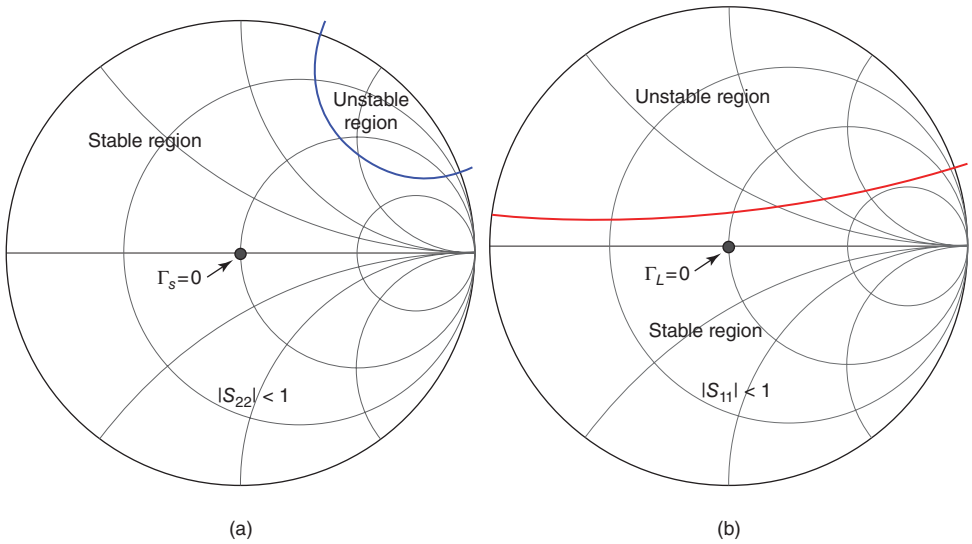


Figure 10.7. Input (a) and output (b) stability circles for the considered MOSFET with identified stable and unstable regions on the Smith chart.

passive input and output terminations. However, in the integrated-circuit environment, the extra MOSFET is transformed through the inductor to present an impedance to the input of the amplifier. This impedance may have a negative resistance, whose magnitude may be greater than the real part of the input impedance of the amplifier, thus resulting in a negative loop resistance at the input network and, hence, causing oscillation at the input. This problem is considered relatively simple and can be easily resolved. However, the unexpected instability is more severe in practical highly dense circuit environments in which both RF and digital circuits containing multiple active devices are integrated. Therefore, even a circuit is designed to achieve its unconditional stability (with any passive terminations), the circuit should be simulated again in the integrated-circuit environment upon which the circuit is integrated into. That is, considering all surrounding circuits, especially those adjacent to it. Furthermore, at the subsystem and system levels, subsystems and systems should be simulated considering constituent integrated circuits altogether to make sure the desired stability or instability is met as designed.

10.3 STABILITY CONSIDERATION FOR *N*-PORT CIRCUITS

The foregoing analysis and discussion of stability for two-port circuits can be extended to those having multiple ports. The stability can be carried out by considering two ports at a time while terminating all unused ports with appropriate impedances or proceed considering all ports simultaneously. Considering all ports together, we can write the "necessary and sufficient conditions" for N-port circuits' unconditional stability under any passive terminating impedance Z_i $(i = 1, 2, \ldots, N)$ as

$$|\Gamma_i| < 1$$
$$|\Gamma_n| = f(S_{ij}, \Gamma_{L1}, \ldots, \Gamma_{Lm}, \ldots, \Gamma_{LN}) < 1, \quad m \neq n$$

Other equations similar to those derived previously for two-port circuits can be derived.

REFERENCES

1. G. Gonzalez, Microwave Transistor Amplifiers, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1984.
2. K. Kurokawa, "Power Waves and the Scattering Matrix," *IEEE Trans. Microw. Theory Tech.*, Vol. MTT-13, No. 3, Mar. 1965.

3. D. Woods, "Reappraisal of the Unconditional Stability Criteria for Active 2-Port Networks in Terms of S Parameters," *IEEE Trans. Circuits Syst.*, Vol. CAS-23, No. 2, Feb. 1976.
4. W. H. Ku, "Unilateral Gain and Stability Criterion of Active Two-ports in Terms of Scattering Parameters," *Proc. IEEE*, Vol. 54, pp. 1617–1618, Nov. 1966.
5. R. P. Meys, "Review and Discussion of Stability Criteria for Linear 2-Ports," *IEEE Trans. Circuits Syst.*, Vol. CAS-37, No. 11, Feb. 1990.
6. M. L. Edwards and J. H. Sinsky, "A Single Stability Parameter for Linear 2-Port Circuits," *IEEE MTT-S Digest*, pp. 885–888, 1992.

PROBLEMS

10.1 Consider a 0.25-μm MOSFET with the following measured *S*-parameters:

Freq. (GHz)	S_{11}	S_{12}	S_{21}	S_{22}
8	$0.9244\angle -45.1^o$	$0.1103\angle 53.4^o$	$1.9542\angle 137.2^o$	$0.7176\angle -43.9^o$
9	$0.9073\angle -50.8^o$	$0.1211\angle 49^o$	$1.8677\angle 132.1^o$	$0.7015\angle -48.8^o$
10	$0.8960\angle -54.4^o$	$0.1275\angle 46.3^o$	$1.8096\angle 129^o$	$0.6914\angle -51.9^o$
11	$0.8797\angle -59.4^o$	$0.1359\angle 42.4^o$	$1.7244\angle 124.2^o$	$0.6771\angle -56.4^o$
12	$0.8693\angle -62.7^o$	$0.1409\angle 40^o$	$1.6685\angle 121.3^o$	$0.6682\angle -59.2^o$

Find out, using the stability condition given in Eq. (10.5), whether the transistor is:

a) Unconditionally stable at these frequencies.

b) Potentially unstable at these frequencies. Draw the input and output stability circles and indicate regions of stability and instability for one frequency.

10.2 A 0.18-μm NMOSFET has the following *S*-parameters at 20 GHz: $S_{11} = 0.818\angle -119.6^o$, $S_{12} = 0.065\angle 20.3^o$, $S_{21} = 5\angle 87.5^o$, and $S_{22} = 0.632\angle -67.4^o$. Ports 1 and 2 are at the gate and drain, respectively, and the source and bulk are connected to the ground. Draw the stability circles and determine the unstable and stable regions on the Smith chart.

10.3 Prove that when a load Z_S is passive, the corresponding reflection coefficient $|\Gamma_S|$ is less than 1.

10.4 Prove that $|\Gamma_{in}| < 1$ corresponds to negative $\text{Re}(Z_{in})$.

10.5 Derive the necessary and sufficient condition given in Eq. (10.2) for unconditional stability.

10.6 Derive the necessary and sufficient condition given in Eq. (10.4) for unconditional stability.

10.7 Derive the necessary and sufficient condition given in Eq. (10.5) for unconditional stability.

10.8 Derive the stability parameter given in Eq. (10.8).

10.9 Consider the MOSFET given in Problem 10.1. Verify the stability conditions obtained in Problem 10.1 using the stability parameter μ in (10.8) and μ' in (10.9).

10.10 Consider the stability condition given in (10.8) and that typically used in (10.5). Which one of these conditions would you prefer to use? Provide logical reasons for your choice.

10.11 Derive (10.13) and (10.14).

10.12 Derive (10.15)–(10.20).

10.13 Consider the MOSFET given in Problem 10.2. If the device is potentially unstable, determine the values of the resistors connected in series with the MOSFET at its input gate and output drain terminals to stabilize it or transform it into unconditional stability.

10.14 Consider the MOSFET given in Problem 10.2. If the device is potentially unstable, determine the feedback network between the gate and drain terminals to produce unconditional stability.

10.15 Consider a 0.25-μm MOSFET, whose *S*-parameters and noise parameters are available, from any foundry or commercially available computer program. Assume the device's input and output terminals are at the gate and drain, respectively, and the source and bulk are tied together to ground.

a) Determine a frequency at which the device is potentially unstable and the input impedance Z_S and output impedance Z_L that make the device stable at both input and output. Design input and output matching networks using two ideal elements (inductor and capacitor) to match to 50-Ω source and load impedance. Calculate the gain or loss, noise figure, and power consumption of the resultant two-port circuit using any available program in a 50-Ω system.

b) Connect a resistor in series with the gate and another one at the drain of the MOSFET to stabilize it. Determine the values of these resistors. Design lumped-element input and output matching networks using two ideal elements (inductor and capacitor) in the same topology as done in Part a to match to 50-Ω source and load impedance. Calculate the gain or loss, noise figure and power consumption of the resultant two-port circuit using any available program in a 50-Ω system. Assume the device is operated from the same voltage supply as in Part a. Compare and comment on the results with those obtained in Part a to assess the effects of the added resistors.

10.16 Consider two 0.13-μm MOSFETs. MOSFET 1 has the following *S*-parameters at 35 GHz: $S_{11} = 0.85\angle -39^{o}$, $S_{12} = 0$, $S_{21} = 3\angle 110^{o}$ and $S_{22} = 0.7\angle -52^{o}$. MOSFET 2 the following *S*-parameters at 35 GHz: $S_{11} = 0.368\angle -96.4^{o}$, $S_{12} = 0.682\angle -92^{o}$, $S_{21} = 0.856\angle -16.8^{o}$, and $S_{22} = 1.089\angle -135.3^{o}$. MOSFET 1 is integrated with MOSFET 2 via a 50-Ω quarter-wave transmission line at 35 GHz as shown in Figure P10.1. Determine the stability of the resulting two-port circuit. Compare the stable condition of this circuit with that using only MOSFET 1 with gate and drain terminated with any passive loads.

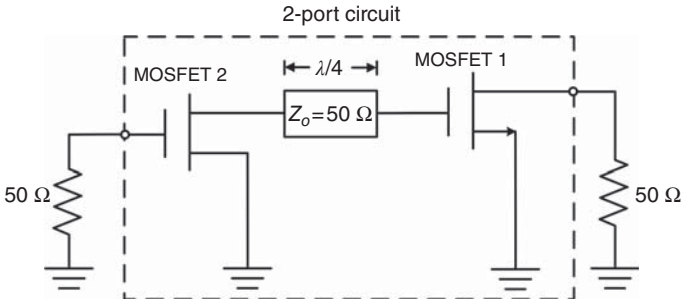


Figure P10.1.

CHAPTER 11

AMPLIFIERS

Amplifiers are one of the most important active components in radio frequency integrated circuits (RFICs) and used virtually in all RFIC systems. Amplifiers are, in principle, categorized into two classes: low noise amplifiers (LNAs) and power amplifiers (PAs). While LNAs are used practically in all receiver subsystems, PAs are used in most transmitters. Depending on applications, an amplifier may operate over a narrow or wide bandwidth. Broadband amplifiers require special design techniques and/or circuit topologies to achieve certain gain, noise figure, linearity, and radio frequency (RF) power, depending on their use as LNA or PA, over the desired bandwidths. This chapter presents the fundamentals and design of RF amplifiers, LNA, PA, balanced amplifiers, and broadband amplifiers.

11.1 FUNDAMENTALS OF AMPLIFIER DESIGN

RF amplifiers are covered in many papers and several books – for example, Reference [1] which is one of the excellent references that provides details of several topics in RF amplifiers. The design of amplifiers can be classified into five categories: design for certain or maximum gain, design for minimum noise figure, design for both gain and noise figure, design for maximum RF power handling, and design for broadband operation. The broadband design category may be considered as part of the other four design categories.

11.1.1 Power Gain

A typical block diagram of general and simple RF amplifiers is shown in Figure 11.1, which consists of an input matching network (IMN), a transistor, and an output matching network (OMN). For the sake of illustration, we use metal oxide semiconductor field effect transistor (MOSFET) for the transistor. $\Gamma_{S(L)}$ and $Z_{S(L)}$ represent the reflection coefficient and input impedance looking into the input (output) matching network, respectively. $\Gamma_{\text{in(out)}}$ and $Z_{\text{in(out)}}$ represent the reflection coefficient and input impedance looking into the input (output) of the MOSFET, respectively. S_{ij} $(i, j = 1, 2)$ stand for the S-parameters of the active device.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

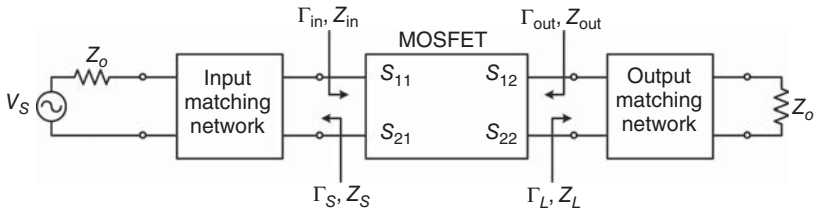


Figure 11.1. Typical block diagram of RF amplifiers. The *S*-parameter block is used to represent the transistor (MOSFET).

There are in general three different kinds of gain for amplifiers: available power gain, G_A, operating power gain, G_O, and transducer power gain, G. They are defined as follow.

$$G_A = \frac{\text{power available from network}}{\text{power available from source}} \tag{11.1}$$

$$G_O = \frac{\text{power delivered to load}}{\text{power delivered to network}} \tag{11.2}$$

and

$$G = \frac{\text{power delivered to load}}{\text{power available from source}} \tag{11.3}$$

The available power gain takes into account the matching between the IMN and the MOSFET and neglects that between the OMN and the device. The operating power, on the other hand, only considers the matching between the device and the OMN. While these power gains give insight into the effects of the individual input or OMN and occasionally used in the analysis, they do not represent the true gain of amplifiers under operation. The transducer power gain includes both the input and output matching and thus indeed represents the actual operating gain of amplifiers. This power gain is used in most measurements and calculations for amplifiers.

The transducer power gain or, in short, power gain of amplifiers, as seen in (11.3), is defined similar to the power gain of any RF active components as the ratio between the power delivered to the load and the power available from the source, which apparently depends on the IMNs and OMNs and the gain of the MOSFET. This gain is derived in Eqs. (A11.23)–(A11.25) of APPENDIX A11 using "signal-flow graphs" and repeated here for convenience:

$$G = \frac{|S_{21}|^2(1-|\Gamma_S|^2)(1-|\Gamma_L|^2)}{|(1-S_{11}\Gamma_S)(1-S_{22}\Gamma_L)-S_{12}S_{21}\Gamma_S\Gamma_L|^2} \tag{11.4}$$

$$G = \frac{1-|\Gamma_S|^2}{|1-\Gamma_{\text{in}}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \tag{11.5}$$

or

$$G = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{\text{out}}\Gamma_L|^2} \tag{11.6}$$

where

$$\Gamma_{\text{in}} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1-S_{22}\Gamma_L} \tag{11.7}$$

and

$$\Gamma_{\text{out}} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \tag{11.8}$$

As seen from Figure 11.1 and (11.5) or (11.6), there are three different gain blocks (the IMN, MOSFET device, and OMN) in the amplifier, and so (11.5) and (11.6) can be rewritten to signify the individual contributions of these gain blocks as

$$G = G_S G_D G_L \tag{11.9}$$

where

$$G_S = \frac{1 - |\Gamma_S|^2}{|1 - \Gamma_{\text{in}}\Gamma_S|^2} \tag{11.10}$$

and

$$G_L = \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \tag{11.11}$$

or

$$G_S = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} \tag{11.12}$$

and

$$G_L = \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_{\text{out}}\Gamma_L|^2} \tag{11.13}$$

are referred to as the input-block gain and output-block gains, respectively, and

$$G_D = |S_{21}|^2 \tag{11.14}$$

stands for the gain of the MOSFET itself which is the gain of an unmatched device. For a given device, S_{11}, S_{22}, and S_{21} are fixed, so the amplifier's gain, as seen from (11.9) to (11.13), depends on both the block gains G_S and G_L or, in turn, both the reflection coefficients Γ_S and Γ_L.

11.1.1.1 Transducer Power Gain with Unilateral Devices. Strictly speaking, all practical MOSFETs are bilateral devices having $S_{12} \neq 0$. However, typical MOSFETs used in RFICs including amplifiers resemble unilateral devices characterized by $S_{12} = 0$, and so an initial amplifier design typically starts with unilateral devices. For unilateral MOSFETs, we obtain from (11.7) and (11.8)

$$\Gamma_{\text{in}} = S_{11} \tag{11.15}$$

and

$$\Gamma_{\text{out}} = S_{22} \tag{11.16}$$

The corresponding power gain, known as the unilateral power gain, is obtained, from (11.5) or (11.6), as

$$G_u = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \tag{11.17}$$

which can be rewritten as

$$G_u = G_{Su} G_D G_{Lu} \tag{11.18}$$

where

$$G_{Su} = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} \tag{11.19}$$

and

$$G_{Lu} = \frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \tag{11.20}$$

while G_D is still given by (11.14).

The unilateral gain, as can be seen from (11.18) to (11.20), depends on G_S and G_L or, in turn, the reflection coefficients Γ_S and Γ_L, respectively. We can then choose G_S and G_L, and hence Γ_S and Γ_L, to achieve a particular gain for the amplifier. Therefore, under unilateral consideration, a fundamental design of amplifiers for a certain gain can start with the design of the IMNs and OMNs corresponding to Γ_S and Γ_L or G_S and G_L. In the following section, we will present an analysis from which Γ_S and Γ_L can be graphically or numerically determined from S_{11}, G_S and S_{22}, G_L, respectively, to achieve a desired amplifier gain.

11.1.2 Gain Design

A common typical design criterion for amplifiers is achieving maximum gain or particular gain. A common gain-design for amplifiers uses constant-gain circles regardless whether the design is conducted with or without a computer-aided design (CAD) program. This design technique is the focus of this section. Also, as discussed in Chapter 10, the stability of amplifiers, as of any active RFIC, is perhaps the design's foremost consideration, as amplifiers must first be stable in order to function properly. The design of amplifiers, for certain gain and/or noise figure, thus needs to revolve around the amplifiers' stability. It is recalled from Chapter 10 that there are two possible states of stability: unconditional stability and potential stability.

11.1.2.1 Gain Design with Unilateral Devices. The analysis starts with unconditional stability and then moves onto potential stability.

Unconditional Stability

The input- and output-block gains in (11.19) and (11.20) are rewritten as

$$G_i = \frac{1 - |\Gamma_i|^2}{|1 - S_{ii}\Gamma_i|^2} \tag{11.21}$$

where $i = Su$ corresponds to $ii = 11$ and $i = Lu$ corresponds to $ii = 22$. For unilateral MOSFETs, as derived in Chapter 10, the unconditional stability is obtained when both $|S_{11}|$ and $|S_{22}|$ are less than 1. We know that the maximum gain is achieved for an amplifier when there is a conjugate match between the IMN and the device's input and between the OMN and the device's output. For unilateral devices, this implies, from Figure 11.1 and (11.15) and (11.16), that

$$\Gamma_S = S_{11}^* \tag{11.22}$$

and

$$\Gamma_L = S_{22}^* \tag{11.23}$$

The corresponding source Z_S and load Z_L are the optimum gain terminations. Equation (11.21) becomes, using (11.22) and (11.23),

$$G_{i,\max} = \frac{1}{|1 - S_{ii}|^2} \tag{11.24}$$

The maximum gain is then obtained from (11.14), (11.18), and (11.24) as

$$G_{u,\max} = \frac{1}{1 - |S_{11}|^2}|S_{21}|^2\frac{1}{1 - |S_{22}|^2} \tag{11.25}$$

When $|\Gamma_i| = 1$, the corresponding input- or output-block gains become $G_{i,\min} = 0$. We can write

$$G_i = \begin{cases} 0, & |\Gamma_i| = 1 \\ G_{i,\max}, & |\Gamma_i| = S_{ii}^* \end{cases} \tag{11.26}$$

We now define a normalized gain for the input and output gain blocks as

$$\begin{aligned} g_i \equiv \frac{G_i}{G_{i,\max}} &= \frac{1 - |\Gamma_i|^2}{|1 - S_{ii}\Gamma_i|^2}(1 - |S_{ii}|^2) \\ &= G_i(1 - |S_{ii}|^2) \end{aligned} \tag{11.27}$$

This normalized gain is bounded as $0 \le g_i \le 1$ corresponds to (11.26) and shows that values for Γ_i can be determined from given values for G_i or g_i, or vice versa, for a given MOSFET. Since Γ_i and S_{ii} are complex numbers, we can let

$$\Gamma_i = U_i + jV_i \tag{11.28}$$

and

$$S_{ii} = A_{ii} + jB_{ii} \tag{11.29}$$

We obtain from (11.27):

$$g_i|1 - S_{ii}\Gamma_i|^2 = (1 - |\Gamma_i|^2)(1 - |S_{ii}|^2) \tag{11.30}$$

Expanding (11.30), we get

$$g_i(1 + |S_{ii}\Gamma_i|^2 - S_{ii}\Gamma_i - S_{ii}^*\Gamma_i^*) = 1 - |\Gamma_i|^2 - |S_{ii}|^2 + |\Gamma_i|^2|S_{ii}|^2 \tag{11.31}$$

which, after rearranging terms, becomes

$$|\Gamma_i|^2(1 - |S_{ii}|^2 + g_i|S_{ii}|^2) - g_iS_{ii}\Gamma_i - g_iS_{ii}^*\Gamma_i^* = 1 - g_i - |S_{ii}|^2 \tag{11.32}$$

Dividing (11.32) by $1 - |S_{ii}|^2(1 - g_i)$ gives

$$|\Gamma_i|^2 - g_i\frac{S_{ii}\Gamma_i - S_{ii}^*\Gamma_i^*}{1 - |S_{ii}|^2(1 - g_i)} = \frac{1 - g_i - |S_{ii}|^2}{1 - |S_{ii}|^2(1 - g_i)} \tag{11.33}$$

which can be rewritten as

$$\Gamma_i\Gamma_i^* - g_i\frac{S_{ii}\Gamma_i - S_{ii}^*\Gamma_i^*}{1 - |S_{ii}|^2(1 - g_i)} + \frac{g_i^2|S_{ii}|^2}{[1 - |S_{ii}|^2(1 - g_i)]^2} = \frac{(1 - g_i - |S_{ii}|^2)[1 - |S_{ii}|^2(1 - g_i)] + g_i^2|S_{ii}|^2}{[1 - |S_{ii}|^2(1 - g_i)]^2} \tag{11.34}$$

or

$$\left|\Gamma_i - \frac{g_iS_{ii}^*}{1 - |S_{ii}|^2(1 - g_i)}\right|^2 = \frac{(1 - g_i - |S_{ii}|^2)[1 - |S_{ii}|^2(1 - g_i)] + g_i^2|S_{ii}|^2}{1 - |S_{ii}|^2(1 - g_i)} \tag{11.35}$$

Equation (11.35) becomes, after some manipulations:

$$\left|\Gamma_i - \frac{g_iS_{ii}^*}{1 - |S_{ii}|^2(1 - g_i)}\right|^2 = \frac{\sqrt{1 - g_i}(1 - |S_{ii}|^2)}{1 - |S_{ii}|^2(1 - g_i)} \tag{11.36}$$

Substituting Γ_i and S_{ii} from (11.28) and (11.29), respectively, into (11.36) gives

$$\left| U_i - \frac{g_i A_{ii}}{1-|S_{ii}|^2(1-g_i)} + j\left[V_i \frac{g_i B_{ii}}{1-|S_{ii}|^2(1-g_i)}\right]\right| = \frac{\sqrt{1-g_i}(1-|S_{ii}|^2)}{1-|S_{ii}|^2(1-g_i)} \tag{11.37}$$

which is equivalent to

$$\left[U_i - \frac{g_i A_{ii}}{1-|S_{ii}|^2(1-g_i)}\right]^2 + \left[V_i + \frac{g_i B_{ii}}{1-|S_{ii}|^2(1-g_i)}\right]^2 = \left[\frac{\sqrt{1-g_i}\,(1-|S_{ii}|^2)}{1-|S_{ii}|^2(1-g_i)}\right]^2 \tag{11.38}$$

Equation (11.38) represents a set of circles for $\Gamma_i(U_i,V_i)$ in the complex Γ_i plane, each corresponding to a fixed value of g_i for a given S_{ii}. These circles are called the input or output "constant-gain circles" corresponding to values of the normalized gain g_i, or un-normalized gain G_i, of the input gain-block or the output gain-block, respectively.

The center $\Gamma_{io}(U_{io},V_{io})$ and radius R_i of these circles are obtained from (11.38) as

$$U_{io} = \frac{g_i A_{ii}}{1-|S_{ii}|^2(1-g_i)} \tag{11.39}$$

$$V_{io} = -\frac{g_i B_{ii}}{1-|S_{ii}|^2(1-g_i)} \tag{11.40}$$

and

$$R_i = \frac{\sqrt{1-g_i}(1-|S_{ii}|^2)}{1-|S_{ii}|^2(1-g_i)} \tag{11.41}$$

Figure 11.2 shows sketch of a constant-gain circle plotted within the Smith chart's outermost circle. The angle between the U_i axis and the line connecting the center O(0,0) of the Smith chart and the center

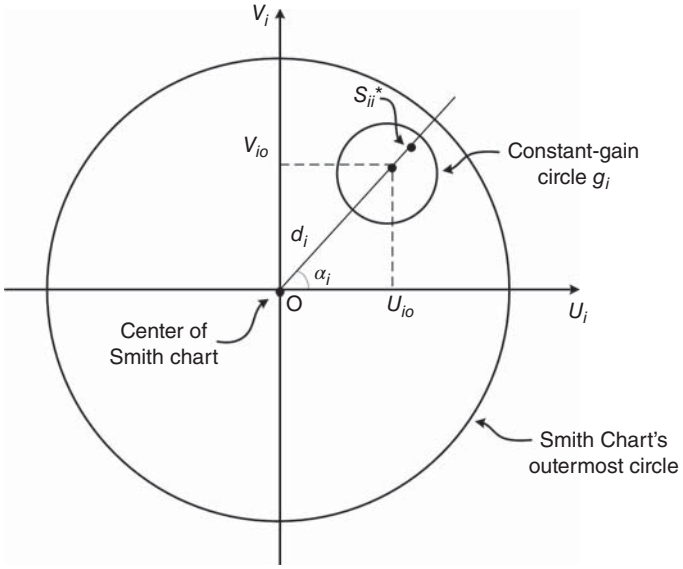


Figure 11.2. A constant-gain circle inside a Smith chart. The center of the (U_i,V_i) coordinates coincides with the center of the Smith chart. *i* stands for Su or Lu corresponding to *ii* representing 11 or 22, respectively.

$\Gamma_{io}(U_{io},V_{io})$ of the constant-gain circle is obtained as

$$\alpha_i = \tan^{-1}\frac{V_{io}}{U_{io}} = \tan^{-1}\left(-\frac{B_{ii}}{A_{ii}}\right) \tag{11.42}$$

utilizing (11.39) and (11.40), which is also the angle between the U_i axis and the line connecting O and $S_{ii}^* = A_{ii} - jB_{ii}$. The distance between the center of the Smith chart and the center of the constant-gain circle can be obtained using (11.39) and (11.40) as

$$d_i = \sqrt{U_{io}^2 + V_{io}^2} = \frac{g_i|S_{ii}|}{1 - |S_{ii}|^2(1 - g_i)} \tag{11.43}$$

Equations (11.42) and (11.43) indicate that the center of the constant-gain circle is located at a distance d_i from the center of the Smith chart on the line connecting that center and the point representing S_{ii}^*.

It is particularly noted that, for maximum gain $G_{i,\max}$, $g_i = 1$ and, from (11.41) and (11.43),

$$\begin{aligned} R_i &= 0 \\ d_i &= |S_{ii}| \end{aligned} \tag{11.44}$$

which indicates that the constant-gain circle for maximum gain is a point located at the same position of S_{ii}^*. This also implies that Γ_i must be located at S_{ii}^* in order to yield maximum gain. Furthermore, when the amplifier is perfectly matched at both the input and output of the MOSFET, that is, $\Gamma_S = \Gamma_L = 0$, we obtain from (11.21), (11.27), (11.41), and (11.43)

$$\begin{aligned} g_i &= 1 - |S_{ii}|^2 \\ G_i &= 1 \\ R_i &= d_i = \frac{|S_{ii}|}{1 + |S_{ii}|^2} \end{aligned} \tag{11.45}$$

which show that the 0-dB constant-gain circle, corresponding to $G_i = 1$, passes through the center of the Smith chart. This result is expected since, while Γ_i always lies on a constant-gain circle, $\Gamma_i = 0$ is always located at the center of the Smith chart. It is noted that this 0-dB gain actually implies no reflection loss through either the input or OMN; that is, perfect match between the source and IMN and between the load and OMN.

Any location on a constant-gain circle would produce Γ_i that will result in the corresponding gain. However, the choice of a specific location for Γ_i is needed and this selection is primarily for the ease of realizing the corresponding matching circuit. In the design of amplifiers, values for G_i can be chosen from the amplifier's specifications, employed MOSFET, (11.14) and (11.18), and from which corresponding values for Γ_i can be calculated using (11.19) and (11.20). As can be seen, different values of Γ_i can be calculated from a given G_i. However, some of these values may not lead to realizable matching networks, which not only prolong but also result in inconvenience in the design process, particularly without a good circuit synthesis computer program. A quick and convenient way in determining a realizable Γ_i is to graphically display the constant-gain circles for various gain values. This technique is useful not only for the insight and visualization of the design, but also for trade-off consideration between gain and noise figures using constant noise figure circles as discussed later. To allow direct determination of Γ_i or Z_i needed to achieve a particular gain, constant-gain circles corresponding to different gains are plotted on a Smith chart, from which appropriate Γ_i or Z_i on these circles corresponding to desired gains can be selected. The radius R_i of each circle is determined from (11.41). The center of the circle is determined using (11.39) and (11.40); it is located at a distance d_i from the center of the Smith chart on the line connecting the Smith chart's center and the point S_{ii}^*. Figure 11.3 shows some constant-gain circles for a 0.18-μm MOSFET on a Smith chart. For a design calling for maximum gain, Γ_i, $G_{i,\max}$ and $G_{u,\max}$ are obtained from (11.22) to (11.25).

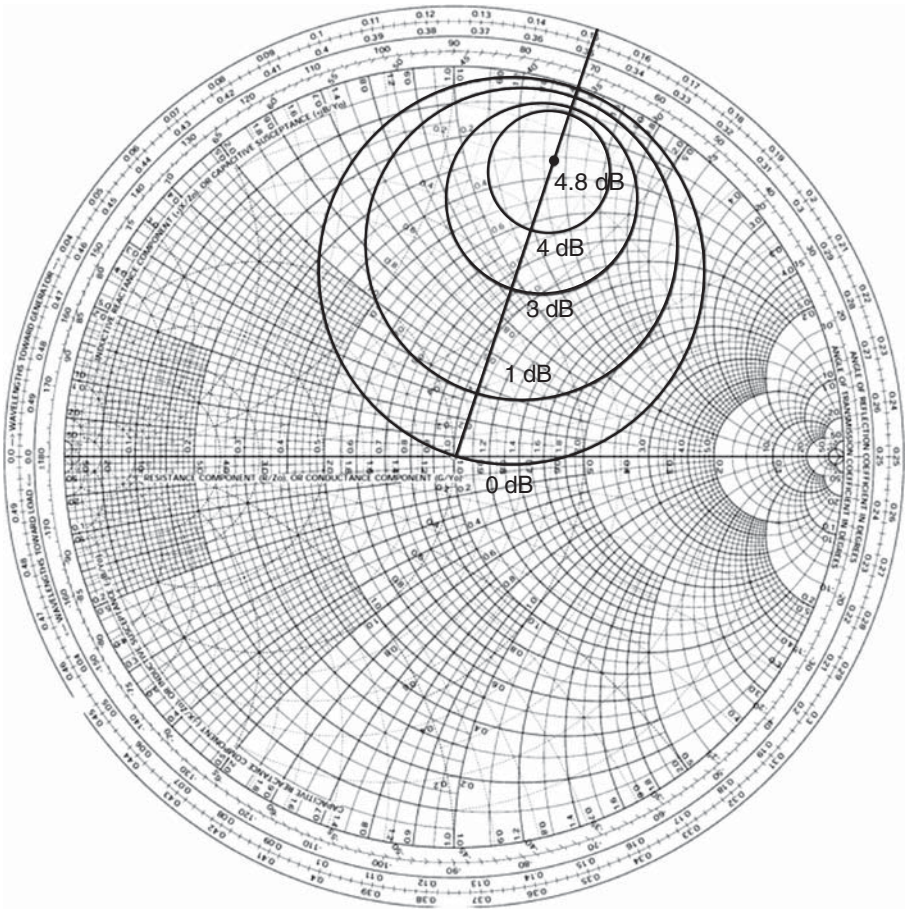


Figure 11.3. Constant-gain circles for a 0.18-μm MOSFET on Smith chart. As an example, any Z_S on the 4-dB circle will produce $G_S = 4$ dB.

Potential Stability

A potential or conditional stability occurs when either $|S_{11}|$ or $|S_{22}|$, or both, is greater than 1, which implies the employed MOSFET is potentially unstable, as discussed in Chapter 10. Under this condition, the constant-gain circles are still characterized by the same radius R_i and center $\Gamma_{io}(U_{io},V_{io})$ at distance d_i from the Smith chart's center and can be plotted similarly. The locations of the centers of these circles, however, are on the line connecting the Smith chart's center and the point $1/S_{ii}$, instead of S_{ii}^* as for the unconditional stability case. It is noted, however, that since the device is potentially unstable, care must be exercised to choose proper Γ_S or Z_S and Γ_L or Z_L on the constant circles so that they fall inside the stable regions within the Smith chart. To avoid possible oscillations due to variations of fabricated devices that may inadvertently put Γ_S and/or Γ_L inside of the unstable regions, Γ_S and Γ_L should be chosen sufficiently far away from their corresponding stability circles. It should also be noted that the final analysis of a designed amplifier must always include calculations confirming the stability of the entire amplifier. For instance, the real parts of both Z_{in} and Z_{out} must be positive across the frequencies of interest. If the real part of Z_{in} or Z_{out} is negative, the corresponding loop real impedance $\text{Re}(Z_{\text{in}} + Z_S)$ or $\text{Re}(Z_{\text{out}} + Z_L)$ must be positive across the desired frequency range.

Unilateral Figure of Merit

Practical MOSFETs have nonzero S_{12} and are thus bilateral devices. A design based on the equations derived earlier assuming unilateral devices, while practically simple and useful to implement, leads to errors. To assess

the magnitude of the resultant error caused by such assumption, the unilateral figure of merit (FOM) is generally used.

We begin the derivation of the unilateral FOM by taking the ratio between the actual gain G of the amplifier employing a bilateral device, given in (11.4), and the unilateral gain G_u, obtained from (11.17), as

$$\frac{G}{G_u} = \frac{1}{|1-U|^2} \tag{11.46}$$

where

$$U = \frac{S_{12}S_{21}\Gamma_S\Gamma_L}{(1-S_{11}\Gamma_S)(1-S_{22}\Gamma_L)} \tag{11.47}$$

For unilateral devices, $U=0$ and G becomes G_u as expected. The values of the ratio G/G_u, as seen from (11.38), are limited by

$$\frac{1}{(1+|U|)^2} < \frac{G}{G_u} < \frac{1}{(1-|U|)^2} \tag{11.48}$$

It is recalled that the maximum unilateral gain $G_{u,\max}$ is obtained when $\Gamma_S = S_{11}^*$ and $\Gamma_L = S_{22}^*$ and this corresponds to maximum possible error. Under this condition, we can write the limit for the gain ratio from (11.48) as

$$\frac{1}{(1+\text{FOM}_A)^2} < \frac{G}{G_{u,\max}} < \frac{1}{(1-\text{FOM}_A)^2} \tag{11.49}$$

where

$$\text{FOM}_A = \frac{|S_{12}S_{21}S_{11}S_{22}|}{(1-|S_{11}|^2)(1-|S_{22}|^2)} \tag{11.50}$$

or

$$\text{FOM}_A = 10\log\left[\frac{|S_{12}S_{21}S_{11}S_{22}|}{(1-|S_{11}|^2)(1-|S_{22}|^2)}\right] \text{ dB} \tag{11.51}$$

is referred to as the "unilateral figure of merit" – the smaller the value for FOM_A, the better approximation for gain using the unilateral gain equations. This parameter should be used to assess the error introduced by the unilateral assumption of a practical device having S_{12} different from zero. The unilateral figures of merit of typical MOSFETs are relatively small and hence acceptable, at least in the initial amplifier design. Figure 11.4 shows the unilateral FOM for a typical 0.18-μm MOSFET from 1 to 20 GHz.

11.1.2.2 Gain Design with Bilateral Devices.

Design Based on Transducer Power Gain

For certain MOSFETs, S_{12} may be sufficiently different from zero that causes non-negligible error in the gain calculation based on the unilateral gain equation (11.17). Amplifier design using such devices thus needs to take into account the nonzero S_{12}.

For a given MOSFET, maximum gain is achieved when the device is conjugately matched to both the IMNs and OMNs. The conjugate of Γ_S and Γ_L can be written, using (11.7) and (11.8), as

$$\Gamma_S^* = \Gamma_{\text{in}} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1-S_{22}\Gamma_L} \tag{11.52}$$

and

$$\Gamma_L^* = \Gamma_{\text{out}} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1-S_{11}\Gamma_S} \tag{11.53}$$

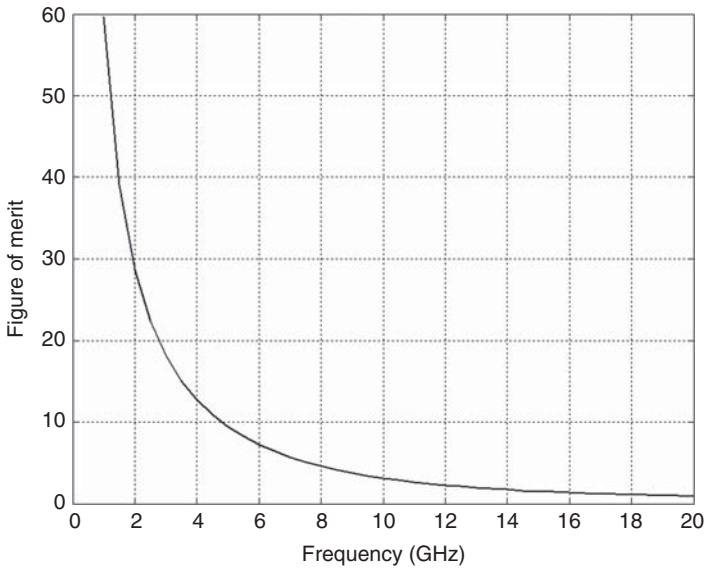


Figure 11.4. Unilateral figure of merit of a typical 0.18-μm MOSFET.

Taking the complex conjugate of (11.53) gives

$$\Gamma_L = S_{22}^* + \frac{S_{12}^* S_{21}^* \Gamma_S^*}{1 - S_{11}\Gamma_S} \tag{11.54}$$

Substituting Γ_L in (11.54) into (11.52), we get

$$\Gamma_S^* = S_{11} + \frac{S_{12}S_{21}\left(S_{22}^* + \frac{S_{12}^* S_{21}^* \Gamma_S^*}{1-S_{11}\Gamma_S}\right)}{1 - S_{22}\left(S_{22}^* + \frac{S_{12}^* S_{21}^* \Gamma_S^*}{1-S_{11}\Gamma_S}\right)} \tag{11.55}$$

Expanding (11.55) and arranging terms give

$$[S_{11}^*(|S_{22}|^2 - 1) - S_{12}^* S_{21}^* S_{22}](\Gamma_S^*)^2 + [(1 - |S_{22}|^2)(1 + |S_{11}|^2) + S_{11}S_{12}^* S_{21}^* S_{22} + S_{11}^* S_{12}S_{21}S_{22}^* - |S_{12}|^2|S_{21}|^2]\Gamma_S^* + [(|S_{22}|^2 - 1)S_{11} - S_{12}S_{21}S_{22}^*] = 0 \tag{11.56}$$

Taking the complex conjugate of (11.56) leads to

$$[S_{11}(|S_{22}|^2 - 1) - S_{12}S_{21}S_{22}^*]\Gamma_S^2 + [(1 - |S_{22}|^2)(1 + |S_{11}|^2) + S_{11}^* S_{12}S_{21}S_{22}^* + S_{11}S_{12}^* S_{21}^* S_{22} - |S_{12}|^2|S_{21}|^2]\Gamma_S + [(|S_{22}|^2 - 1)S_{11}^* - S_{12}^* S_{21}^* S_{22}] = 0 \tag{11.57}$$

which is a quadratic equation for Γ_S. Solving (11.57), we can obtain the optimum source reflection coefficient corresponding to maximum gain as

$$\Gamma_{SO} = \frac{A_1 \pm \sqrt{A_1^2 - 4|B_1|^2}}{2B_1} \tag{11.58}$$

where

$$A_1 = 1 + |S_{11}|^2 - |S_{22}|^2 - |\Delta|^2 \tag{11.59}$$

$$B_1 = S_{11} - S_{22}^*\Delta \tag{11.60}$$

with

$$\Delta = S_{11}S_{22} - S_{12}S_{21} \tag{11.61}$$

Similarly, we can derive the following quadratic equation for Γ_L from (11.52) and (11.53):

$$[S_{22}(|S_{11}|^2 - 1) - S_{12}S_{21}S_{11}^*]\Gamma_L^2 + [(1 - |S_{11}|^2)(1 + |S_{22}|^2) + S_{11}^*S_{12}S_{21}S_{22}^* + S_{11}S_{12}^*S_{21}^*S_{22} - |S_{12}|^2|S_{21}|^2]\Gamma_L + [(|S_{11}|^2 - 1)S_{22}^* - S_{12}^*S_{21}^*S_{11}] = 0 \tag{11.62}$$

which, upon solving, gives the optimum load reflection coefficient corresponding to maximum gain as

$$\Gamma_{LO} = \frac{A_2 \pm \sqrt{A_2^2 - 4|B_2|^2}}{2B_2} \tag{11.63}$$

where

$$A_2 = 1 + |S_{22}|^2 - |S_{11}|^2 - |\Delta|^2 \tag{11.64}$$

$$B_2 = S_{22} - S_{11}^*\Delta \tag{11.65}$$

The optimum source and load reflection coefficients can thus be completely determined from (11.58) and (11.63), respectively, using the device's *S*-parameters. Note that the "minus" sign in (11.58) and (11.63) is for the unconditional stability case.

The maximum gain can be obtained from (11.4) using the optimum source and load reflection coefficients Γ_{SO} and Γ_{LO} as

$$G_{\max} = \frac{|S_{21}|^2(1 - |\Gamma_{SO}|^2)(1 - |\Gamma_{LO}|^2)}{|(1 - S_{11}\Gamma_{SO})(1 - S_{22}\Gamma_{LO}) - S_{12}S_{21}\Gamma_{SO}\Gamma_{LO}|^2} \tag{11.66}$$

which can be rewritten, upon substituting (11.58)–(11.65) into (11.66), in terms of the stability factor K given in Eq. (10.2) as

$$G_{\max} = \left|\frac{S_{21}}{S_{12}}\right|(K - \sqrt{K^2 - 1}) \tag{11.67}$$

where K is repeated here for convenience:

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + \Delta^2}{2|S_{12}S_{21}|} \tag{11.68}$$

When $K = 1$, the maximum gain becomes

$$G_{\text{sta,max}} = \left|\frac{S_{21}}{S_{12}}\right| \tag{11.69}$$

which is the absolute maximum gain that an amplifier can attain using a stable device. This particularly gain is known as "the maximum stable gain." Although, theoretically, the gain for an amplifier using potentially unstable devices needs only to be equal or smaller than the maximum stable gain, in practice proper gain must be selected to be sufficiently smaller than the maximum stable gain to make sure not only the source and load impedances to fall within their respective stable region, but are also sufficiently far away from the respective stability circle for a more stable design.

For the gain-design of amplifiers, Γ_S and Γ_L are typically chosen according to a desired gain. It is recalled for the unilateral case that, for given device and desired amplifier gain, various G_S and G_L can be selected and corresponding gain circles can be drawn, from which proper Γ_S and Γ_L corresponding to G_S and G_L can be chosen, respectively. In the unilateral case, G_S and G_L are independent to each other, and so their gain

circles can be determined and plotted independently. The unilateral gain-design process is thus relatively simple.

For the bilateral care, however, the gain-design process is more involved. As can be seen in (11.11), G_L and hence the corresponding gain circles are a function of only Γ_L and S_{22}. Γ_L can thus be directly and easily determined from a gain circle. On the other hand, G_S, as seen in (11.7) and (11.10), depends not only on the device's S-parameters and Γ_S, but also on Γ_L and hence G_L. The gain circles for G_S cannot be thus drawn independently from the gain circles for G_L. Since the determination of Γ_L is independent of Γ_S while the calculation of Γ_S depends upon Γ_L, Γ_L must be chosen so that proper Γ_S can be obtained which, together with Γ_L, result in G_L and G_S that lead to a desired G. This process indeed involves iterations and is inconvenient.

In reality, most amplifiers employ transistors such as MOSFETs with very small S_{12} magnitude, which resemble well unilateral devices. Transistors with non-negligible S_{12} magnitude are not typically used, except at extremely high frequencies where only such devices are available to RFIC designers. Since typically used devices approximate well the unilateral nature and advanced CAD programs are commercially available, the simple design method based on unilateral devices can be used to yield good initial design, which can subsequently be optimized taking into account the device's bilateral behavior to obtain good final results using a CAD program.

Design Based on Operating Power Gain

In case the unilateral design procedure is neither suitable nor desired, a bilateral design can be implemented. However, to avoid the tedious (iterative) procedure based on the transducer power gain as mentioned in the foregoing section, under bilateral consideration, it is common to design amplifiers using constant-gain circles based on the available power gain G_A or operating power gain G_O defined earlier in (11.1) or (11.2), respectively. These gains can be derived using the signal-flow-graph technique described in APPENDIX A11 as

$$G_A = \frac{1-|\Gamma_S|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1}{1-|\Gamma_{\text{out}}|^2} \tag{11.70}$$

$$G_O = \frac{1}{1-|\Gamma_{\text{in}}|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \tag{11.71}$$

As can be seen, only Γ_L is involved in determining G_O, while only Γ_S is needed for G_A. Thus a simple procedure using constant-gain circles similar to that described for the unilateral case can be implemented for amplifiers employing unconditionally stable or potentially stable bilateral transistors using the available or operating power gain. Such a procedure based on the operating power gain is described as follows.

The operating power gain expression (11.71) can be rewritten making use of Γ_{in} in (11.7) as

$$G_O = |S_{21}|^2\frac{(1-|\Gamma_L|^2)}{|1-S_{22}\Gamma_L|^2-|S_{11}-\Delta\Gamma_L|^2} \tag{11.72}$$

where Δ is given in (11.61), or

$$G_O = |S_{21}|^2 g_o \tag{11.73}$$

where

$$\begin{aligned} g_o &= \frac{(1-|\Gamma_L|^2)}{|1-S_{22}\Gamma_L|^2-|S_{11}-\Delta\Gamma_L|^2} \\ &= \frac{(1-|\Gamma_L|^2)}{1-|S_{11}|^2+|\Gamma_L|^2(|S_{22}|^2-|\Delta|^2)-2\text{Re}(\Gamma_L B_2)} \end{aligned} \tag{11.74}$$

is defined as the normalized operating power gain, with B_2 being given in (11.65). Expanding (11.74) gives

$$|\Gamma_L|^2[1+g_o(|S_{22}|^2-|\Delta|^2)]-2g_o\text{Re}(\Gamma_L B_2) = 1-g_o(1-|S_{11}|^2) \tag{11.75}$$

or, after rearrangement,

$$|\Gamma_L|^2 - \frac{g_o \Gamma_L B_2}{1 + g_o(|S_{22}|^2 - |\Delta|^2)} - \frac{g_o \Gamma_L^* B_2^*}{1 + g_o(|S_{22}|^2 - |\Delta|^2)} = \frac{1 - g_o(1 - |S_{11}|^2)}{1 + g_o(|S_{22}|^2 - |\Delta|^2)} \tag{11.76}$$

Now comparing (11.78) with the expression

$$|z|^2 - zz_0^* - z^* z_0 = r^2 - |z_0|^2 \tag{11.77}$$

obtained from the equation for a circle of z with radius r and center at z_0

$$|z - z_0|^2 = r^2 \tag{11.78}$$

we recognize that (11.76) represents a circle for Γ_L with center located at

$$C_{go} = \frac{g_o B_2^*}{1 + g_o(|S_{22}|^2 - |\Delta|^2)} \tag{11.79}$$

and radius of

$$R_{go} = \frac{(1 - 2K|S_{12}S_{21}|g_o + |S_{12}S_{21}|^2 g_o^2)^{1/2}}{|1 + g_o(|S_{22}|^2 - |\Delta|^2)|} \tag{11.80}$$

where K is given in (11.68). This circle is called the constant operating power-gain circle and depends on the S-parameters of the device and the normalized operating power gain. For a given device, any load reflection coefficient Γ_L or impedance Z_L on a constant operating power-gain circle (and inside a stable region for potentially unstable devices) can be selected to produce a corresponding operating power gain.

Letting R_{go} from (11.80) equal to zero, corresponding maximum operating power-gain $G_{O,\max}$ or $g_{o,\max}$, leads to

$$|S_{12}S_{21}|^2 g_{o,\max}^2 - 2K|S_{12}S_{21}|g_{o.\max} + 1 = 0 \tag{11.81}$$

Solving (11.81) for $g_{o,\max}$ under the assumption of unconditional stability ($K > 1$ and $|\Delta| < 1$ as seen in Chapter 10) gives

$$g_{o,\max} = \frac{1}{|S_{12}S_{21}|}(K - \sqrt{K^2 - 1}) \tag{11.82}$$

Substituting $g_{o,\max}$ in (11.82) into (11.79) gives the optimum load reflection coefficient that produces the maximum operating power gain $G_{O,\max}$ as

$$\Gamma_{ML} = C_{go,\max} = \frac{g_{o,\max} B_2^*}{1 + g_{o,\max}(|S_{22}|^2 - |\Delta|^2)} \tag{11.83}$$

A design method for unconditionally stable devices similar to that based on the transducer power gain for unilateral devices can be used. For given operating power gains G_O, the constant operating power-gain circles can be drawn using (11.79) and (11.80) on a Smith chart, on which a load reflection coefficient Γ_L or impedance Z_L can be selected to produce a particular gain. Typically, an optimum source reflection coefficient Γ_S or impedance Z_S is chosen according to $\Gamma_S = \Gamma_{\text{in}}^*$, where Γ_{in} is given in (11.52), to produce maximum power delivery to the device.

For potentially unstable transistors, a similar design procedure can also be implemented. The constant operating power-gain circles are drawn using (11.79) and (11.80) for different G_O's, from which Γ_L (or impedance Z_L) can be selected according to a desired gain G_O which must be less than the maximum stable

gain $G_{\text{sta,max}}$ given in (11.69). However, since the device is potentially unstable, the output stability circle with the center and radius given in Eqs. (10.17) and (10.18),

$$C_L = (C_{LR}, C_{LI}) = \frac{(S_{22} - \Delta S_{11}^*)^*}{|S_{22}|^2 - |\Delta|^2} \tag{11.84}$$

and

$$R_L = \left| \frac{S_{12}S_{21}}{|S_{22}|^2 - |\Delta|^2} \right| \tag{11.85}$$

respectively, must also be drawn in the same Smith chart, and the selected Γ_L must lie within the stable region and inside the Smith chart, yet sufficiently far away from the output stability circle to avoid possible shift into the unstable region due to device variations. To determine the source reflection coefficient Γ_S (or impedance Z_S), the input stability circle with the center and radius given respectively in (10.19) and (10.20),

$$C_S = (C_{SR}, C_{SI}) = \frac{(S_{11} - \Delta S_{22}^*)^*}{|S_{11}|^2 - |\Delta|^2} \tag{11.86}$$

and

$$R_s = \left| \frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2} \right| \tag{11.87}$$

is drawn in the same Smith chart and used to check the selection possibility of the reflection coefficient Γ_S (or impedance Z_S). A source reflection coefficient $\Gamma_S = \Gamma_{\text{in}}^*$ according to (11.52) that produces maximum power delivery to the device is typically chosen. This reflection coefficient, however, must lie within the input stability region (and sufficiently far away from the stability circle) to be useful. Otherwise, a different gain for G_O ($<G_{\text{sta,max}}$) must be used and a corresponding Γ_L is selected within the stable region of, yet sufficiently far away from, the output stability circle, from which $\Gamma_S = \Gamma_{\text{in}}^*$ is determined which must fall inside the stable region of the input stability circle and be sufficiently located away from the circle, considering possible degradation of the gain and matching of the amplifier. Depending on transistors, the selected gain G_O may need to be much smaller than the maximum stable gain $G_{\text{sta,max}}$ in order to achieve stability considering possible variations of fabricated devices used in the amplifier.

11.2 LOW NOISE AMPLIFIERS

LNAs are used in RF receivers' front-end to amplify signals from receiving antennas to a certain level while minimizing noise effects on the receivers. Therefore, both gain and noise figure are important in LNA design. In general, design for maximum gain is different from design for minimum noise figure. Other important considerations for LNA design are linearity, which dictates the amplifier's ability to accommodate large signals without distortion and to handle interferences, power consumption, and match to external circuits or components at the LNA's input and output ports. Low power consumption is particularly needed for portable devices. Although linearity is less concerned in LNA's as compared to PAs due to typically weak receiving signals, it should also be considered in the LNA design to avoid possible unwanted nonlinear effects, particularly when there are possibilities of large input signals.

11.2.1 Noise Figure Fundamentals

We begin the analysis by considering a "noisy RF amplifier" immersed within a noisy setting – altogether as a simple "noise model" as shown in Figure 11.5(a). The noise model consists of two noisy components: a "noisy resistor" R_N, representing the input noise to the amplifier from all external elements and environment, and the actual "noisy amplifier" which contains all the internal noise contributed by the amplifier itself. Note

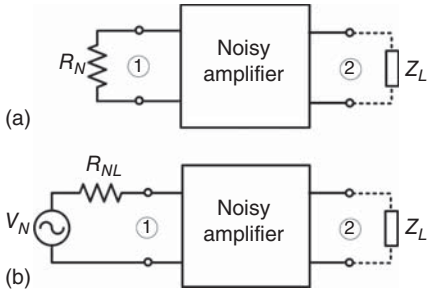


Figure 11.5. A noisy RF amplifier (a) and its equivalence (b).

that Z_L is the amplifier's output termination and does not affect the noise of the (nonterminated) amplifier itself. The total output noise power at port 2 (without considering Z_L) is contributed by the output noise power from R_N (i.e., the amplified input noise power) and the output noise power produced by the amplifier itself. As these noise powers are independent of each other, the amplifier's output noise power is the sum of these noise powers. It is noted that, in practice, there is always noise contributed externally to the amplifier and hence there exists noise power at the amplifier's output even though there is no real signal applied to the amplifier. The design of the amplifier is normally not concerned with this kind of noise, but rather the noise generated internally in the amplifier. The noisy resistor R_N produces thermal or Johnson noise voltage, and so electrically equivalent to a generator having a noise voltage in series with an internal (ideal) noiseless resistor R_{NL} as shown in Figure 11.5(b).

The (rms) noise voltage produced by the resistor R_N is

$$V_N = \sqrt{4kTBR_N} \tag{11.88}$$

where $k = 1.374 \times 10^{-23}$ J/K is the Boltzmann's constant, T is the resistor's noise temperature in kelvin (K),[1] and B is the noise bandwidth in hertz (Hz), which is the absolute RF bandwidth over which the amplifier operates. V_N depends on the bandwidth and is thus considered "white noise." The maximum available noise power that can be generated from R_N is

$$P_N = \frac{V_N^2}{4R_N} = kTB \tag{11.89}$$

making use of (11.88). As can be seen from (11.89) the inherent input noise power to amplifiers operating over wide bandwidths can be sufficiently large that degrades the amplifiers' noise performance significantly.

We now define the noise figure or noise factor of an amplifier as the normalized total output noise power of the amplifier with respect to the (inherent) output white noise power:

$$F \equiv \frac{\text{total output noise power}}{\text{output noise power from } R_N} = \frac{P_{No}}{P_{Ni}G_A} \tag{11.90}$$

where $P_{Ni} = kTB$ is the input noise power and G_A is the available power gain of the amplifier defined as

$$G_A \equiv \frac{P_{So}}{P_{Si}} \tag{11.91}$$

with P_{Si} and P_{So} being the available signal power at the input and output of the amplifier, respectively. Noise figure is typically given in decibels as 10log F. It is important to note that the noise figure of an amplifier characterizes the "noise" of the amplifier by itself regardless of any noise (e.g., noise from R_N) that can possibly enter the amplifier. Equation (11.90) suggests that the noise figure of amplifiers reduces as the gain

[1]The standard noise temperature, corresponding to room temperature, is 290 K.

is increased, so amplifiers should be designed to have high gain in order to lower their noise figure. Note that, as we will see later, maximum gain does not correspond to minimum noise figure in general and so the implication indicated by (11.90) should be taken relatively. The noise figure defined in (11.90) can be rewritten, using (11.91), as

$$F = \frac{P_{Si}/P_{Ni}}{P_{So}/P_{No}} = \frac{\text{signal-to-noise ratio at input}}{\text{signal-to-noise ratio at output}} \tag{11.92}$$

The signal-to-noise (S/N) ratio is a term commonly used to compare the relative strength of a signal with respect to the noise power at a particular location in an electronic component or system. Noise figure is one of the two most important metrics, besides gain, for LNA. Noise figure is also important in other RF components such as mixers and RF systems. As can be inferred from (11.92), it is an important "FOM" for LNA, in particular, and other RF components/systems, in general, and is used to signify possible degradation in the S/N of devices, components or systems. If a device, component or system has no noise then $F = 1$ or 0 dB, and the output noise, and hence S/N, is due entirely to the source driving the device, component or system, the external elements, and the environment. It is also noted that Figure 11.5 can also be used to represent other noisy two-port (or multiport) RF components or subsystems, such as receivers, and hence their noise analysis can be done similarly.

The noise figure of complementary metal oxide silicon (CMOS) RF amplifiers, or of any noisy two-port active RFIC, is contributed by three noise sources with each considered thermal noise produced by an equivalent noisy resistor: the noise source represented by resistance R_N and the gate and drain noises of the employed MOSFETs. This noise figure can be derived as

$$F = F_{\min} + \frac{r_n}{g_s}[(g_s - g_{so})^2 + (b_s - b_{so})^2] \tag{11.93}$$

where $r_n \equiv R_N/Z_o$ is the normalized noise resistance with Z_o being the amplifier's terminating impedance, typically 50 Ω, $y_s \equiv Y_S/Y_o = g_s + jb_s$ is the normalized source admittance with $Y_o = 1/Z_o$, and $y_{so} \equiv Y_{So}/Y_o = g_{so} + jb_{so}$ is the normalized optimum source admittance corresponding to the minimum noise figure $F_{\min}$; that is, when $Y_S = Y_{So}$, $F = F_{\min}$. The admittance Y_S and Y_{So} are the admittances looking into the source from the device's input and relate to the respective reflection coefficient Γ_S and Γ_{So}, as seen in Figure 11.1, as

$$Y_S = \frac{1 - \Gamma_S}{1 + \Gamma_S} \tag{11.94}$$

and

$$Y_{So} = \frac{1 - \Gamma_{So}}{1 + \Gamma_{So}} \tag{11.95}$$

Substituting (11.94) and (11.95) into (11.93), we get

$$F = F_{\min} + \frac{4r_n|\Gamma_S - \Gamma_{So}|^2}{(1 - |\Gamma_S|^2)|1 + \Gamma_{So}|^2} \tag{11.96}$$

r_n, $F_{\min}$, Γ_{So} (or Y_{So}) are known as the noise parameters of MOSFETs and can be measured or calculated. The calculations for these parameters are computationally intensive for complex devices and require accurate parameters of the device physics and geometry. Therefore, in practice, they are typically obtained through measurements of devices. The measured values are normally accurate and can be obtained by varying the source admittance Y_S (using an impedance tuner) and measuring the corresponding F, Y_S, $F_{\min}$ and Y_{So}, and fitting the measured data to certain equations to determine r_n. Noise parameters are typically provided by CMOS foundries or transistor manufacturers. For a given MOSFET, one can obtain these parameters at certain frequency and, from which, Γ_S or Y_S can be determined and used in designing the IMN. Noise figure

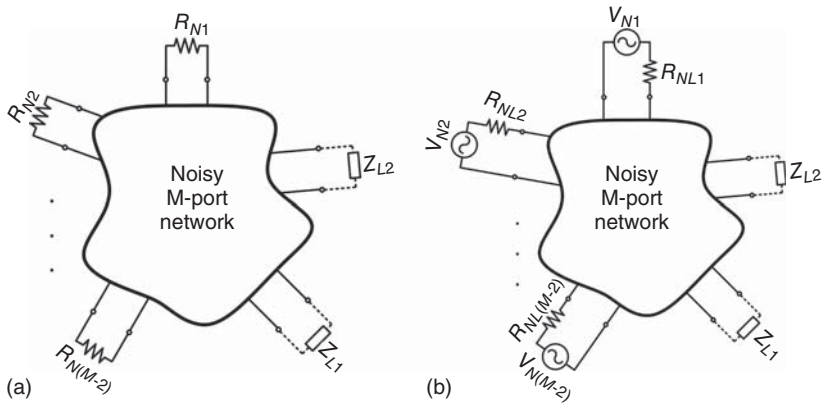


Figure 11.6. A noisy M-port network (a) and its equivalence (b).

depends on frequency for given devices, but in general a noise figure from 2 to 3 dB is considered very well for RF amplifiers, depending on operating frequencies.

It is particularly noted that a noisy M-port network can be reduced to a noisy two-port network with proper terminations at unused ports as shown in Figure 11.6. The noise analysis of the resultant two-port network can thus be carried out similarly to that described for two-port networks with unused ports properly terminated and noise generated by these terminations properly taken into account. The internal noise is the noise generated by the entire terminated network; that is, from elements between the two ports and from others including unused ports' terminations connecting to these two ports. The external noise contributions, as in the two-port case, are those coming to the input port from all external elements and environment. The analysis of multiport RF active components such as mixers can thus be performed by simply extending the two-port analysis.

11.2.2 MOSFET Noise Parameters

The noise parameters of MOSFETs are approximately given by [2]

$$F_{\min} \simeq 1 + \frac{2}{\sqrt{5}} \frac{f}{f_T} \sqrt{\gamma \delta (1 - |c|^2)} \tag{11.97}$$

$$r_N \simeq \frac{\gamma}{\alpha g_{mo}} \tag{11.98}$$

$$Y_{So} \simeq 2\pi f C_{gs} \left[\alpha \sqrt{\frac{\delta}{5\gamma} \left(1 - |c|^2\right)} - j \left(1 + \alpha |c| \sqrt{\frac{\delta}{5\gamma}} \right) \right] \tag{11.99}$$

where

$$\alpha = \frac{g_{mo}}{g_{do}} \tag{11.100}$$

and is equal to 1 for long-channel devices and progressively decreases as the channel length reduces; $c = j0.395$ for long-channel devices and is assumed to be the same for short-channel devices; $g_{mo} = \partial I_{ds} / \partial V_{gs}|_{V_{ds}=\text{constant}}$ is the (DC) transconductance; $g_{do} = 1/R_{dso}$ is the drain–source conductance at $V_{ds} = 0$ V; C_{gs} is the gate–source capacitance; f is frequency; f_T is the device's cut-off frequency; and

$$\gamma = \begin{cases} 1 & \text{for } V_{ds} = 0 \text{ V and long-channel device} \\ 2/3 & \text{for saturation region and long-channel device} \\ 2 - 3 & \text{or larger for short-channel NMOS in saturation} \end{cases} \tag{11.101}$$

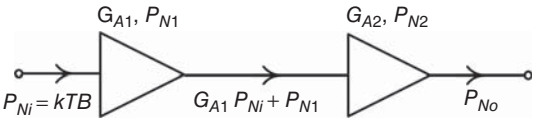


Figure 11.7. Block diagram of a two-stage amplifier or two cascaded amplifiers.

$$\delta = \begin{cases} 2 & \text{for } V_{ds} = 0\,\text{V and long-channel device} \\ 4/5 & \text{for saturation region and long-channel device} \\ 4-6 & \text{or larger for short-channel NMOS in saturation} \end{cases} \tag{11.102}$$

Equation (11.97) shows that the noise figure reduces as the device's cut-off frequency f_T increases, leading to the desire of using MOSFETs with f_T as high as possible or, for a given device, operating frequency much smaller than f_T for achieving noise figure as low as possible. Equations (11.97)–(11.102) are useful for estimating the noise parameters of MOSFETs when they are not provided by the foundry or cannot be measured.

11.2.3 Noise Figure of Multistage Amplifiers

We consider a two-stage amplifier or two amplifiers connected in cascade, as shown in Figure 11.7, where G_{A1}, G_{A2} and P_{N1}, P_{N2} are the available power gains of the first and second stages or amplifiers and the output noise powers produced by the internal noise in these stages or amplifiers, respectively, and P_{Ni} is the input noise or the noise from the equivalent noisy resistor R_N. The total noise power at the output of the amplifier can be written as

$$P_{No} = (G_{A1}P_{Ni} + P_{N1})G_{A2} + P_{N2} \tag{11.103}$$

assuming the input of the first stage and the output of the second stage are perfectly matched and there is also a perfect match between the output of the first stage and the input of the second stage. The noise figure defined in (11.90) can be rewritten, upon replacing G_A with $G_{A1}G_{A2}$ and substituting P_{No} from (11.103), as

$$F = \frac{P_{No}}{P_{Ni}G_{A1}G_{A2}} = 1 + \frac{P_{N1}}{P_{Ni}G_{A1}} + \frac{P_{N2}}{P_{Ni}G_{A1}G_{A2}} \tag{11.104}$$

or

$$F = F_1 + \frac{F_2 - 1}{G_{A1}} \tag{11.105}$$

where

$$F_1 = 1 + \frac{P_{N1}}{P_{Ni}G_{A1}} \tag{11.106}$$

and

$$F_2 = 1 + \frac{P_{N2}}{P_{Ni}G_{A2}} \tag{11.107}$$

are the noise figures of the first and second stage or amplifier, respectively.

Now we consider an N-stage amplifier or N cascaded amplifiers as shown in Figure 11.8. The noise figure can be derived by extending the foregoing two-stage analysis, assuming perfect impedance match at the input and output of the first and last stage, respectively, and between the inter-stages, as

$$F = F_1 + \frac{F_2 - 1}{G_{A1}} + \frac{F_3 - 1}{G_{A1}G_{A2}} + \cdots + \frac{F_N - 1}{G_{A1}G_{A2}G_{A3}\cdots G_{A(N-1)}} \tag{11.108}$$

Figure 11.8. Block diagram of an N-stage amplifier or N cascaded amplifiers.

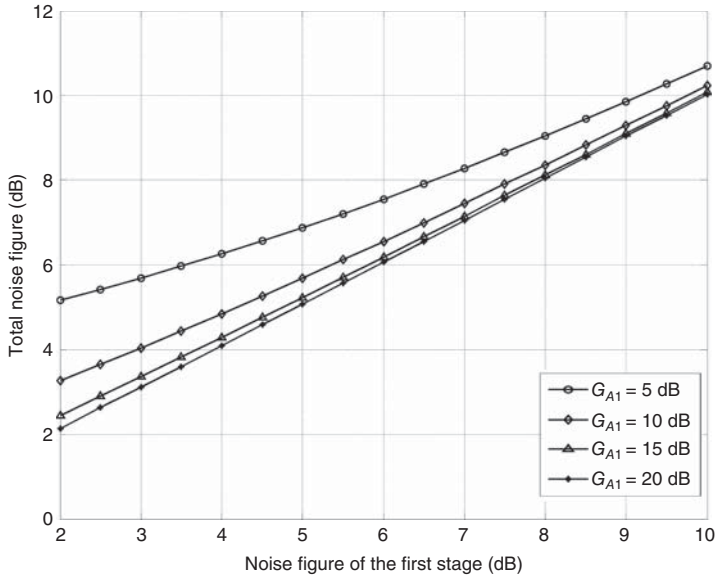


Figure 11.9. Total noise figure of a three-stage RF amplifier as a function of the noise figure and gain of the first stage. Assume the noise figure and gain of the second stage are 8 and 20 dB and those of the third stage are 10 and 25 dB, respectively.

where G_{An} and F_n $(n = 1, 2, \ldots, N)$ are the available power gain and noise figure of the nth stage or amplifier. It is seen in (11.108) that the noise figure of each stage, except that of the first stage, is reduced by 1 (or 0 dB). This result comes from the removal of the noise contribution produced by R_N, which is correct due to the fact that there is only one input noise power from R_N that enters the amplifier via the first stage, provided that perfect impedance match is maintained. The result is also expected since the noise figure of the amplifier should depend only on the noise of its stages, not external noise sources. Equation (11.108) can also be used to determine the noise figure of other components, subsystems, or systems consisting of cascaded networks, for example, an RF receiver. Examination of (11.108) reveals that the overall noise figure is primarily affected by the noise figures of the earlier stages or amplifiers, particularly the first one, provided that they have sufficiently high gain. Therefore, in the design of low noise, high gain amplifiers, it is common to design the first stage for low noise and high gain and the other stages for high gain only. In typical RF receivers, the first component is always an LNA and hence it should be designed for low noise figure and high gain. The high gain of the LNA helps reduce (or even eliminate) the noise contributions of subsequent components to the overall noise figure of the receiver and, hence, relaxing the design requirements for these components. When multiple amplifiers are used in low noise receivers, the first RF amplifier before the mixer or the first IF amplifier after the mixer should have low noise figure and high gain, while high gain is only needed for following amplifiers. Figure 11.9 illustrates this point by showing the overall noise figure of a three-stage RF amplifier versus noise figure of the first stage for various gains of the first stage. As can be seen, the overall noise figure depends strongly on the noise figure and gain of the first stage. The overall noise figure in fact approaches that of the first stage as its gain is increased.

11.2.4 Noise-Figure Design

As seen in (11.96), the noise figure of an amplifier depends on several parameters, one of them being the reflection coefficient Γ_S seen at the input terminal of the transistor or the impedance Z_S presented to that

terminal. Theoretically, for a given MOSFET, Z_S can be chosen to achieve any noise-figure value for the amplifier. We define a noise-figure parameter f_i corresponding to the noise figure F_i and obtain it, utilizing (11.96), as

$$f_i \equiv \frac{|\Gamma_S - \Gamma_{So}|^2}{1 - |\Gamma_S|^2} = \frac{F_i - F_{\min}}{4r_n}|1 + \Gamma_{So}|^2 \tag{11.109}$$

from which, we obtain

$$(\Gamma_S - \Gamma_{So})(\Gamma_S^* - \Gamma_{So}^*) = f_i - f_i|\Gamma_S|^2 \tag{11.110}$$

Expanding (11.81) and making use of

$$\Gamma_{So}\Gamma_S^* + \Gamma_S\Gamma_{So}^* = 2\text{Re}(\Gamma_S\Gamma_{So}^*) \tag{11.111}$$

give

$$|\Gamma_S|^2(1 + f_i) + |\Gamma_{So}|^2 - 2\text{Re}(\Gamma_S\Gamma_{So}^*) = f_i \tag{11.112}$$

Multiplying (11.111) with $(1 + f_i)$ gives

$$|\Gamma_S|^2(1 + f_i)^2 + |\Gamma_{So}|^2 - 2(1 + f_i)\text{Re}(\Gamma_S\Gamma_{So}^*) = f_i^2 + f_i(1 - |\Gamma_{So}|^2) \tag{11.113}$$

and then dividing by $(1 + f_i)^2$, we get

$$|\Gamma_S|^2 + \frac{|\Gamma_{So}|^2}{(1 + f_i)^2} - \frac{2\text{Re}(\Gamma_S\Gamma_{So}^*)}{1 + f_i} = \frac{f_i^2 + f_i(1 - |\Gamma_{So}|^2)}{(1 + f_i)^2} \tag{11.114}$$

which can be rewritten as

$$\left|\Gamma_S - \frac{\Gamma_{So}}{1 + f_i}\right|^2 = \frac{f_i^2 + f_i(1 - |\Gamma_{So}|^2)}{(1 + f_i)^2} \tag{11.115}$$

For a given device, Γ_{So} can be determined and so (11.115) represents a set of circles in the Γ_S-plane with f_i as the parameter. Each value of Γ_S corresponds to a particular noise figure F_i and hence f_i according to (11.109), and so these circles are called "constant noise-figure circles." The centers C_{Fi} and radii R_{Fi} of these circles can be obtained from (11.115) as

$$C_{Fi} = \frac{\Gamma_{So}}{1 + f_i} \tag{11.116}$$

and

$$R_{Fi} = \frac{1}{1 + f_i}\sqrt{f_i^2 + f_i(1 - |\Gamma_{So}|^2)} \tag{11.117}$$

As can be seen from (11.116), the centers of the constant noise-figure circles lie on the line connecting the center of the Smith chart and the point Γ_{So}. When the noise figure is minimum, the noise-figure parameter f_i is equal to zero according to (11.109) and correspondingly, from (11.116) and (11.117), $C_{Fi} = \Gamma_{So}$ and $R_{Fi} = 0$. The resultant noise-figure circle then becomes a point located at Γ_{So}. In order to plot the constant noise-figure circles, r_n, Γ_{So}, and $F_{\min}$ are needed. Figure 11.10 illustrates some constant noise-figure circles obtained for a 0.18-μm MOSFET with $r_n = 27.67\,\Omega$, $\Gamma_{So} = 0.713\angle 61.574°$, and $F_{\min} = 1.069$ dB at 15 GHz. The minimum noise-figure circle is located at $\Gamma_{So} = 0.713\angle 61.574°$. Any point on a particular noise-figure circle, such as the 5-dB circle, would give the corresponding noise figure.

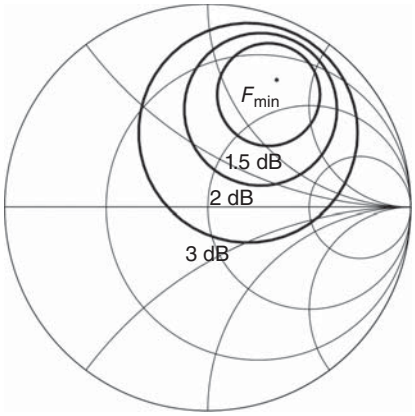


Figure 11.10. Constant noise-figure circles for a 0.18-μm MOSFET.

A procedure for plotting the constant noise-figure circles on a Smith chart for a given device with known r_n, Γ_{So}, and $F_{\min}$ can be derived from the foregoing formulation as:

1. Locate Γ_{So} on a Smith chart and draw a line from the Smith chart's center to Γ_{So}. This center also represents the minimum noise-figure circle corresponding to $F_{\min}$.
2. From the amplifier's noise figure specifications, determine F_i and hence f_i according to (11.109).
3. Calculate the center location C_{Fi} and radius R_{Fi} corresponding to F_i using (11.116) and (11.117), respectively.
4. Draw the constant noise-figure circle centered at C_{Fi} with radius R_{Fi}.
5. Repeat steps 2 to 4 for other F_i's.

It is noted that in the noise-figure design, the required design parameter for achieving a certain noise-figure is Γ_S and hence Z_S, which is used, once determined, to design the IMN to match the MOSFET to the source impedance, typically 50 Ω. Γ_L and hence Z_L in generally does not affect the amplifier's noise figure, as implied in (11.93), and so they are only selected to provide the maximum gain $G_{L,\max}$ corresponding to Γ_S that gives the required noise figure. For minimum noise figure, $\Gamma_S = \Gamma_{So}$ and hence Γ_L can be determined from (11.8) as

$$\Gamma_L = \Gamma_{\text{out}}^* = \left(S_{22} + \frac{S_{12}S_{21}\Gamma_{So}}{1 - S_{11}\Gamma_{So}} \right)^* \tag{11.118}$$

The OMN for the desired Γ_L can then be designed to transform the transistor's output impedance to the load impedance, which is also normally 50 Ω. It should be noted that for unconditionally stable devices, any Γ_S and Γ_L can be chosen corresponding to the required noise figure and gain, respectively. On the other hand, if the device is potentially unstable, Γ_S must be selected within the stable region of the input (or source) stability circle before Γ_L is selected, which must also be in the stability region of the output (or load) stability circle.

11.2.5 Design for Gain and Noise Figure

A common design requirement for LNAs is simultaneously achieving both high gain and low noise figure. Such design for LNAs employing unilateral or bilateral devices can be done with Smith charts using the following procedures utilizing both constant gain and noise-figure circles.

For Unilateral Devices

1. Plot constant noise-figure and constant transducer gain (G_S) circles in the Γ_S-plane on a Smith chart. For potentially unstable devices, also plot the input stability circle in the same Smith chart.
2. Choose Γ_S or Z_S inside the Smith chart to achieve a desired noise figure or a compromise between noise figure and G_S. For potentially unstable devices, make sure the selected Γ_S or Z_S lies within the stable region and is sufficiently far away from the input stability circle.
3. Calculate Γ_{out} using (11.8) with the device's S-parameters and chosen value of Γ_S or Z_S.
4. Design an IMN to realize Γ_S or Z_S.
5. Choose Γ_L or Z_L according to $\Gamma_L = \Gamma_{\text{out}}^*$ for maximum gain and design a corresponding OMN. For potentially unstable devices, also plot output stability circle in the same Smith chart and make sure the selected Γ_L or Z_L lies within the stable region and is sufficiently far away from the output stability circle. If not, reselect Γ_L or Z_L to produce stability while considering the resultant gain and matching.

For Bilateral Devices

1. Plot constant noise-figure circles and constant operating power-gain circles on a Smith chart. For potentially unstable devices, also plot the input and output stability circles in the same Smith chart.
2. Choose Γ_L or Z_L inside the Smith chart to achieve a desired noise figure or a compromise between noise figure and gain. For potentially unstable devices, make sure the selected Γ_L or Z_L lies within the corresponding stable region and is sufficiently far away from the output stability circle.
3. Choose Γ_S or Z_S according to $\Gamma_S = \Gamma_{\text{in}}^*$. For potentially unstable devices, make sure the selected Γ_S or Z_S lies within the corresponding stable region and is sufficiently far away from the input stability circle. Otherwise, a different value for the operating power gain must be used and a corresponding Z_L is selected, from which $\Gamma_S = \Gamma_{\text{in}}^*$ is determined which must fall inside the stable region of the input stability circle and be sufficiently located away from the circle, considering possible degradation of the gain and matching of the amplifier.
4. Design the IMNs and OMNs to realize Γ_S or Z_S and Γ_L or Z_L, respectively.

The above procedure for bilateral devices is based on the operating power gain. A similar procedure can also be implemented using the available gain.

11.3 DESIGN EXAMPLES

In this section, we show two amplifier design examples: one based on the unilateral design procedure for maximum (unilateral) gain and another one on bilateral design. These designs use the following S-parameters of a 0.18-μm MOSFET at 15 GHz:

S_{11}	S_{12}	S_{21}	S_{22}
$0.844\angle -70.545°$	$0.153\angle 41.004°$	$2.302\angle 123.372°$	$0.583\angle -54.216°$

11.3.1 Unilateral Amplifier Design

In this design, we assume the device is unilateral, even though it is actually bilateral at 15 GHz as can be seen from $|S_{12}| = 0.153$. Accordingly, the design proceeds with the selection of the source and load reflection coefficients for maximum unilateral gain as

$$\Gamma_S = S_{11}^* = 0.844\angle 70.545°$$
$$\Gamma_L = S_{22}^* = 0.583\angle 54.216°$$

The maximum unilateral gain is obtain from (11.25) as

$$G_{u,\max} = \frac{1}{1-|S_{11}|^2}|S_{21}|^2\frac{1}{1-|S_{22}|^2} = 27.9 = 14.45 \text{ dB}$$

Although the magnitudes of S_{11} and S_{22} are less than 1, the device may be potentially unstable. To check the stability, we calculate the stability parameters according to Eqs. (10.3) and (10.4) as

$$|\Delta| = S_{11}S_{22} - S_{12}S_{21} = 0.5025$$

and

$$K = \frac{1-|S_{11}|^2-|S_{22}|^2+|\Delta|^2}{2|S_{12}S_{21}|} = 0.2844$$

Since $K<1$, the device is potentially unstable and we need to examine the stability circles. Accordingly, we calculate the centers and radii for the input and output stability circles at 15 GHz using Eqs. (10.19), (10.20) and (10.17), (10.18) as

$$C_S = (C_{SR}, C_{SI}) = \frac{(S_{11}-\Delta S_{22}^*)^*}{|S_{11}|^2-|\Delta|^2} = 1.422\angle 87.8^\circ$$

$$R_s = \left|\frac{S_{12}S_{21}}{|S_{11}|^2-|\Delta|^2}\right| = 0.766$$

and

$$C_L = (C_{LR}, C_{LI}) = \frac{(S_{22}-\Delta S_{11}^*)^*}{|S_{22}|^2-|\Delta|^2} = 4.42\angle 100.9^\circ$$

$$R_L = \left|\frac{S_{12}S_{21}}{|S_{22}|^2-|\Delta|^2}\right| = 4.03$$

respectively. Figure 11.11 shows these stability circles at 15 GHz. As $|S_{11}|<1$ and $|S_{22}|<1$, the stable regions are outside of the stability circles and inside the Smith chart. As displayed on Figure 11.11, the source reflection coefficient Γ_S (and hence source impedance Z_S) is located within the unstable region of the transistor, while the load reflection coefficient Γ_L (and hence load impedance Z_L) is in the stable region but very close to the output stability circle. Therefore, the amplifier with these source and load impedances may not be stable under maximum gain at 15 GHz. To assure stability, Γ_S and Γ_L must be located in the corresponding stable regions. Under this condition, however, the maximum gain of 14.45 dB as calculated earlier cannot be achieved. The problem of not being able to achieve the maximum unilateral gain at 15 GHz is due to the fact that the transistor is bilateral. The calculated unilateral FOM according to (11.50) is

$$\text{FOM}_A = \frac{|S_{12}S_{21}S_{11}S_{22}|}{(1-|S_{11}|^2)(1-|S_{22}|^2)} = 1.078$$

which is large, indicating that the assumption of unilateral device is clearly not valid. As such, we need to consider the employed transistor as a bilateral device and design the amplifier using a bilateral procedure as described in the next example in Section 11.3.2.

For the sake of showing a unilateral design, we continue with the amplifier design to achieve stability. As we are not concerned with the noise figure here, we only need to choose certain gains for the input and output gain blocks. It is noted that the constant-gain circles corresponding to these gains must not lie completely in the unstable regions. Part of these circles needs to be located within the stable region to allow proper selections of Γ_S and Γ_L that provide stability. The input and output constant-gain circles can be plotted using

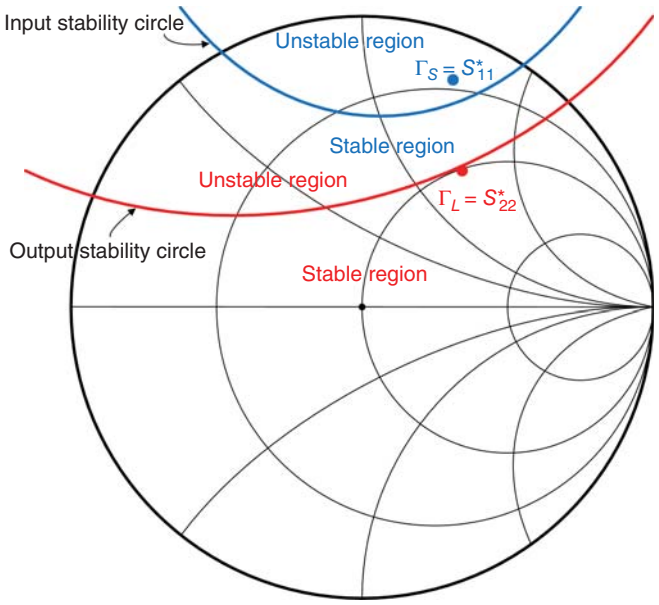


Figure 11.11. Stability circles.

(11.39)–(11.41) and (11.29) for different gains. Selecting a normalized gain g_{Su} of 0.75, we can calculate the center $\Gamma_{Suo}(U_{Suo},V_{Suo})$ and radius R_{Su} of the corresponding input gain circle as

$$U_{Suo} = \frac{g_{Su}A_{11}}{1-|S_{11}|^2(1-g_{Su})} = 0.243$$

$$V_{Suo} = -\frac{g_{Su}B_{11}}{1-|S_{11}|^2(1-g_{Su})} = 0.689$$

and

$$R_{Su} = \frac{\sqrt{1-g_{Su}}(1-|S_{11}|^2)}{1-|S_{11}|^2(1-g_{Su})} = 0.227$$

Similarly, by selecting $g_{Lu} = 0.7$, we can determine the center and radius of the corresponding output gain circle as

$$U_{Luo} = \frac{g_{Lu}A_{22}}{1-|S_{22}|^2(1-g_{Lu})} = 0.266$$

$$V_{Luo} = -\frac{g_{Lu}B_{22}}{1-|S_{22}|^2(1-g_{Lu})} = 0.369$$

and

$$R_{Lu} = \frac{\sqrt{1-g_{Lu}}(1-|S_{22}|^2)}{1-|S_{22}|^2(1-g_{Lu})} = 0.403 = 0.227$$

Figure 11.12 shows these circles along with the input and output stability circles on a Smith chart.

To proceed with the design, we choose Γ_S and Γ_L within the stable regions and sufficiently far away from the input and output stability circles, respectively, as shown in Figure 11.12, as

$$\Gamma_S = 0.75\angle 69^\circ$$

$$\Gamma_L = 0.32\angle 37^\circ$$

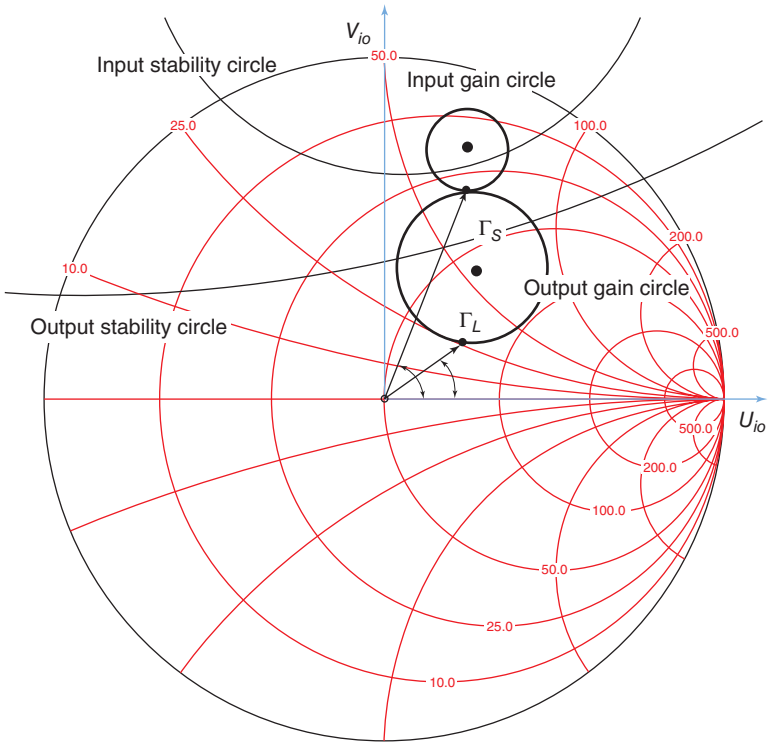


Figure 11.12. Input and output gain circles on a Smith chart.

The IMNs and OMNs to match 50 Ω to Z_S and Z_L, respectively, can be designed following the design procedure discussed in Section 6.2. The gain of the amplifier can be calculated from (11.17) as 13.55 dB which is about 1 dB lower than the maximum unilateral gain.

11.3.2 Bilateral Amplifier Design

In this design, we consider the device as bilateral as indicated by its nonzero S_{12} and follow the design procedure described in Section Design Based on Operating Power Gain for potentially unstable transistors. We first calculate the maximum stable gain given in (11.69) as

$$G_{\text{sta,max}} = \left|\frac{S_{21}}{S_{12}}\right| = 15.07 = 11.78\,\text{dB}$$

The amplifier's operating gain, G_O, should be smaller than 11.78 dB to maintain stability. To determine a proper load (Γ_L) and source (Γ_S) reflection coefficients, we plot the constant operating power-gain circles for various gains, output stability circle, and input stability circle on the same Smith chart, and examine possible locations for Γ_L and Γ_S. From this analysis, we see that when G_O is larger than 10 dB, the source reflection coefficient obtained from $\Gamma_S = \Gamma_{\text{in}}^*$ in (11.52) is always inside the unstable region of the input stability circle, while for 6 dB $< G_O <$ 10 dB, $\Gamma_S = \Gamma_{\text{in}}^*$ is very close to the input stability circle and hence is not suitable for the amplifier. Figure 11.13 shows the results for G_O of 6, 10, and 11.5 dB. In this figure, for each operating gain, Γ_L is chosen on the gain circle and inside the stable region of the output stability circle, while being away from the circle, from which $\Gamma_S = \Gamma_{\text{in}}^*$ is calculated. The amplifier, therefore, cannot be designed to achieve a stable gain of more than 10 dB. In fact, the gain should be kept below 6 dB for stability guarantee.

Based on the above analysis, we choose $G_O = 5.5$ dB. Figure 11.14 shows the 5.5-dB operating power-gain circle, input and output stability circles, and selected locations for Γ_L (and hence Z_L) and Γ_S (and hence Z_S). Γ_L is first chosen as $\Gamma_L = 0.695\angle -45.45°$, which results in $Z_L = 51 - j97.5\ \Omega$. Γ_S is then determined as

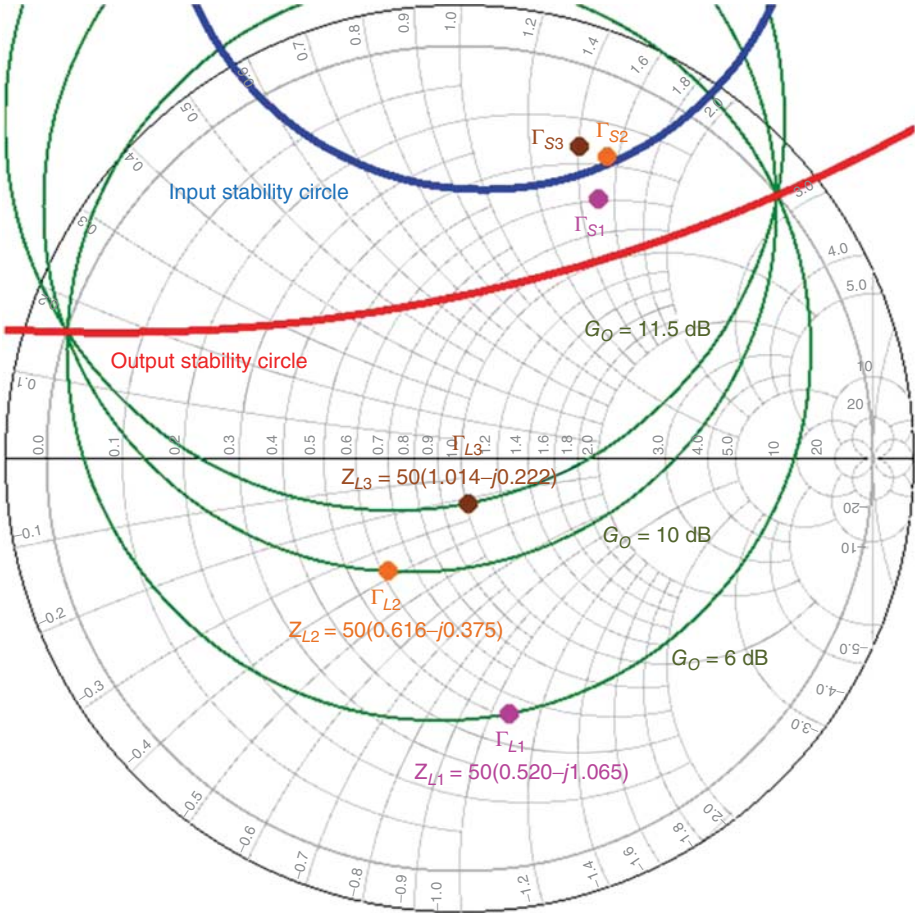


Figure 11.13. Constant operating power-gain circles for $G_O = 6$, 10, and 11.5 dB, stability circles, and possible locations for Γ_{Si}, Z_{Si} and Γ_{Li}, Z_{Li}, where $i = 1$, 2, and 3 corresponds $G_O = 6$, 10, and 11.5 dB, respectively.

$\Gamma_S = \Gamma_{\text{in}}^* = \left(S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1-S_{22}\Gamma_L}\right)^* = 0.635\angle 66.67°$ which gives $Z_S = 33.15 + j64.75\ \Omega$. Figure 11.15 shows the amplifier's schematic with the input and output lumped-element matching networks, designed using the procedure in Section 6.2, that match 50 Ω to Z_S and Z_L, respectively.

Figure 11.16 shows the calculated gain of the designed amplifier. The gain is 5.46 dB which is close to the designed value of 5.5 dB. Figure 11.17 shows the real parts of the impedances looking into the gate (Z_{in}) and drain (Z_{out}) of the MOSFET versus frequency. These are all positive from below 1 GHz to more than 20 GHz, indicating that the designed amplifier is stable at these frequencies.

11.4 POWER AMPLIFIERS

Low PAs, such as the LNA covered in Section 11.2, are designed to operate under small driving signals and are classified as small-signal amplifiers. PAs, on the other hand, are designed to work under large-signal conditions and hence can be referred to as large-signal amplifiers. They are considered the most important component in transmitters.

11.4.1 Power-Amplifier Parameters

In addition to the two fundamental parameters, gain and voltage standing wave ratio (VSWR), typically used for amplifiers, PAs are characterized by other parameters including linearity, power handling, DC-RF

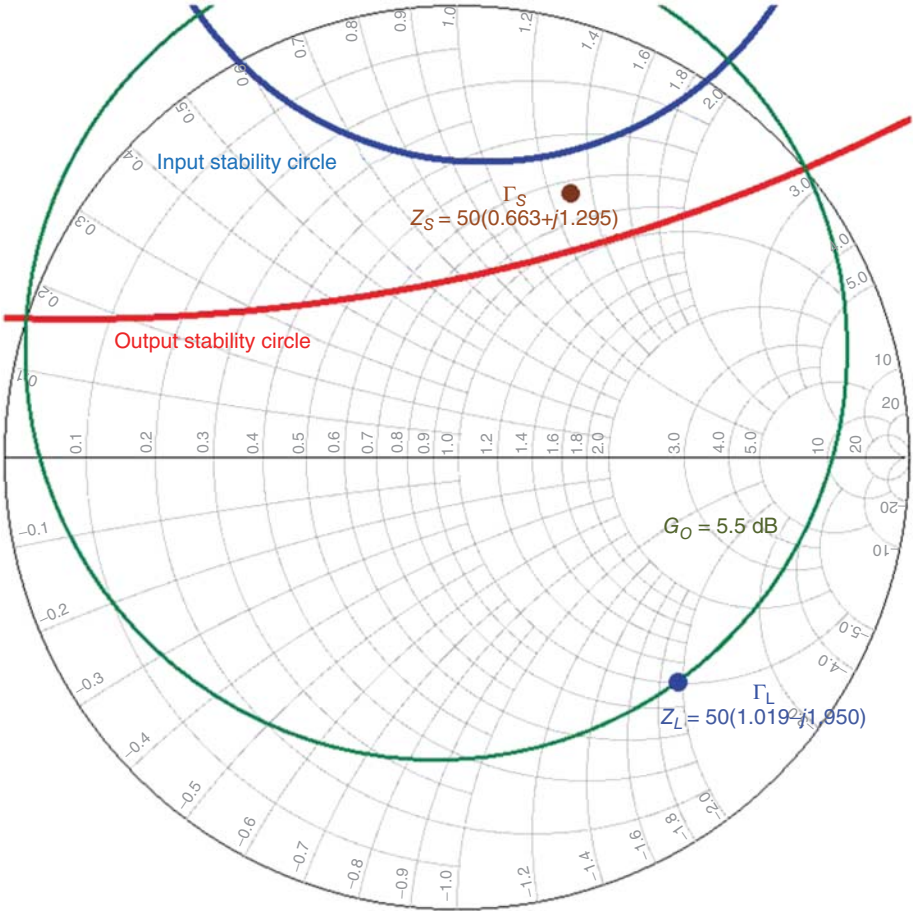


Figure 11.14. Constant operating power-gain circle for $G_O = 5.5$ dB, stability circles, and chosen locations for Γ_S, Z_S and Γ_L, Z_L.

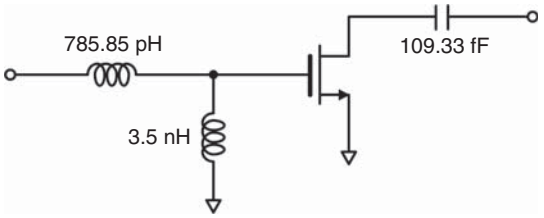


Figure 11.15. Amplifier schematic. Bias networks are not shown.

conversion efficiency and power-added efficiency. Since PAs are typically operated with high DC current levels, their efficiency is a very important design consideration, particularly for portable devices. Distortion, which is related to linearity, is also very important due to typically large input signals for PAs.

The DC-RF conversion efficiency measures the efficiency in converting DC power to RF power and is defined as

$$\eta_{\text{dc}} = \frac{P_{\text{out}}}{P_{\text{dc}}} \tag{11.119}$$

where P_{out} and P_{dc} represent the RF output power and (input) DC power, respectively. As can be seen, this efficiency parameter ignores the presence of RF input signal and, while useful in classifying different classes of PAs, is not a "true" efficiency FOM describing a particular PA under operation.

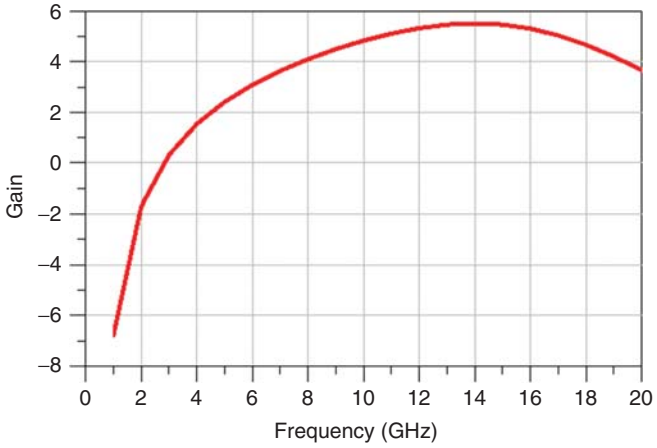


Figure 11.16. Amplifier gain.

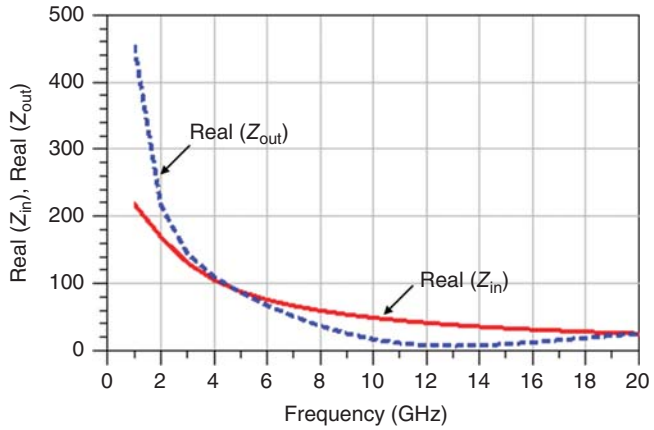


Figure 11.17. Real parts of Z_{in} and Z_{out}.

The power-added efficiency measures the efficiency in converting DC power into RF power generated by the amplifier and is defined as

$$\eta_a = \frac{P_{out} - P_{in}}{P_{dc}} \tag{11.120}$$

where P_{in} is the RF input power. Dividing the numerator and denominator in (11.120) by P_{out} and utilizing (11.119) gives

$$\eta_a = \left(1 - \frac{1}{G}\right)\eta_{dc} \tag{11.121}$$

where $G = P_{out}/P_{in}$ is the transducer power gain defined in (11.3). The power-added efficiency takes into account the ability of PAs in amplifying RF input signals as well as the contribution of supplied DC power to RF output power, and thus describes the amplifiers' efficiency characteristic more proper than the DC-RF conversion efficiency parameter. As can be seen, the power-added efficiency increases when both the gain and DC-RF efficiency are boosted up and it provides a tradeoff between the gain and DC-RF efficiency. However, since it involves amplifiers operating under a particular RF signal condition, it cannot be used to specify different types of PAs and is thus not used in comparing various power-amplifier classes.

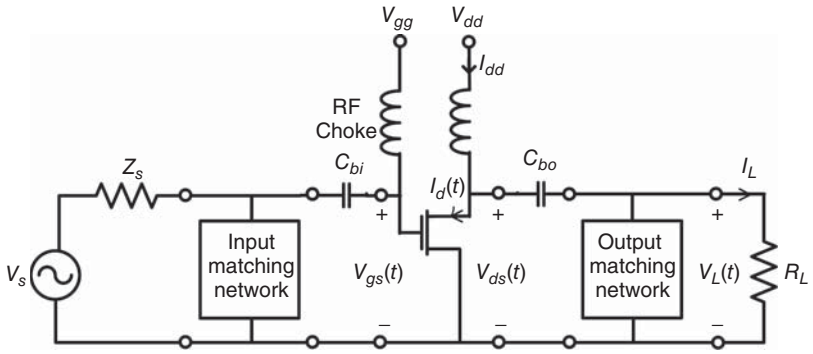


Figure 11.18. A general power amplifier topology.

11.4.2 Power-Amplifier Types

PAs are classified in different classes such as A, B, C, D, E, F, and AB depending on their operation. Among them, Class A, B, and C amplifiers are the fundamental types from which other variants are derived.

Consider a general PA as shown in Figure 11.18 which consists of a MOSFET, DC blocking capacitors, IMNs and OMNs, and RF chokes for biasing networks. The OMN is assumed to behave as a perfectly matched lossless band-pass filter passing only signals at fundamental frequencies and reject all harmonics. We also assume that the output blocking capacitor is lossless the transistor has no resistive parasitics at the drain and source terminals. Regardless of the amplifier's matching and the transistor's impedance characteristics, the power generation capability and efficiency of the amplifier is controlled by the transistor itself which depends upon its bias condition, specifically the drain bias voltage (V_{dd}) for a given gate bias[2] (V_{gg}). The bias condition specifies the region where the output current at the drain terminal lies: cut-off (or off) region, active region, or saturation region. The cut-off region corresponds to the transistor being biased at cut-off; that is when the gate–source bias voltage V_{gg} is less than the threshold voltage V_t. Under this bias condition, the transistor behaves approximately as an open circuit with nearly no current flow across the drain and source. The saturation region is the region corresponding to the transistor being biased beyond cut-off. In this region, the transistor behaves approximately as a resistor having a small resistance R. This results when the drain–source bias voltage V_{dd} is less than I_dR, where I_d is the drain or drain–source current. It is noted that the saturation region referred to here is different from the saturation used in typical descriptions of CMOS transistors such as that in Chapter 9, in which the saturation refers to the region where the drain current is almost constant as the drain–source bias voltage is increased. The active region corresponds to the transistor being biased between cut-off and saturation, which results when the applied gate–source bias voltage is larger than the threshold voltage and the drain–source bias voltage is greater than I_dR. In this region, the drain current is approximately proportional to $(V_{gs} - V_t)$ as discussed in Chapter 9. A PA is classified as Class A, B, or C amplifier when the output current is in the active, cut-off or saturation region, respectively. As the transistor's bias condition affects its efficiency, Class A, B, and C amplifiers have different levels of efficiency.

Under RF operation, we assume a continuous sinusoidal signal[3] is applied to the gate of the MOSFET in the amplifier shown in Figure 11.18. The transistor amplifies the input signal at the fundamental frequency and generates other signals at the harmonic frequencies. The (drain) current flowing into the drain of the transistor consists of two components: the DC (input) bias current and the RF (output) current. The RF current relates to the RF (output) voltage at the drain. While the RF voltage is a sinusoidal waveform, assuming no distortion occurred in the amplification process, the RF current can be sinusoidal or disrupted sinusoidal depending on the transistor's operating point, as will be seen later for different amplifier classes. However, points formed by pairs of their corresponding discrete values lie on a curve known as the AC load line which characterizes the

[2]These are the bias voltages across drain–source and gate–source, respectively, assuming the RF chokes have no resistance and the transistor has no resistances at its terminals.

[3]For driving signals of different waveforms, such as square wave, the basics of the discussion presented here still apply. However, the results for parameters such as power and efficiency are different to those for sinusoidal waveforms.

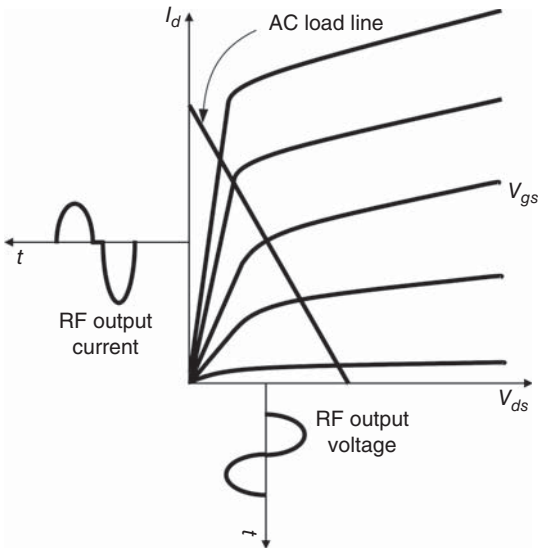


Figure 11.19. Typical MOSFET I–V curves and AC load line. I_d and V_{ds} are the current and voltage at the drain, respectively, and V_{gs} is the gate control voltage. The RF output voltage ($V_{ds,\mathrm{rf}}$) is assumed sinusoidal and the RF output current ($I_{d,\mathrm{rf}}$) is assumed as a disrupted sinusoid.

relation between the RF voltage at the drain (and hence the RF voltage at the load) and the corresponding RF current. This load line is useful in specifying the transistor's operation state. Figure 11.19 illustrates a typical AC load line superimposed over the I–V curves of a MOSFET. The operating point of the transistor, which dictates its operating condition, must satisfy both the I–V characteristics and the AC load line, and hence must lie at the intersection of these lines. It is particularly noted that the I–V curves and the load line are only valid in their respective DC and RF domain, and so the coordinates of the operating points should be read in DC or RF values when considering the I–V curves or load line, respectively. Therefore, strictly speaking, it is not suitable to use these together or, in another word, the operating points cannot satisfy both the I–V curves and load line simultaneously. In order for this to happen, the frequency must be sufficiently low that the DC characteristics of the transistor are approximately valid. At RF frequencies, this assumption, however, does not hold well. Nevertheless, the combined load-line-I–V-curves as shown Figure 11.19 can be used to describe qualitatively the operation of a transistor with respect to its I–V and RF characteristics. For the purpose of studying the characteristics of power-amplifier classes regardless of their operating frequencies, we can assume the frequencies are substantially low, approaching DC, and, under this assumption, both the AC load line and I–V curves can be used simultaneously.

11.4.2.1 Class-A Amplifier. Class A PA operates essentially as a linear PA, in which the transistor functions in the (linear) active region. Class A PA's operating condition is thus established by setting the bias condition for the transistor so that it operates between the cut-off and saturation.

Under an RF sinusoidal excitation, the RF portions of the drain voltage and current at the fundamental frequency arrive completely at the load resistor after passing through the (assumed lossless) output blocking capacitor and MN. For Class A amplifiers not operated under very large input RF signals, harmonics generated are not significant and do not affect much the power and efficiency of the amplifiers; hence we can neglect harmonics in the drain voltage and current. We also assume that the distortion is negligibly small that does not alter the RF output waveforms. Under this consideration, we can write the total drain current–voltage and current as

$$V_{ds}(t) = V_{ds,\mathrm{dc}} + V_{ds,\mathrm{rf}}(t) = V_{dd} + V_{dm}\sin\omega_o t \tag{11.122}$$

$$I_d(t) = I_{d,\mathrm{dc}} + I_{d,\mathrm{rf}}(t) = I_{dd} - I_{dm}\sin\omega_o t \tag{11.123}$$

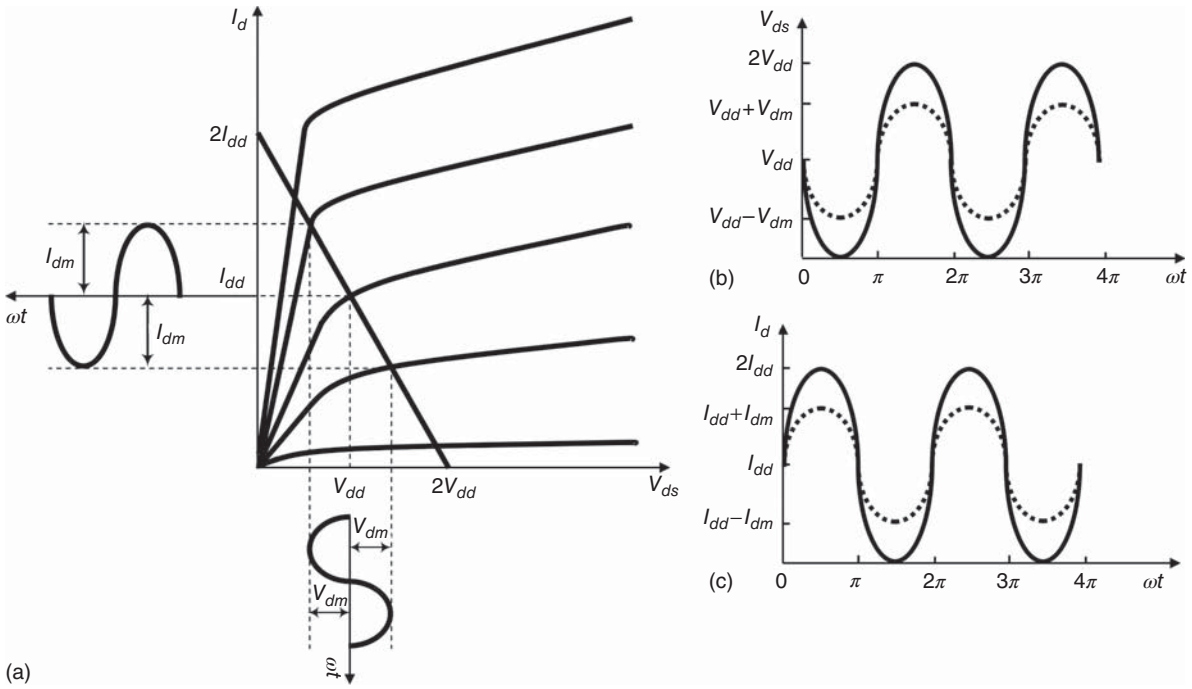


Figure 11.20. Operating characteristics of the Class A amplifier: I–V curves superimposed on the AC load line (a); drain voltage (b); and drain current (c). The dashed curves in (b) and (c) represent the general waveforms and the solid curves denote the waveforms corresponding to the maximum voltage ($2V_{dd}$) and current ($2I_{dd}$) swings.

where the first parts, V_{dd} and I_{dd}, are the (constant) DC bias voltage and current, respectively, and the second parts, $V_{d,\mathrm{rf}} = V_{dm} \sin \omega_o t$ and $I_{d,\mathrm{rf}} = -I_{dm} \sin \omega_o t$, are the RF voltage at and the RF current flowing into the drain, respectively. The transistor is thus operated essentially as a current source $I_d(t)$ constantly conducting current during 360° spanning each cycle, and hence over the entire operation time. Its conduction angle is thus 360°. The RF voltage and current are related by

$$V_{d,\mathrm{rf}}(t) = V_L(t) = -I_{d,\mathrm{rf}}(t) R_L \tag{11.124}$$

where $V_L(t)$ is the voltage across the load resistor R_L. The waveforms of both the RF voltage and current are sinusoidal with 180° out of phase. The AC load line, describing $I_{d,\mathrm{rf}}$ as a function of $V_{ds,\mathrm{rf}}$, has a slope equal to $-1/R_L$. Figure 11.20 shows the AC load line superimposed on the I–V curves along with the RF voltage and current waveforms, illustrating the operation of the Class A amplifier. The voltage and current waveforms represent the respective RF components in (11.122) and (11.123). In these waveforms, we assume the voltage and current vary from zero to $2V_{dd}$ and $2I_{dd}$, respectively. However, the maximum voltage ($2V_{dm}$) is always slightly less than $2V_{dd}$ to avoid saturation for the transistor and the maximum current ($2I_{dm}$) is always slightly smaller than $2I_{dd}$ due to the "knee" naturally occurred on the I–V curve at the saturation region. As the voltage and current waveforms vary symmetrically around V_{dd} and I_{dd}, respectively, their minimum is therefore slightly greater than zero.

The RF output power delivered to the load resistor can be written as

$$P_{\mathrm{rf}} = \frac{V_{dm}^2}{2R_L} \tag{11.125}$$

The DC power supplied to the drain, or the DC input power, is always constant regardless of the RF input signal and is obtained as

$$P_{\mathrm{dc}} = V_{dd} I_{dd} = \frac{V_{dd}^2}{R_L} \tag{11.126}$$

As the DC bias voltage V_{dd} is always different from zero, the DC input power is finite and dissipated into heat all the time, even when the transistor is not excited with an RF signal, hence degrading the amplifier's efficiency. As mentioned earlier, for linear operation, the transistor must be operated between the cut-off and saturation. To prevent it from cutting off and reaching saturation, the RF current and voltage amplitudes must satisfy

$$\begin{aligned} I_{dm} &\leq I_{dd} \\ V_{dm} &\leq V_{dd} \end{aligned} \tag{11.127}$$

The RF output power delivered to the load thus reaches a maximum when $V_{dm} = V_{dd}$ and $I_{dm} = I_{dd}$, or the peak instantaneous drain voltage ($V_{ds,\max}$) and current ($I_{d,\max}$) are $2V_{dd}$ and $2I_{dd}$, respectively. This maximum power is therefore obtained as

$$P_{\text{rf,max}} = \frac{1}{2} V_{dd} I_{dd} = \frac{1}{8} V_{ds,\max} I_{d,\max} \tag{11.128}$$

The RF output power can be rewritten from (11.125) as

$$P_{\text{rf}} \leq \frac{V_{dd}^2}{2R_L} \tag{11.129}$$

The maximum RF output power is obtained when the transistor is biased at (I_{dd},V_{dd}) lying about half-way between the cut-off and saturation on the AC load line, as seen in Figure 11.20(a). The DC-RF conversion efficiency can now be obtained from (11.126) and (11.129) as

$$\eta_{\text{dc}} = \frac{P_{\text{rf}}}{P_{\text{dc}}} \leq \frac{1}{2} \tag{11.130}$$

The maximum DC-RF conversion efficiency is thus merely 50%,[4] signifying that as least 50% of the DC power is dissipated into heat instead of being converted into RF power. This relatively poor efficiency and large power dissipation are the main drawbacks of Class A amplifiers, particularly making them unsuitable for portable wireless devices. In reality, this maximum efficiency, however, never reaches 50% due to the fact that the maximum RF voltage (V_{dm}) and current (I_{dm}) are always slightly smaller than the DC supply voltage (V_{dd}) and current (I_{dd}), respectively. Practical Class A PAs even has DC-RF conversion efficiency much less than 50% due to various factors and unavoidable effects such as loss and mismatch of the OMN, loss of other elements in the output such as blocking capacitor, and transistor's resistive parasitics and nonlinear effects.

Neglecting the RF input power and considering only the RF output power P_{rf} with respect to the DC input power P_{dc}, we can define the average power dissipation in the transistor as the difference between P_{dc} and P_{rf}, and obtain from (11.125) and (11.126):

$$P_d \equiv P_{\text{dc}} - P_{\text{rf}} = \frac{1}{R_L} \left(V_{dd}^2 - \frac{V_{dm}^2}{2} \right) \tag{11.131}$$

Figure 11.21 illustrates the relation between the DC-RF conversion efficiency η_{dc}, P_d, P_{rf}, and P_{dc}, and their behavior as a function of the RF output voltage swing V_{dm}. While the DC power always remains constant for a supplied DC voltage V_{dd}, as noted earlier, the efficiency and RF power increases and the power dissipation decreases as the drain voltage swing is increased, reaching a maximum efficiency of 50% and a maximum power of $V_{dd}^2/2R_L$ at $V_{dm} = V_{dd}$. It is particularly noted that the efficiency depends substantially on the amplitude of the drain voltage swing – the higher the swing the better the efficiency. It is apparent that the Class A amplifiers always dissipate power regardless of the driving RF input signal as expected from its operating

[4]The maximum efficiency is different for different driving waveforms. For instance, for square waves, the efficiency can approach 100%.

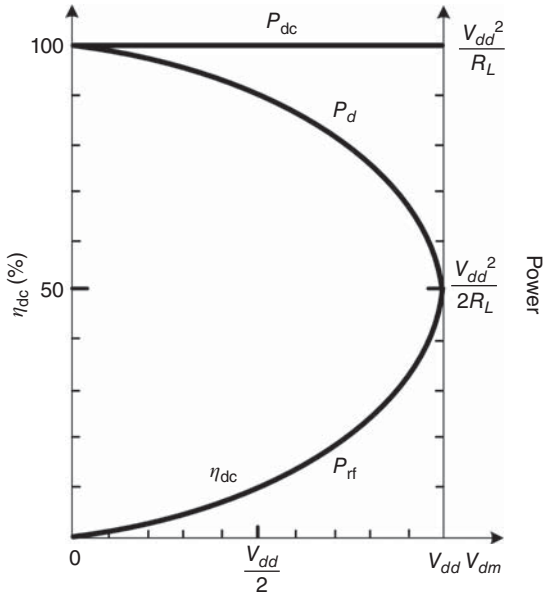


Figure 11.21. Characteristics of DC-RF conversion efficiency, power dissipation, RF output power, and DC input power in Class A PAs.

characteristics and mentioned earlier. In fact, the power dissipated is maximized when there is no RF input signal at all. This indicates a danger for Class A operation as the dissipated power may be greater than the transistor's maximum allowable power dissipation and hence causing damage to the device. Large power consumption when there is no RF input signal is perhaps the major problem with Class-A amplifiers for most practical applications, where "standby" periods occur frequently. In some devices such as cellular phones, intermittent RF input signals due to typical conversations are inevitable. For devices needed to be "on" at all time, the all-time power dissipation is absolutely undesirable, as this will shorten the operating hours of batteries and increase the chance of failure for not only the amplifier transistors, but also other semiconductors contained in the devices, due to excess heat. Due to their low efficiency and large power consumption, Class A amplifiers are thus rarely used in practice for RF applications, particularly for portable wireless devices. However, if Class A amplifiers are employed in RF systems, they should only be used for small-power driver amplifiers or initial stages of multistage PAs with low supply voltages, whose power dissipation is sufficiently low to be conveniently handled. Furthermore, slightly offsetting bias conditions from that for ideal Class-A amplifiers are typically used to improve the amplifier performance.

11.4.2.2 Class-B Amplifier. As for Class A, Class-B PA operates as a linear PA, in which its output signal is a linear function of the input signal. In Class-B amplifiers, the gate–source bias voltage V_{gg} is set to be less than the threshold voltage V_t (typically $V_{gg} = V_t$), making the transistor operate essentially at cut-off with zero drain–source current. Under RF sinusoidal excitations, the transistor in Class-B amplifiers, however, is operated alternately between on and off in each half of the driving signal's sinusoidal cycle. Another word, the transistor only conducts current within one half or 180° of each cycle, and hence its conduction angle is only 180°.

Under ideal conditions with RF sinusoidal excitations, the total voltage and current at the drain can be written as

$$V_{ds}(t) = V_{ds,\mathrm{dc}} + V_{ds,\mathrm{rf}}(t) = V_{dd} + V_{dm}\sin\omega t \tag{11.132}$$

$$I_d(t) = I_{d,\mathrm{rf}}(t) = \begin{cases} I_{dm}\sin\omega t, & 2(n-1)\pi \le \omega t \le (2n-1)\pi \quad \text{for} \quad n = 1, 2, \ldots, N \\ 0, & \text{otherwise} \end{cases} \tag{11.133}$$

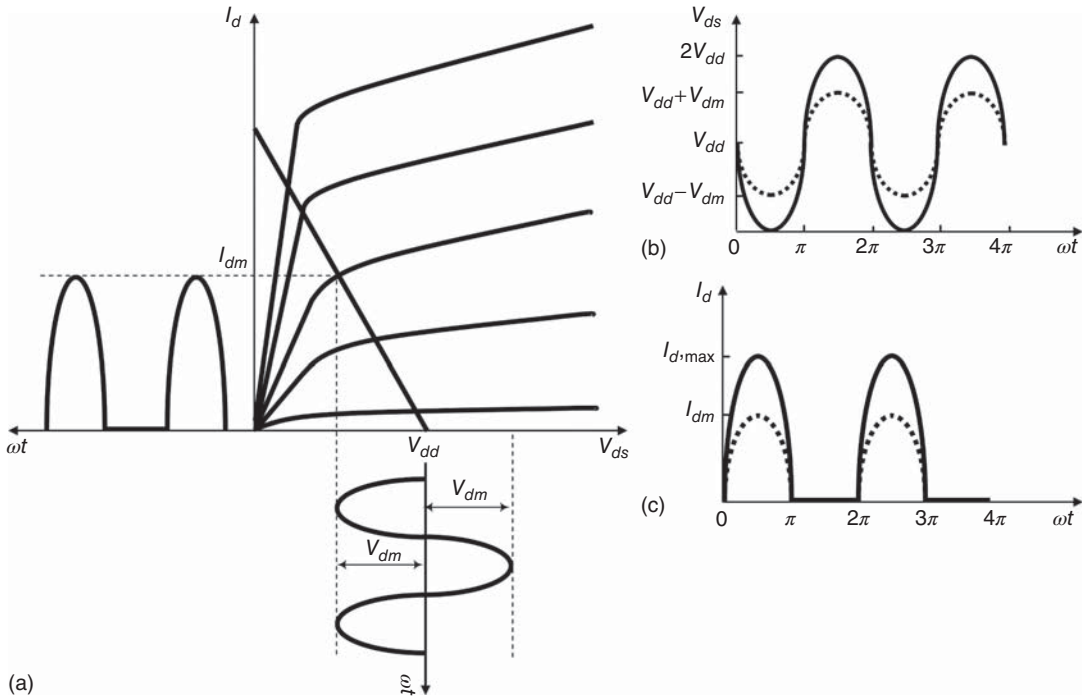


Figure 11.22. Operating characteristics of the Class B amplifier: I–V curves superimposed on the AC load line (a); drain voltage (b); and drain current (c). The dashed curves in (b) and (c) represent the general waveforms and the solid curves denote the waveforms corresponding to the maximum voltage ($2V_{dd}$) and current ($I_{d,\max}$) swings.

where n indicates nth cycle of the driving signal. These expressions show that, while the RF driving signal is continuous at the operating frequency ω_o, the transistor acts as a current source only during each half of the RF excitation's cycles, whose current forms a pulse train with a (base) pulse-width of π/ω and a duty cycle of 50%. The transistor is essentially turned on and off alternately during each cycle. According to Fourier series, the current would therefore contain inherently large harmonics, leading to possible distortion in the output signal. These harmonics, however, are assumed to be suppressed by the OMN as mentioned earlier. Figure 11.22 shows a roughly representative AC load line superimposed on the transistor's I–V curves, along with the RF voltage and current waveforms from (11.132) and (11.133), illustrating the operation of the Class-B amplifiers.

Using Fourier series, we can derive the fundamental-frequency component of the drain current as

$$\begin{aligned} I_{\text{rf}} &= \frac{1}{\pi}\int_0^{2\pi} I_d(t)\sin\omega t d(\omega t) = \frac{1}{\pi}\int_0^{\pi} (I_{dm}\sin\omega t)\sin\omega t d(\omega t) \\ &= \frac{1}{\pi}\int_0^{\pi} I_{dm}\sin^2\omega t d(\omega t) = \frac{I_{dm}}{\pi}\int_0^{\pi}\frac{1-\cos(2\omega t)}{2} d(\omega t) \\ &= \frac{I_{dm}}{2} \end{aligned} \tag{11.134}$$

from which,

$$I_{dm} = 2I_{\text{rf}} = \frac{2V_{dm}}{R_L} \tag{11.135}$$

where V_{dm} is also the amplitude of the RF voltage across the load resistor at the fundamental frequency. The RF output power delivered to the load resistor can be obtained as

$$P_{\text{rf}} = \frac{1}{2}V_{dm}I_{\text{rf}} = \frac{V_{dm}^2}{2R_L} \tag{11.136}$$

This power is the same as the RF output power in Class-A amplifiers, which is expected as the load resistance and current are the same for both classes of amplifiers at the fundamental frequency. Due to large harmonics generated in the drain current, the fundamental power of Class-B amplifiers, however, is lower than that of Class A-amplifier. Utilizing Fourier analysis, the DC component or the average of the RF drain current is obtained by integrating the RF drain current over a period as

$$I_{d,\mathrm{dc}} = \frac{1}{2\pi}\int_0^{2\pi} I_d(t)d(\omega t) = \frac{1}{2\pi}\int_0^{\pi} 2I_{\mathrm{rf}}\sin\omega t d(\omega t)$$
$$= \frac{2I_{\mathrm{rf}}}{\pi} = \frac{2V_{dm}}{\pi R_L} \tag{11.137}$$

making use of (11.135). This DC component is the DC input current to the drain. The DC input power to the drain is then obtained as

$$P_{\mathrm{dc}} = V_{dd}I_{d,\mathrm{dc}} = \frac{2V_{dm}V_{dd}}{\pi R_L} \tag{11.138}$$

As the transistor is operated in the cutoff region, it does not consume any DC power until driven by an RF signal, which is in contrast to Class-A amplifiers that consume DC power at all times. Under an RF excitation, the transistor, however, only dissipates DC power during a half of each RF cycle. The DC-RF efficiency is thus expected to be better than that of Class-A amplifiers.

The RF output power delivered to the load reaches a maximum when the peak instantaneous drain voltage $V_{ds,\max} = 2V_{dd}$ and the instantaneous drain current attains its peak amplitude of $I_{d,\max}$. The maximum RF power is thus

$$P_{\mathrm{rf,max}} = \frac{1}{2}V_{dd}I_{d,\max} = \frac{1}{4}V_{ds,\max}I_{d,\max} \tag{11.139}$$

The DC-RF conversion efficiency can be obtained from (11.136) and (11.138) as

$$\eta_{\mathrm{dc}} = \frac{P_{\mathrm{rf}}}{P_{\mathrm{dc}}} = \frac{\pi}{4}\frac{V_{dm}}{V_{dd}} \le \frac{\pi}{4} = 0.785 \tag{11.140}$$

recognizing that the maximum possible value for the magnitude of the RF voltage (V_{dm}) is V_{dd}. The maximum DC-RF conversion efficiency is thus 78.5%[5] which is much more than that of Class A-amplifiers, making Class-B amplifiers a better choice in term of efficiency. In practical circuits, the efficiency, however, is much less than 78.5% due to various factors and unavoidable effects caused by nonideal conditions. Class-B amplifiers consume at least 21.5% of the supplied DC power.

Neglecting the RF input power and considering only the RF output power P_{rf} with respect to the DC input power P_{dc}, the power dissipated in the transistor is obtained as the difference between P_{dc} and P_{rf} from (11.136) and (11.138) as

$$P_d = \frac{V_{dm}}{R_L}\left(\frac{2V_{dd}}{\pi} - \frac{V_{dm}}{2}\right) \tag{11.141}$$

Taking the derivative of P_d with respect to V_{dm}, letting it equal to zero, and solving for V_{dm}, we get

$$V_{dm} = \frac{2V_{dd}}{\pi} \tag{11.142}$$

which specifies the value of V_{dm} at which the dissipated power is maximum. Substituting (11.142) into (11.141) gives the maximum power dissipation of

$$P_{d,\max} = \frac{2}{\pi^2 R_L}V_{dd}^2 \tag{11.143}$$

[5]The maximum efficiency is different for different driving waveforms. For instance, for square waves, the efficiency can approach 100%.

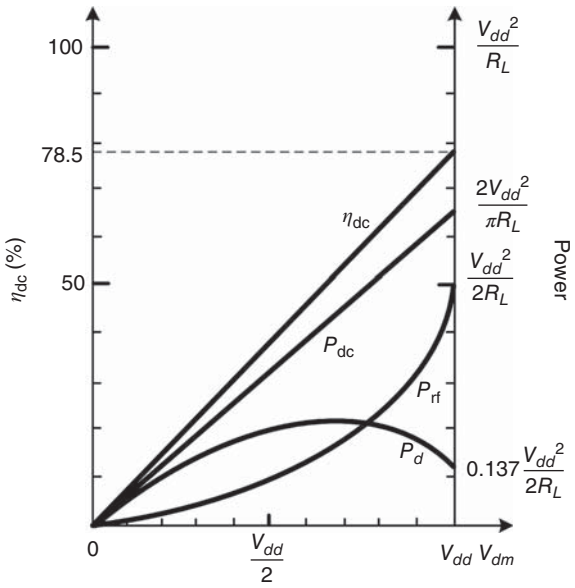


Figure 11.23. Characteristics of DC-RF conversion efficiency, power dissipation, RF output power, and DC input power in Class-B PAs.

Figure 11.23 illustrates the relation between the DC-RF conversion efficiency η_{dc}, P_d, P_{rf} and P_{dc}, and their behavior as a function of the RF output voltage swing V_{dm}. The input DC power and efficiency increase linearly with respect to the drain voltage swing, reaching a maximum power of $2V_{dd}^2/\pi R_L$ and a maximum efficiency of 78.5%. The RF output power and the power dissipation follow parabolic and hyperbolic curves, reaching maximum values of $V_{dd}^2/2R_L$ and $2V_{dd}^2/\pi^2 R_L$, respectively. For a given bias voltage, the efficiency depends solely on the amplitude of the drain voltage swing. In contrast to Class-A amplifiers, Class-B amplifiers dissipate no power when they are not driven with RF signals. As the drain voltage swing is increased, the dissipated power increases to a maximum value of $2V_{dd}^2/\pi^2 R_L$ at $V_{dm} = 2V_{dd}/\pi$ and then reduces along a hyperbolic curve. As compared to Class-A amplifiers, Class-B amplifiers are more suitable for higher power applications, where both higher efficiency and lower power consumption facilitate the handling of high power. They are especially useful for pulse-signal amplification, particularly those having low duty cycles, due to small average power consumption and current over the entire operating time. In typical usages, the bias condition for Class-B amplifiers is also adjusted from the setting for ideal Class-B amplifiers, making them actually operated as between Class-A and Class-B types, for enhanced performance.

As mentioned earlier, the resultant half-cycle operation of the RF drain current causes significant harmonic content in the output signal, which inevitably degrades the fundamental-frequency output power and causes possible signal distortion. In order to alleviate these problems, Class-B PAs are typically employed in a push–pull configuration. Push–pull amplifier is described in Section 11.5.3 and consists of two transistors driven oppositely in phase, in which one transistor conducts current while the other is off during one half of each of the driving sinusoidal cycles, resulting in complete elimination of even harmonics (under ideal conditions) and hence significant reduction in signal distortion as compared to a single-transistor Class-B counterpart.

11.4.2.3 Class-C Amplifier. It is recognized during the analysis of Class-B amplifiers that the efficiency of PAs improves as the conduction time of the transistor's RF drain current is reduced. It is thus natural to reduce the drain current's conduction duration further from the half-cycle period of Class-B amplifiers to achieve better efficiency. One apparent approach would be to bias the transistor so that it is operated beyond cut-off with no DC current flowing across drain and source. This results in the so-called Class-C amplifiers. In Class-C amplifiers, the bias condition is set so that the transistor only conducts current within duration less than a half-cycle for each of the driving cycles, or with a conduction angle of less than 180°. While

Class-A and B amplifiers are linear amplifiers, Class-C amplifiers function as nonlinear devices, in which the output signal amplitude does not change linearly with respect to the input signal. The output waveform, however, resembles a sinusoid corresponding to a sinusoidal input signal, assuming the OMN filters out all the generated harmonics.

Again, we assume a sinusoidal driving signal and can write the drain voltage and current, under ideal conditions, as

$$V_{ds}(t) = V_{ds,\mathrm{dc}} + V_{ds,\mathrm{rf}}(t) = V_{dd} + V_{dm}\cos\omega t \tag{11.144}$$

$$I_d(t) = I_{d,\mathrm{rf}}(t) - I_{dd} = \begin{cases} I_{dm}\left[\sin\omega t - \sin\left(\dfrac{\pi-\theta}{2}\right)\right], & \dfrac{(4n-3)\pi-\theta}{2} \le \omega t \le \dfrac{(4n-3)\pi+\theta}{2} \quad \text{for } n = 1, 2, \ldots, N \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} I_{dm}\left[\sin\omega t - \cos\left(\dfrac{\theta}{2}\right)\right], & \dfrac{(4n-3)\pi-\theta}{2} \le \omega t \le \dfrac{(4n-3)\pi+\theta}{2} \quad \text{for } n = 1, 2, \ldots, N \\ 0, & \text{otherwise} \end{cases} \tag{11.145}$$

where θ represents the conduction angle ($\theta < 180°$), as indicated in Figure 11.24, and n denotes the nth cycle. Equation (11.145) models the drain current, considering that it only exists within the angle $\theta < \pi$ and thus the amplitude must be reduced by an amount corresponding to the angle $(\pi - \theta)/2$ from the edge of the cycle. This amplitude reduction is represented by $I_{dd} = I_{dm}\cos(\theta/2)$. As can be seen, while the RF driving signal is continuous, the transistor acts as a current source only during less than half of each RF cycle. The current thus forms a pulse train with a pulse-width of θ/ω and a duty cycle less than 50%. As the conduction angle is reduced, this duty cycle can become substantially small, making the average drain current extremely small, thus rendering the amplifier ineffective. Switching the transistor on and off alternately inevitably produces substantial harmonics in the drain current, and hence distortion. Again, we assume the harmonics are filtered out by the OMN, so that the voltage and current reaching the load resemble sinusoids. These waveforms are indeed purely sinusoidal when all the harmonics are completely removed by the OMN, leaving only the fundamental components arriving at the load. Figure 11.24 shows an approximate AC load line drawn on the transistor's I–V curves, along with the RF voltage and current waveforms from (11.144) and (11.145), describing the characteristics of Class-C amplifiers.

The fundamental-frequency component of the drain current can be derived from

$$\begin{aligned} I_{\mathrm{rf}} &= \frac{1}{\pi}\int_0^{2\pi} I_d(t)\sin\omega t d(\omega t) = \frac{1}{\pi}\int_{(\pi-\theta)/2}^{(\pi+\theta)/2} I_{dm}\left(\sin\omega t - \cos\frac{\theta}{2}\right)\sin\omega t d(\omega t) \\ &= \frac{1}{\pi}\int_{(\pi-\theta)/2}^{(\pi+\theta)/2} I_{dm}\left(\sin^2\omega t - \cos\frac{\theta}{2}\sin\omega t\right) d(\omega t) \\ &= \frac{I_{dm}}{\pi}\int_{(\pi-\theta)/2}^{(\pi+\theta)/2}\left[\frac{1}{2} - \frac{1}{2}\cos(2\omega t) - \cos\frac{\theta}{2}\sin\omega t\right] d(\omega t) \\ &= \frac{I_{dm}}{\pi}\left\{\frac{\theta}{2} - \frac{1}{4}\left[\sin(\pi+\theta) - \sin(\pi-\theta)\right] + \cos\frac{\theta}{2}\left[\cos\left(\frac{\pi+\theta}{2}\right) - \cos\left(\frac{\pi-\theta}{2}\right)\right]\right\} \\ &= \frac{I_{dm}}{2\pi}(\theta - \sin\theta) \end{aligned} \tag{11.146}$$

upon applying $\sin(a \pm b) = \sin a\cos b \pm \sin b\cos a$. From (11.146), we obtain

$$I_{dm} = \frac{2\pi I_{\mathrm{rf}}}{\theta - \sin\theta} \tag{11.147}$$

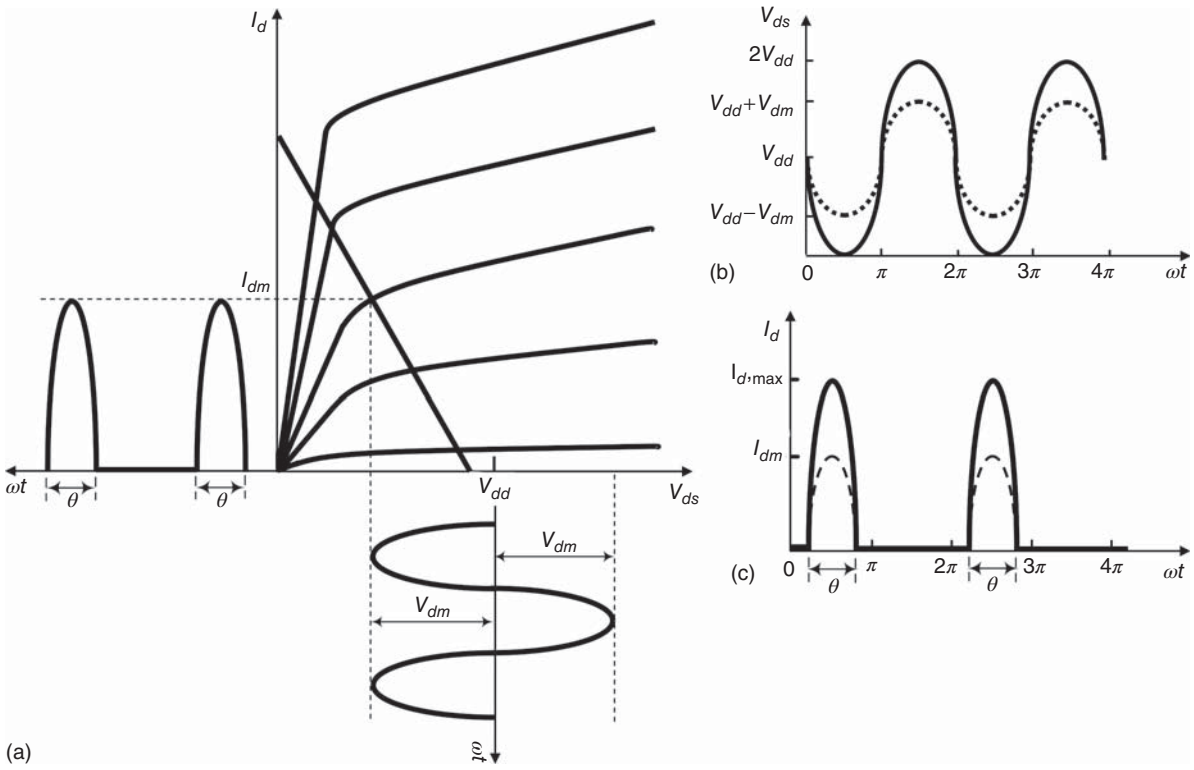


Figure 11.24. Operating characteristics of Class-C amplifiers: I–V curves superimposed on the AC load line (a); drain voltage (b); and drain current (c). The dashed curves in (b) and (c) represent the general waveforms and the solid curves denote the waveforms corresponding to the maximum voltage ($2V_{dd}$) and current ($I_{d,\max}$) swings. θ is the conduction angle, within which the drain current exists.

The RF output power delivered to the load resistor is obtained as

$$P_{\rm rf} = \frac{1}{2} V_{dm} I_{\rm rf} = \frac{(\theta - \sin\theta) V_{dm} I_{dm}}{4\pi} \tag{11.148}$$

using (11.147), where V_{dm} is also the amplitude of the RF voltage across the load resistor at the fundamental frequency under ideal conditions. This power depends on the conduction angle and is equal to the RF output power of Class-A and B amplifiers when the conduction angle is 180°. Accordingly, the load resistance and current are the same as those for Class A and B at the fundamental frequency for 180° conduction period. The DC component or the average of the drain current supplied to the drain is derived as

$$\begin{aligned} I_{d,\rm dc} &= \frac{1}{2\pi}\int_0^{2\pi} I_d(t)\, d(\omega t) = \frac{1}{2\pi}\int_{(\pi-\theta)/2}^{(\pi+\theta)/2} I_{dm}\left[\sin\omega t - \cos\frac{\theta}{2}\right] d(\omega t) \\ &= -I_{dm}\left[\cos\omega t + (\omega t)\cos\frac{\theta}{2}\right]\Bigg|_{(\pi-\theta)/2}^{(\pi+\theta)/2} \\ &= \frac{I_{dm}}{\pi}\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right) = \frac{2\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right)}{\theta - \sin\theta} I_{\rm rf} \end{aligned} \tag{11.149}$$

making use of (11.147). The DC input power to the drain is then determined as

$$P_{\rm dc} = V_{dd} I_{d,\rm dc} = \frac{1}{\pi}\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right) V_{dd} I_{dm} \tag{11.150}$$

As the transistor is operated beyond its cutoff, it does not consume any DC power until an RF signal arrives. Under an RF excitation, the transistor, however, only dissipates DC power within the time (θ/ω) that the transistor conducts in each sinusoidal cycle. The DC-RF efficiency is thus expected to be better than those for Class-A and B amplifiers.

The DC-RF conversion efficiency can be obtained from (11.148) and (11.150) as

$$\eta_{\rm dc} = \frac{P_{rf}}{P_{\rm dc}} = \frac{\theta - \sin\theta}{4\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right)} \cdot \frac{V_{dm}}{V_{dd}} \leq \frac{\theta - \sin\theta}{4\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right)} \tag{11.151}$$

The maximum efficiency results when the peak magnitude of the RF voltage (V_{dm}) reaches V_{dd}. The maximum DC-RF conversion efficiency depends on the conduction angle as expected. When the conduction angle is 180°, the efficiency reduces to that of Class-B amplifiers. Equation (11.151) indicates that the conversion efficiency increases as the conduction angle is reduced. Reducing the conduction angle, however, has adverse effect of decreasing the output power, which, for very small conduction time, diminishes substantially to be useful. Taking the limit of the maximum efficiency, $\eta_{\rm dc,max}$, as θ approaches 0 and applying L'Hospital rule, we get

$$\lim \eta_{\rm dc,max} = \lim \frac{1 - \cos\theta}{\theta \sin\frac{\theta}{2}} = \lim \frac{2\sin\frac{\theta}{2}}{\theta} = \lim \cos\frac{\theta}{2} = 1 \tag{11.152}$$

which shows that the maximum efficiency, under ideal circuit condition, can reach as high as 100% when the conduction angle is reduced to zero. Under that operating condition, however, the RF output power vanishes, as can be seen from (11.148). Therefore, Class-C amplifiers should never be operated near zero-conduction-angle. The existence of the conduction angle in amplifiers, however, provides a trade-off between efficiency and output power, thus enabling RFIC designers to optimize the performance of amplifiers with respect to a particular specification.

As in previous cases, the power dissipated in the transistor is obtained as the difference between the DC input power $P_{\rm dc}$ and the RF output power $P_{\rm rf}$. From (11.148) and (11.150), we get

$$P_d = \frac{V_{dm}}{R_L}\left[\frac{2\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right)}{(\theta - \sin\theta)}V_{dd} - \frac{V_{dm}}{2}\right] \tag{11.153}$$

making use of (11.147). Taking the derivative of P_d with respect to V_{dm}, letting it equal to zero, and solving for V_{dm}, we get

$$V_{dm} = \frac{2\left(\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}\right)}{(\theta - \sin\theta)}V_{dd} \tag{11.154}$$

which specifies the value of V_{dm} at which the dissipated power is maximum. Substituting (11.154) into (11.153) gives the maximum power dissipation of

$$P_{d,\rm max} = \frac{2}{R_L}\left(\frac{\sin\frac{\theta}{2} - \frac{\theta}{2}\cos\frac{\theta}{2}}{\theta - \sin\theta}\right)^2 V_{dd}^2 \tag{11.155}$$

It is noted that all equations derived for Class-C amplifiers reduced to those for Class-A and Class-B amplifiers when θ equals to 2π and π, respectively.

Figure 11.25 illustrates the relation between the DC-RF conversion efficiency $\eta_{\rm dc}$, P_d, $P_{\rm rf}$ and $P_{\rm dc}$, and their behavior as a function of the RF output voltage swing V_{dm} for different conduction angles. As for Class-B amplifiers, Class-C amplifiers do not consume power when they are not driven with RF signals and

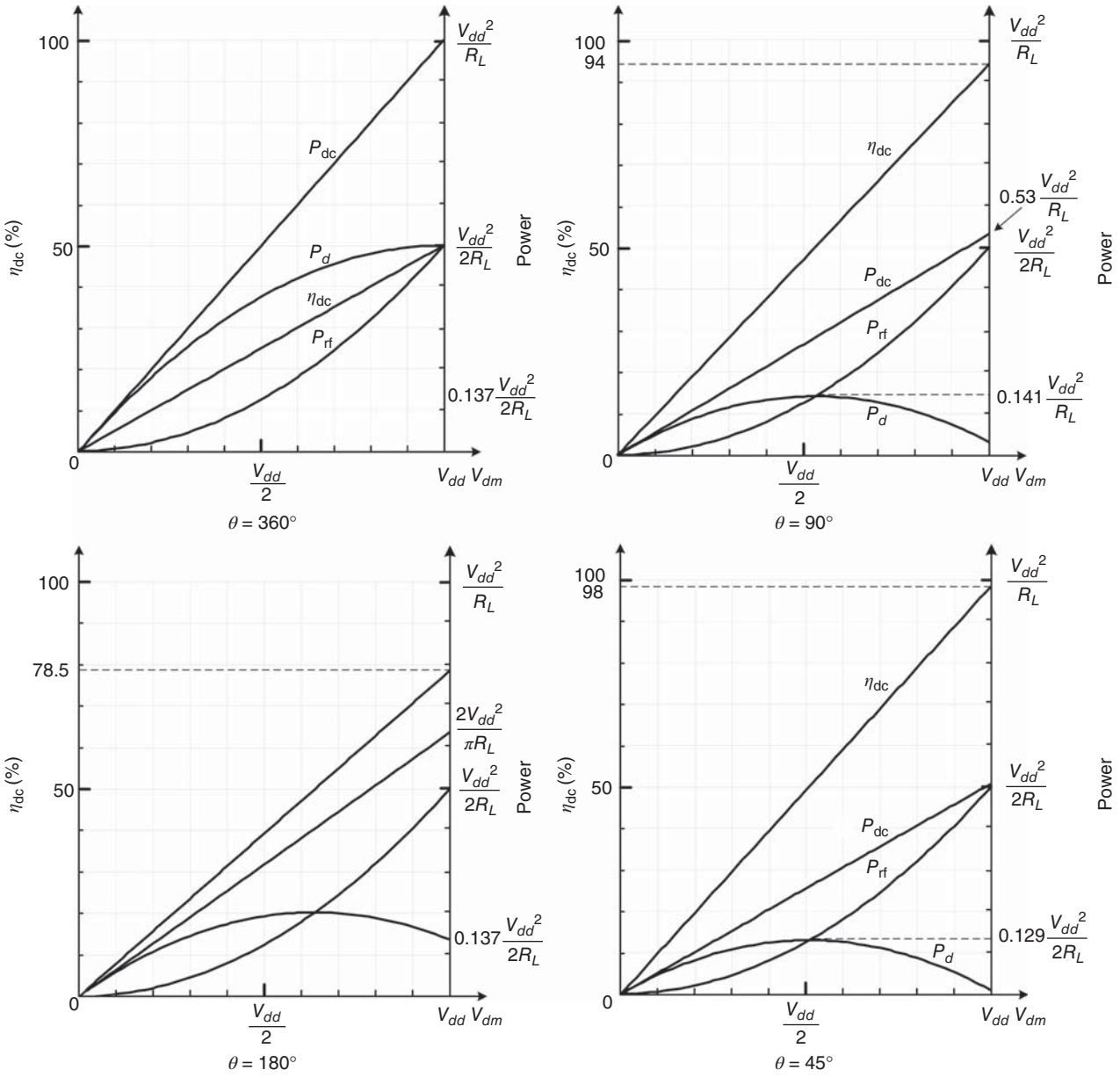


Figure 11.25. Characteristics of DC-RF conversion efficiency, power dissipation, RF output power, and DC input power in Class B PAs with the conduction angle θ as parameter. $\theta = 360°$ and 180°. correspond to Class-A and Class-B amplifiers, respectively.

are suitable for high power applications, where both high efficiency and low power consumption facilitate the handling of high power. They are especially useful for pulse-signal amplification, particularly those having low duty cycles, due to small average power consumption and current over the entire operating time. Figure 11.26 shows the maximum DC-RF conversion efficiency as a function of the conduction angle.

11.4.2.4 Class-AB Amplifier. It is recognized during the analysis of Class-A and B amplifiers that the efficiency of PAs improves as the conduction time of the transistor's RF drain current is reduced. It is thus natural to reduce the drain current's conduction duration further from the half-cycle period of Class-B amplifiers to achieve better efficiency. One apparent approach would be to bias the transistor so that it is operated between Class A and B. This results in the so-called Class-AB amplifiers. In Class-AB amplifiers, the bias condition is set so that the transistor only conducts current within duration less than a half of each of the driving cycles.

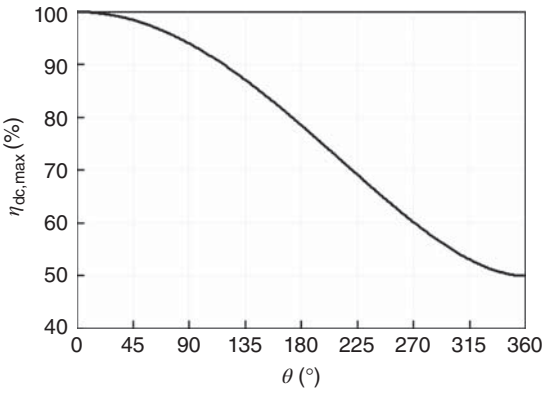


Figure 11.26. Maximum DC-RF conversion efficiency versus conduction angle.

11.5 BALANCED AMPLIFIERS

Single-ended amplifiers are unbalanced structure and widely used for both low noise and PAs. Balanced amplifiers are another amplifier type with unique characteristics not found in single-ended counterparts. They are also attractive for both low noise and PAs and in fact preferred for certain RF applications and systems. Typical balanced amplifiers are considered a combination of two identical single-ended amplifiers and their design typically starts with the design of the constituent amplifier, whose single-ended characteristics are retained in the properties of the balanced amplifiers. The most widely used balanced amplifiers are perhaps differential amplifier, balanced amplifier employing 90° hybrids or couplers, and push–pull amplifier employing 180° hybrids, couplers, baluns, or transformers. The balanced amplifier employing 90° hybrids is commonly known as "balanced amplifier" among microwave engineers, and the name "90° balanced amplifier" will be used hereafter to indicate this kind of amplifier. Differential amplifiers are conceivably more stable than single-ended amplifiers and facilitate integration with other balanced RF structures such as balanced mixers and balanced-type antennas (e.g., antennas with balanced feeds like coplanar strips). Ninety-degree balanced amplifiers have inherent wideband characteristics dictated not by the individual single-ended amplifiers but by the employed 90° hybrids, and are particularly attractive for broadband applications. Push–pull amplifiers have bandwidth depending on both 180° hybrids and constituent single-ended amplifiers and can also be designed for wideband applications using broadband hybrids.

11.5.1 Differential Amplifiers

11.5.1.1 Basics. A single-ended amplifier has one input and one output port. Single-ended amplifiers, while being popular, have several drawbacks resulting from their unsymmetrical topology such as parasitic inductance and resistance to ground and inability to reject unwanted signals such as noise. Differential amplifiers are balanced structures originally invented for amplification using vacuum tubes and have become one of the most preferred amplifier topologies due to various advantages such as (common-mode) rejection of unwanted signals. Differential amplifiers have two inputs and two outputs. Generally, a differential amplifier can be considered consisting of two single-ended amplifiers, as in other balanced amplifiers. However, a differential amplifier is different from other balanced amplifiers due to both internal connection between the two constituent single-ended amplifiers and external connection to other components.

Figure 11.27(a) shows the core of differential amplifiers, which consists of two source-coupled transistors with two input signals, V_{i1} and V_{i2}, and two output signals, V_{o1} and V_{o2}. The two input signals are provided from a preceding balanced component or a single-ended component using a balun (or 180° out-of-phase power divider), as commonly known among microwave engineers, or a transformer commonly referred to in low frequency analog circuits. Similarly, the two output signals can be fed directly to a balanced component or combined using a balun for use with a single-ended component.

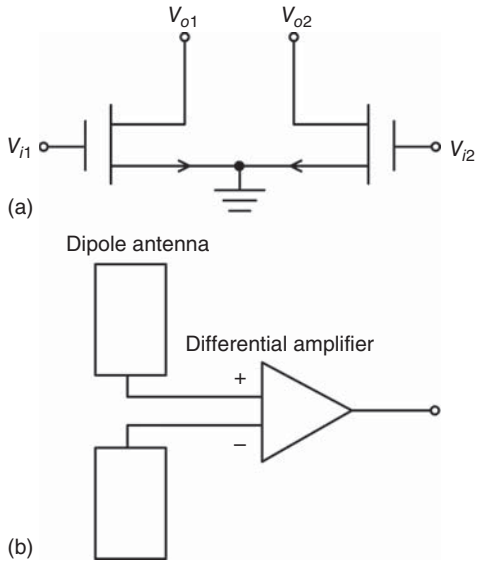


Figure 11.27. (a) Differential amplifier core with two N-MOSFETs and (b) illustration of direct integration between a differential amplifier and a (balanced) dipole antenna; a single output, combined from two differential outputs, is assumed for the amplifier.

Intuitively, from the core topology, differential amplifiers should have various advantages as compared to their single-ended counterparts. They possess inherent rejection or immunity from "signal disturbances" common to both inputs, known as "common-mode signals," such as common noise from power supplies or unwanted couplings from nearby circuits. This feature is very desirable in practical RFICs, especially those integrated directly with digital circuits on the same chip. It should be noted, however, that, even for ideal or perfectly balanced differential amplifiers, perfect common-mode rejection only occurs when the perturbed signals appear equally on both input signals. In order for this to happen, the physical structures delivering the perturbed signals must be perfectly symmetrical with respect to the two input ports. Differential amplifiers have larger output voltage swings and higher power handling. They have improved linearity and hence larger dynamic range. There exists a "virtual ground" for the input differential signals at the connection between the sources of the transistors in differential amplifiers, thus avoiding parasitic inductance and resistance between the source connection and the actual ground as existed in single-ended amplifiers. Differential amplifiers possess no "DC offset" at their output, allowing direct connection of stages in multistage differential amplifiers or of multiple differential amplifiers without using coupling capacitors in between. Differential amplifiers also facilitate direct integration with balanced components such as balanced antennas or doubly balanced mixers without the need of baluns, as shown in Figure 11.27(b), which is important in some system integrations.

Differential amplifiers, however, also have several shortcomings compared to their single-ended counterparts. They occupy larger area, theoretically twice as much. It should be noted, however, that, in practice, the inherent immunity from common-mode noises or other unwanted signals facilitates optimization of differential amplifiers' layout, which may result in size smaller than that of single-ended amplifiers in certain circuit environments, particularly those involving highly dense analog and digital integrated circuits. They consume twice as large of power consumption or bias current for same noise figure and/or gain. On the other hand, in order to maintain same power consumption, the gain of differential amplifiers is 3 dB lower while the noise figure is higher.

Practical differential amplifiers employ a transistor current source or a linear resistor at the source connection to minimize the dependence of bias currents upon the common-mode levels of the two input signals, which affects the amplifiers' gain and output voltage swings. Figure 11.28 shows a simplified differential amplifier with a constant current source I_{CS}. The bias voltage V_{dd} and load resistors at the drains are also included. It should be noted that other types of load resistors, such as diodes or transistors acting as current sources, may also be used. To increase the differential amplifier's gain and isolation, multiple transistors connected

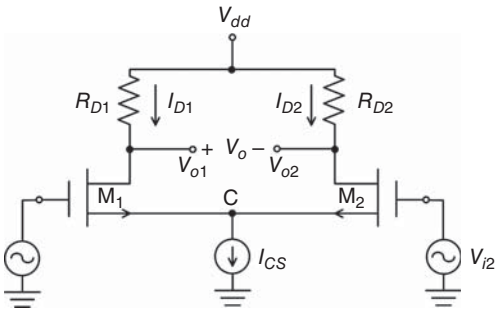


Figure 11.28. A simplified differential amplifier employing a constant current source.

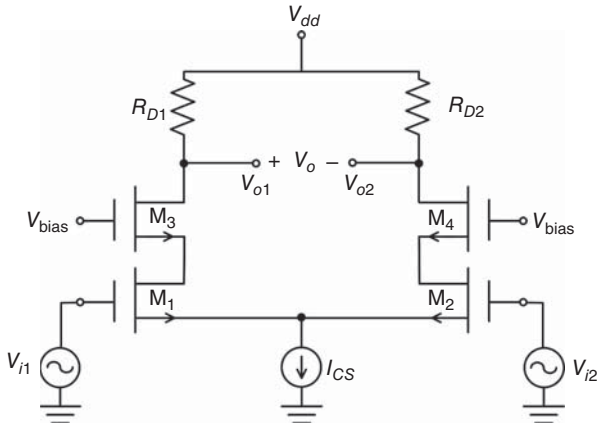


Figure 11.29. A simplified cascoded differential amplifier.

in cascode, similar to those used in single-ended amplifiers, may also be used in each half of the amplifier, as shown in Figure 11.29, at the cost of higher power consumption.

11.5.1.2 Analysis

DC Offset

Offset, or typically known as DC offset, is a phenomenon in RFICs caused by nonzero output voltage when the input voltage is set to zero. The DC-offset level is equal to the nonzero output voltage. In amplifiers, the offset is also amplified along with input signal to the amplifier, causing possible problems in amplifier's operation and/or subsequent components and hence system performance. For instance, the offset level in multistage amplifiers or cascaded amplifiers may be significantly amplified that degrades the linearity of the amplifiers themselves and/or subsequent components. Zero- or near-zero-offset is a desired feature in amplifiers and other RFICs.

Referring to Figure 11.28, we can write

$$V_{o1} = V_{dd} - R_{D1}I_{D1} \tag{11.156}$$

and

$$V_{o2} = V_{dd} - R_{D2}I_{D2} \tag{11.157}$$

and from which, obtain the differential output voltage as

$$V_o = R_{D2}I_{D2} - R_{D1}I_{D1} \tag{11.158}$$

For identical transistors and drain resistors with equal input signals or zero differential input voltage (i.e., $V_{i1} = V_{i2}$), $I_{D1} = I_{D2} = I_{CS}/2$, and hence the differential output voltage is equal to zero, thus avoiding offset. This zero-offset allows direct cascades of differential pairs or multiple differential amplifiers without coupling capacitors – a unique feature in differential amplifiers.

Perfectly Balanced Amplifiers

In ideal situations, the two input signals to differential amplifiers are purely differential with equal amplitude and 180° out of phase. Practical differential amplifiers, however, do not have perfectly differential input signals. Nondifferential excitations of differential amplifiers, either in unequal amplitudes or phases not perfectly 180° difference, degrade circuit performance such as reduction of the common-mode rejection ability. To illustrate this imperfect yet practical operation, without loss of generality, we consider a perfectly balanced differential amplifier using the amplifier topology shown in Figure 11.28 with identical transistors ($M_1 \equiv M_2 = M$) and drain resistors ($R_{D1} = R_{D2} = R_D$). We also assume that the two transistors are operated exactly under the same condition, so that all their electrical parameters are identical, and the amplifier's layout is perfectly symmetrical with respect to the central vertical line going through the source connection point C. Under these conditions, the considered differential amplifier has perfect symmetry – both electrically and physically.

We begin the analysis by expressing the two different input voltages as

$$V_{i1} = \frac{1}{2}(V_{i1} + V_{i2}) + \frac{1}{2}(V_{i1} - V_{i2}) = V_{ic} + \frac{1}{2}V_{id} \tag{11.159}$$

and

$$V_{i2} = \frac{1}{2}(V_{i1} + V_{i2}) + \frac{1}{2}(V_{i2} - V_{i1}) = V_{ic} - \frac{1}{2}V_{id} \tag{11.160}$$

where

$$V_{ic} = \frac{1}{2}(V_{i1} + V_{i2}) \tag{11.161}$$

and

$$V_{id} = V_{i1} - V_{i2} \tag{11.162}$$

As can be seen, V_{ic} appears in both of the inputs, thus representing the voltage portion common to these inputs and is referred to as the common-mode input voltage. On the other hand, V_{id} appears oppositely in the input voltages and is thus classified as the differential-mode input voltage.

The output voltages corresponding to V_{i1} and V_{i2} can be written, respectively, as

$$V_{o1} = V_{oc} + \frac{1}{2}V_{od} \tag{11.163}$$

and

$$V_{o2} = V_{oc} - \frac{1}{2}V_{od} \tag{11.164}$$

where

$$V_{oc} = \frac{1}{2}(V_{o1} + V_{o2}) \tag{11.165}$$

and

$$V_{od} = V_{o1} - V_{o2} \tag{11.166}$$

are the common-mode output voltage (corresponding to V_{ic}) and the differential-mode output voltage (corresponding to V_{id}), respectively. Equations (11.163) and (11.164) also show that each output signal contains

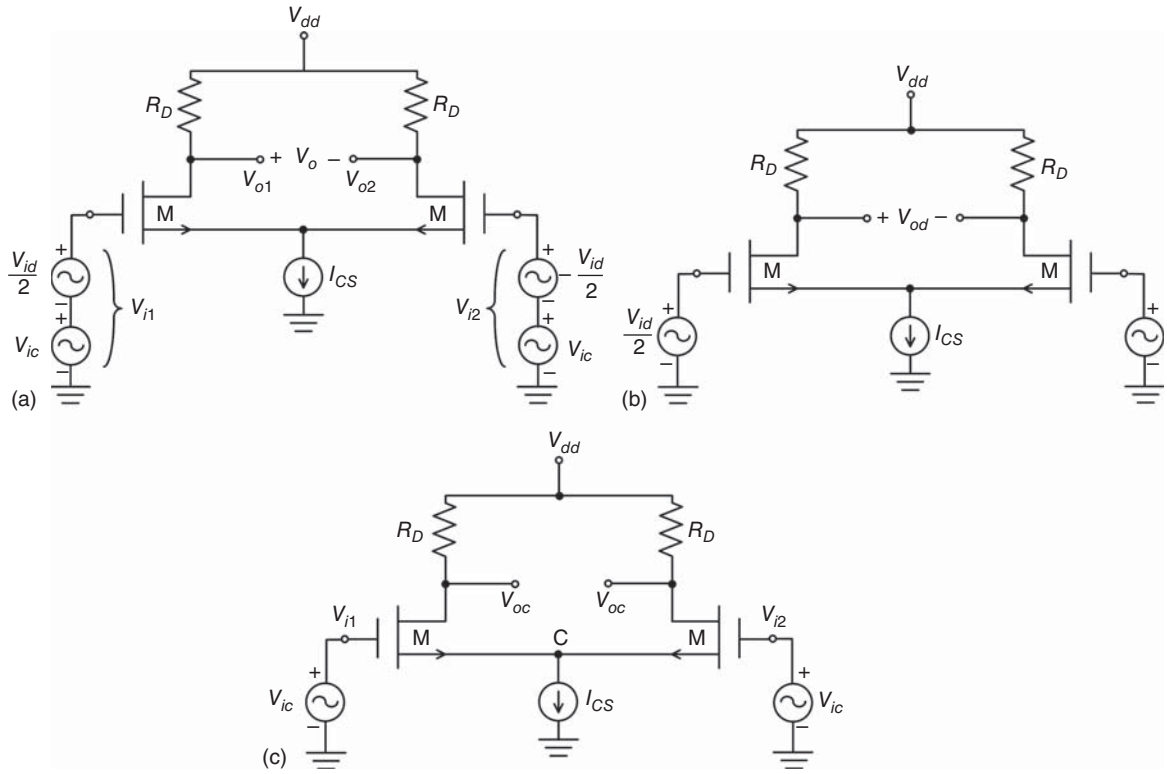


Figure 11.30. Differential amplifier with input signals decomposed into differential and common parts (a) and its breakdown into two separate amplifiers with differential input only (b) and common input only (c).

two parts: one common part to both output signals (V_{oc}) and one differential part ($\pm V_d/2$) having equal magnitude and 180° phase difference.

Graphically, we can then redraw the differential amplifier as shown in Figure 11.30(a) and further decompose it into two circuits: Figure 11.30(b) with two equal but 180° out-of-phase excitations (differential mode) and Figure 11.30(c) with two identical input voltages (common mode). We assume that the two transistors are operated identically so all their electrical parameters are exactly the same and the circuit's layout is symmetrical with respect to the central line. The amplifier circuit itself thus has perfect symmetry, both electrically and physically, without considering the driving input signals. Figure 11.30(b) and (c) with opposite and identical excitations lend themselves perfectly to the odd- and even-mode analysis, respectively, discussed in Section 8.2.2 for symmetrical passive microwave components. We can then analyze them utilizing the odd- and even-mode analysis.

Differential Mode. Identical to the odd-mode operation discussed in Section Odd-Mode Analysis of Chapter 8, under the differential mode, there exists an electrical wall or short circuit along the central plane of the amplifier circuit. The amplifier circuit in Figure 11.30(b) can hence be decomposed into two separate identical subcircuits as shown in Figure 11.31 – each operating independently. Note that the current source I_{CS} is removed due to the short circuit at the connection between the source terminals, which acts as a "virtual ground." The analysis of the amplifier under the differential mode can thus only need to be performed on each half of the amplifier.

Using scattering parameters, we can write for the circuits shown in Figure 11.31(a) and (b):

$$\begin{bmatrix} b_1^d \\ b_2^d \end{bmatrix} = \begin{bmatrix} S_{11}^d & S_{12}^d \\ S_{21}^d & S_{22}^d \end{bmatrix} \begin{bmatrix} a_1^d \\ a_2^d \end{bmatrix} \tag{11.167}$$

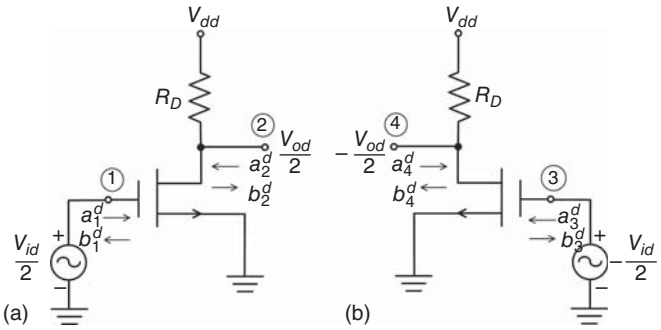


Figure 11.31. (a, b) Two equivalent half-circuits of Figure 11.30(b). Note that $V_{od2} = -V_{od4} = V_{od}/2$.

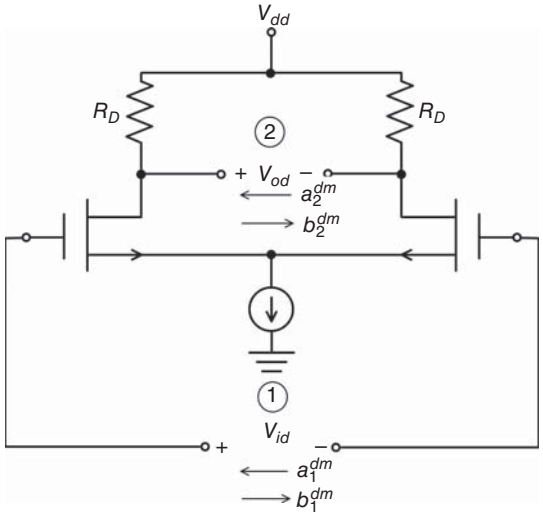


Figure 11.32. Differential amplifier under differential-mode excitation.

and

$$\begin{bmatrix} b_3^d \\ b_4^d \end{bmatrix} = \begin{bmatrix} S_{11}^d & S_{12}^d \\ S_{21}^d & S_{22}^d \end{bmatrix} \begin{bmatrix} a_3^d \\ a_4^d \end{bmatrix} \tag{11.168}$$

respectively, where b's and a's represent the reflected and incident voltages, respectively, and S_{ij}^d $(i, j = 1, 2)$ are the S-parameters of one half-circuit of the differential amplifier under the differential-mode operation. We now superimpose Figure 11.31(a) and (b) and add the current source I_{CS} to reconstruct the differential amplifier under the differential-mode excitation as shown in Figure 11.32, which is indeed equivalent to Figure 11.30(b). Without taking into account the effect from I_{CS}, we can write for Figure 11.32, considering the two ports 1 and 2, as

$$\begin{bmatrix} b_1^{dm} \\ b_2^{dm} \end{bmatrix} = \begin{bmatrix} S_{11}^{dm} & S_{12}^{dm} \\ S_{21}^{dm} & S_{22}^{dm} \end{bmatrix} \begin{bmatrix} a_1^{dm} \\ a_2^{dm} \end{bmatrix} \tag{11.169}$$

where S_{ij}^{dm} $(i, j = 1, 2)$ are the whole differential amplifier's S-parameters under the differential mode. The reflected input and output voltages can be expressed using (11.167) and (11.168) as

$$\begin{aligned} b_1^{dm} &= b_1^d - b_3^d = S_{11}^d(a_1^d - a_3^d) + S_{12}^d(a_2^d - a_4^d) \\ &= S_{11}^d a_1^{dm} + S_{12}^d a_2^{dm} \end{aligned} \tag{11.170}$$

and

$$b_2^{dm} = b_2^d - b_4^d = S_{21}^d(a_1^d - a_3^d) + S_{22}^d(a_2^d - a_4^d)$$
$$= S_{21}^d a_1^{dm} + S_{22}^d a_2^{dm} \tag{11.171}$$

respectively, making use of

$$a_1^{dm} = a_1^d - a_3^d \tag{11.172}$$

and

$$a_2^{dm} = a_2^d - a_4^d \tag{11.173}$$

Comparing (11.170) and (11.171) to (11.169) gives

$$S_{11}^{dm} = S_{11}^d S_{12}^{dm} = S_{12}^d$$
$$S_{21}^{dm} = S_{21}^d S_{22}^{dm} = S_{22}^d \tag{11.174}$$

which show that the gain, isolation, and return loss of a differential amplifier operated under the differential mode (or a differential amplifier with pure differential inputs of $V_{i1} = -V_{i2}$) is equal to those of its half-circuit, which is basically a single-ended amplifier, provided that same transistor, bias current, and other circuit elements are used in the two halves. It is noted, however, that the magnitude of the input signal for the differential amplifier is twice of that needed for the corresponding single-ended amplifier and their power consumptions are equal. It should also be noted that the analysis is valid for differential amplifiers with pure differential excitation which, however, only occurs in ideal situations – not in real conditions under which differential amplifiers are operated.

Common Mode. Under the common mode, as shown in Figure 11.30(c), the differential amplifier is excited with two identical signals and, like operation under the even mode discussed in Section Even-Mode Analysis of Chapter 8, there is no current flowing across the two output terminals, and a magnetic wall or open circuit appears along the central plane. The amplifier can hence be separated into two independent circuits, as shown in Figure 11.33, and analyzed using either half-circuit. The *S*-parameters for each of the half-circuit, between ports 1(2) and 3(4), can be described as

$$\begin{bmatrix} b_{1(3)}^c \\ b_{2(4)}^c \end{bmatrix} = \begin{bmatrix} S_{11}^c & S_{12}^c \\ S_{21}^c & S_{22}^c \end{bmatrix} \begin{bmatrix} a_{1(3)}^c \\ a_{2(4)}^c \end{bmatrix} \tag{11.175}$$

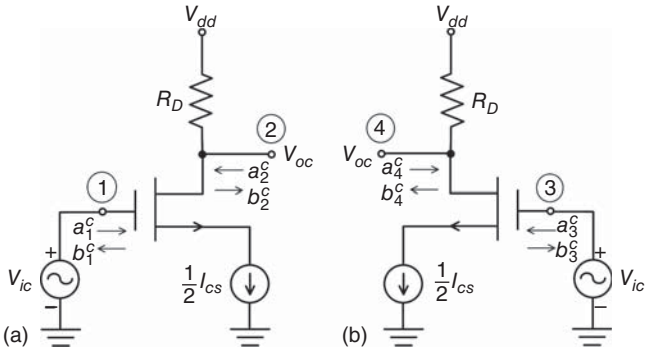


Figure 11.33. (a, b) Differential amplifier decomposed into two identical circuits under the common-mode operation. Note that $V_{oc2} = V_{oc4} = V_{oc}$.

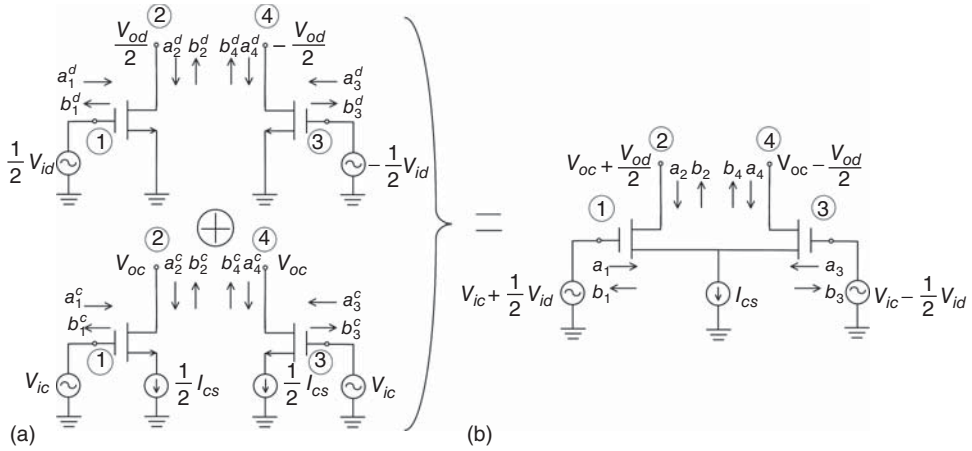


Figure 11.34. (a) Superposition of the differential-amplifier circuits operating under the differential and common modes and (b) the resultant circuit. The drain resistors and V_{dd} are omitted for brevity.

where S_{ij}^c $(i, j = 1, 2)$ are the S-parameters of one half-circuit of the differential amplifier in the common mode. Although there is contribution from each of the common-mode halves, such as the gain S_{21}^c of each half, there is no contribution from the combined two common-mode halves across the two output terminals for the differential amplifier, and thus not affecting the differential-mode output for perfectly balanced differential amplifiers. This common-mode rejection is a unique property of differential amplifiers.

***S*-Parameters.** The entire differential amplifier is now analyzed by superimposing the differential- and common-mode circuits.[6] Figure 11.34 illustrates the superposition of the differential- and common-mode amplifiers and the resultant differential amplifier. The S-parameter equation of the differential amplifier is written from Figure 11.34(b) as

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \tag{11.176}$$

The incident and reflected voltages of the differential amplifier are related to those of the decomposed circuits as

$$a_i = a_i^d + a_i^c \tag{11.177}$$

and

$$b_i = b_i^d + b_i^c \tag{11.178}$$

where $i = 1, 2, 3, 4$, respectively. We let

$$V_{ic} = KV_{id} \tag{11.179}$$

which leads to

$$a_1^c = Ka_1^d \tag{11.180}$$

[6]Superposition is only applicable to linear circuits and therefore, in general, can only be used for small-signal operations that produce no or negligible nonlinearities. It is noted, however, that superposition may also be used for large signals at certain bias voltages under the consideration that at these specific bias conditions, the devices are represented as linear networks.

Making use of (11.177) and (11.180), we obtain

$$a_1^d = \frac{1}{1+K} a_1 \tag{11.181}$$

Substituting (11.181) into (11.180) gives

$$a_1^c = \frac{K}{1+K} a_1 \tag{11.182}$$

The input and output voltages in the differential and common modes are related by the small-signal differential-mode voltage gain G_d and common-mode voltage gain G_c defined as

$$G_d \equiv \frac{V_{od}}{V_{id}} \tag{11.183}$$

and

$$G_c \equiv \frac{V_{oc}}{V_{ic}} \tag{11.184}$$

respectively. In (11.183) and (11.184), it is implicitly implied that $V_{ic} = 0$ and $V_{id} = 0$, respectively, or the differential- and common-mode subcircuits are completely isolated from each other. Utilizing (11.183) and (11.184) along with (11.179), we can write

$$V_{oc} = KG_c V_{id} \tag{11.185}$$

and

$$V_{od} = \frac{G_d}{KG_c} V_{oc} \tag{11.186}$$

from which, we obtain

$$a_2^d = \frac{G_d}{KG_c} a_2^c \tag{11.187}$$

Making use of (11.177) and (11.187), we get

$$a_2^c = \frac{KG_c}{G_d + KG_c} a_2 \tag{11.188}$$

Substituting (11.188) into (11.187) gives

$$a_2^d = \frac{G_d}{G_d + KG_c} a_2 \tag{11.189}$$

The reflected voltage at port 1 can be determined using (11.178), (11.175), and (11.167) as

$$b_1 = b_1^c + b_1^d = (S_{11}^c a_1^c + S_{12}^c a_2^c) + (S_{11}^d a_1^d + S_{12}^d a_2^d) \tag{11.190}$$

Substituting (11.181), (11.182), (11.188), and (11.189) into (11.190) results in

$$b_1 = \frac{1}{1+K}(KS_{11}^c + S_{11}^d) a_1 + \frac{1}{G_d + KG_c}(KG_c S_{12}^c + G_d S_{12}^d) a_2 \tag{11.191}$$

Similarly, we can derive the reflected voltage at port 2 as

$$b_2 = \frac{1}{1+K}(KS_{21}^c + S_{21}^d) a_1 + \frac{1}{G_d + KG_c}(KG_c S_{22}^c + G_d S_{22}^d) a_2 \tag{11.192}$$

We also have from the common- and differential-mode operations:

$$\begin{aligned} a_1^c &= a_3^c a_2^c = a_4^c \\ a_1^d &= -a_3^d a_2^d = -a_4^d \end{aligned} \tag{11.193}$$

We can derive the reflected voltage at port 3, using (11.178), (11.175), (11.168), and (11.193), as

$$b_3 = S_{11}^c a_1^c + S_{12}^c a_2^c - S_{11}^d a_1^d - S_{12}^d a_2^d \tag{11.194}$$

which becomes, after substituting (11.181), (11.182), (11.188), and (11.189),

$$b_3 = \frac{1}{1+K}(KS_{11}^c - S_{11}^d)a_1 + \frac{1}{G_d + KG_c}(KG_c S_{12}^c - G_d S_{12}^d)a_2 \tag{11.195}$$

Similarly, the reflected voltage at port 4 is obtained as

$$b_4 = \frac{1}{1+K}(KS_{21}^c - S_{21}^d)a_1 + \frac{1}{G_d + KG_c}(KG_c S_{22}^c - G_d S_{22}^d)a_2 \tag{11.196}$$

The S-parameters of the differential amplifier can now be obtained from (11.191), (11.192), (11.195), and (11.196) as

$$\begin{aligned} S_{11} = S_{33} &= \tfrac{1}{1+K}(KS_{11}^c + S_{11}^d) \qquad S_{12} = S_{34} = \tfrac{1}{G_d+KG_c}(KG_c S_{12}^c + G_d S_{12}^d) \\ S_{21} = S_{43} &= \tfrac{1}{1+K}(KS_{21}^c + S_{21}^d) \qquad S_{22} = S_{44} = \tfrac{1}{G_d+KG_c}(KG_c S_{22}^c + G_d S_{22}^d) \\ S_{31} = S_{13} &= \tfrac{1}{1+K}(KS_{11}^c - S_{11}^d) \qquad S_{32} = S_{14} = \tfrac{1}{G_d+KG_c}(KG_c S_{12}^c - G_d S_{12}^d) \\ S_{41} = S_{23} &= \tfrac{1}{1+K}(KS_{21}^c - S_{21}^d) \qquad S_{42} = S_{24} = \tfrac{1}{G_d+KG_c}(KG_c S_{22}^c - G_d S_{22}^d) \end{aligned} \tag{11.197}$$

recognizing that

$$\begin{aligned} S_{33} = S_{11} \quad S_{34} = S_{12} \quad S_{23} = S_{41} \\ S_{43} = S_{21} \quad S_{44} = S_{22} \quad S_{14} = S_{32} \end{aligned} \tag{11.198}$$

from the symmetry property of S-parameters and

$$S_{13} = S_{31} \quad S_{24} = S_{42} \tag{11.199}$$

from the reciprocal property. It is noted that the transmissions between ports 1 and 2 and between ports 3 and 4 do not follow the reciprocal property due to active devices (i.e., $S_{21} \neq S_{12}$ and $S_{43} \neq S_{34}$).

Gains and Other Figures of Merit. Under small-signal operations, the input and output voltages of the differential amplifier shown in Figure 12.12 are related by

$$\begin{bmatrix} V_{o1} \\ V_{o2} \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{11} & G_{12} \end{bmatrix} \begin{bmatrix} V_{i1} \\ V_{i2} \end{bmatrix} \tag{11.200}$$

where G_{ij} $(i, j = 1, 2)$ are the small-signal voltage-gain parameters. Substituting (11.159), (11.160), (11.163), and (11.164) into (11.200) and expanding it give

$$V_{oc} + \frac{V_{od}}{2} = G_{11}\left(V_{ic} + \frac{V_{id}}{2}\right) + G_{12}\left(V_{ic} - \frac{V_{id}}{2}\right) \tag{11.201}$$

$$V_{oc} - \frac{V_{od}}{2} = G_{21}\left(V_{ic} + \frac{V_{id}}{2}\right) + G_{22}\left(V_{ic} - \frac{V_{id}}{2}\right) \tag{11.202}$$

Adding and subtracting (11.201) and (11.202) and then solving for V_{oc} and V_{od} yield

$$\begin{bmatrix} V_{oc} \\ V_{od} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(G_{11} + G_{12} + G_{21} + G_{22}) & \frac{1}{4}(G_{11} - G_{12} + G_{21} - G_{22}) \\ G_{11} + G_{12} - G_{21} - G_{22} & \frac{1}{2}(G_{11} - G_{12} - G_{21} + G_{22}) \end{bmatrix} \begin{bmatrix} V_{ic} \\ V_{id} \end{bmatrix} \tag{11.203}$$

It is seen that both the input signals, V_{ic} and V_{id}, contribute to the outputs V_{oc} and V_{od}, as expected. Specifically, V_{id} relates to V_{od} and V_{ic} relates to V_{oc} by the differential- and common-mode voltage gains defined in (11.194) and (11.195), respectively. These voltage gains are obtained from (11.203) as

$$G_d \equiv \left.\frac{V_{od}}{V_{id}}\right|_{V_{ic}=0} = \frac{1}{2}(G_{11} - G_{12} - G_{21} + G_{22}) \tag{11.204}$$

and

$$G_c \equiv \left.\frac{V_{oc}}{V_{ic}}\right|_{V_{id}=0} = \frac{1}{2}(G_{11} + G_{12} + G_{21} + G_{22}) \tag{11.205}$$

The relations between V_{ic} and V_{od} and between V_{id} and V_{oc} are given in the common-to-differential-mode and differential-to-common-mode voltage gains defined as

$$G_{cd} \equiv \left.\frac{V_{od}}{V_{ic}}\right|_{V_{id}=0} = G_{11} + G_{12} - G_{21} - G_{22} \tag{11.206}$$

$$G_{\mathrm{dc}} \equiv \left.\frac{V_{oc}}{V_{id}}\right|_{V_{ic}=0} = \frac{1}{4}(G_{11} - G_{12} + G_{21} - G_{22}) \tag{11.207}$$

respectively, making use of (11.203).

In differential amplifiers, the desired gain is G_d and those undesired are G_c, G_{dc} and G_{cd}. G_{dc} is an important parameter as it not only measures the magnitude of the undesired conversion from the input differential signal to the output common-mode signal, which signifies an unwanted loss of the input differential signal that is otherwise converted to the desired differential output signal, but also the relative magnitude of the output common-mode signal from the input differential signal. This undesired voltage level becomes real input signal to subsequent stages of differential amplifiers or other succeeding components such as mixers, thereby degrading the system performance or, if sufficiently high, can corrupt the entire system. G_{cd} is also an important parameter signifying the conversion from the input common-mode signal to the output differential-mode signal. The converted signal, which is undesired, will enter subsequent stages of the differential amplifier, another differential amplifier, or other active components, as real signal and causes various potential problems such as increasing noise (if the common signal is noise.) Both of G_{dc} and G_{cd} are important design parameters in differential amplifiers.

One of the basic criteria in the design of differential amplifiers is maximizing G_d while minimizing G_c, G_{cd} and G_{dc} as much as possible. For (ideal) perfectly balanced differential amplifiers with ideally pure differential input signals ($V_{i1} = -V_{i2}$), the output signals are also purely differential ($V_{o1} = -V_{o2}$). Under these ideal conditions, we have

$$\begin{aligned} G_{11} &= G_{22} \\ G_{12} &= G_{21} \\ V_{oc} &= V_{ic} = 0 \end{aligned} \tag{11.208}$$

which result in undesired gains

$$G_{\rm dc} = G_{cd} = G_c = 0 \tag{11.209}$$

leaving only the desired gain $G_d \neq 0$. When $G_{cd} = G_{dc} = 0$, there is no conversion from input common-mode signal into output differential signal and no conversion from input differential signal into an output common-mode signal.

In practice, input signals are not perfectly differential and there exists a common signal to both inputs. The existence of a common signal results in nonzero G_c, G_{cd}, and G_{dc}, even with perfectly balanced differential amplifiers. Accordingly, besides the desired gain G_d and other parameters such as noise figure, there are three other figures of merit that determine quality of differential amplifiers. Among these figures of merit, the most important one is the "common-mode-rejection ratio" (CMRR), that measures the desired gain G_d relative to the undesired gain G_c, defined as

$$\text{CMRR} \equiv \left|\frac{G_d}{G_c}\right| \tag{11.210}$$

The two other figures of merit, although less important, are also useful in assessing the differential amplifier's performance. They are the differential-common-conversion-rejection ratio (DCRR) and common-differential-conversion-rejection ratio (CDRR) defined as

$$\text{DCRR} \equiv \left|\frac{G_d}{G_{cd}}\right| \tag{11.211}$$

and

$$\text{CDRR} \equiv \left|\frac{G_d}{G_{\rm dc}}\right| \tag{11.212}$$

The DCRR measures the differential-mode gain with respect to the magnitude of the conversion from the input common-mode signal to the output differential-mode signal. The CDRR compares the differential-mode gain with the conversion gain between the input differential- and output common-mode signals.

Consider one half of the differential-mode amplifier schematic as shown in Figure 11.31(a). This circuit is a simple common-source amplifier whose output voltage is related to that at the input as [3]

$$\frac{V_{od}}{2} = -g_m R \frac{V_{id}}{2} \tag{11.213}$$

which yields

$$G_d = \frac{V_{od}}{V_{id}} = -g_m R \tag{11.214}$$

where g_m is the MOSFET's transconductance and R is the parallel combination of R_D and R_o, with R_o being the output resistance of the MOSFET. It is noted that the voltage between the body and source (V_{bs}) of the transistors is equal to the voltage across the current source I_{CS}, which is constant for a pure differential input signal. Therefore, the body-effect transconductance $g_{mb} = \partial I_d / \partial V_{bs}$ can be neglected.

The common-mode half-circuit shown in Figure 11.33(a) is basically a common-source amplifier with degeneration and its output voltage is [3]

$$V_{oc} = -G_m R V_{ic} \tag{11.215}$$

where

$$G_m \simeq \frac{g_m}{1 + 2R_{CS}(g_m + g_{mb})} \tag{11.216}$$

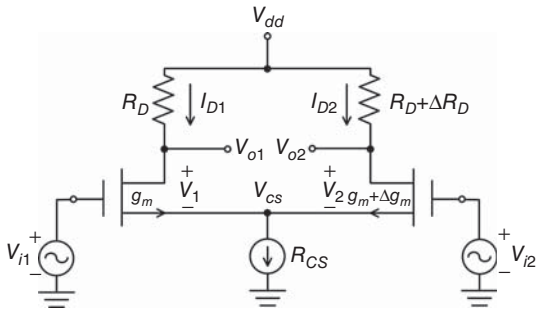


Figure 11.35. A differential amplifier with drain-resistor and transistor mismatches.

with R_{CS} being the Norton-equivalent resistance of the current source I_{CS}. The common-mode voltage gain can then be obtained as

$$G_c = -G_m R \simeq -\frac{g_m R}{1 + 2R_{CS}(g_m + g_{mb})} \tag{11.217}$$

Substituting (11.214) and (11.217) into (11.210) gives

$$\text{CMRR} \simeq 1 + 2R_{CS}(g_m + g_{mb}) \tag{11.218}$$

Mismatch Analysis. Practical differential amplifiers are imbalanced due to various reasons such as mismatches of transistors, resistors, inductors, etc., from processing and layout unsymmetry, etc. Imbalanced differential amplifiers cause problems, which can be severe if the degree of imbalance is large. Imbalance produces nonzero G_{cd} and G_{dc}, that give rises to undesired conversions of input common (differential) signals to output signals which, if sufficiently large, will corrupt the amplifiers, successive components, and/or systems containing them. Another mismatch is caused by applied input voltages; that is, the magnitude of V_{i1} is different from that of V_{i2} and/or their phase difference is not 180°. These unbalanced input voltages result in a common-mode signal, which may be significant if the imbalance is large. The resulting common-mode signal not only reduces the desired input voltage level (from the applied V_{i1} and V_{i2}), but also may cause disruption in the differential amplifiers, following components, and/or systems if largely converted into an output differential signal.

To illustrate the effects of mismatch, we consider a differential amplifier circuit under small-signal operation, as shown in Figure 11.35, in which we assume mismatches between the two drain resistors and MOSFET's. The transistors are represented by their transconductances g_m and $g_m + \Delta g_m$. The mismatch between the two resistors is denoted by ΔR_D. We also assume the constant current source has a finite (Norton-equivalent) resistance R_{CS}. The drain currents can be obtained from Figure 11.35 as

$$I_{D1} = g_m V_1 = g_m(V_{i1} - V_{CS}) \tag{11.219}$$

and

$$I_{D2} = (g_m+\Delta g_m)V_2 = (g_m+\Delta g_m)(V_{i2} - V_{CS}) \tag{11.220}$$

where

$$V_{CS} = R_{CS}(I_{D1} + I_{D2}) \tag{11.221}$$

is the voltage at the source connection. Substituting (11.219) and (11.220) into (11.221) and solving for V_{CS}, we get

$$V_{CS} = \frac{R_{CS}[g_m V_{i1} + (g_m + \Delta g_m)V_{i2}]}{1 + R_{CS}(2g_m + \Delta g_m)} \tag{11.222}$$

which becomes, upon using (11.159) and (11.160),

$$V_{CS} = R_{CS}\frac{(2g_m + \Delta g_m)V_{ic} - \frac{\Delta g_m}{2}V_{id}}{1 + (2g_m + \Delta g_m)R_{CS}} \tag{11.223}$$

The output voltages of the amplifier can be written as

$$V_{o1} = -R_D I_{D1} = -R_D g_m(V_{i1} - V_{CS}) \tag{11.224}$$

and

$$V_{o2} = -(R_D + \Delta R_D)I_{D2} = -(R_D + \Delta R_D)(g_m+\Delta g_m)(V_{i2} - V_{CS}) \tag{11.225}$$

from which, we can derive the differential-mode output voltage as

$$V_{od} = V_{o1} - V_{o2} = -g_m R_D V_{id} + (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\left(V_{ic} - V_{CS} - \frac{1}{2}V_{id}\right) \tag{11.226}$$

Substituting (11.223) into (11.224) yields

$$\begin{aligned} V_{od} = &-\left\{ g_m R_D + \frac{1}{2}(R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\left[1 - \frac{R_{CS}\Delta g_m}{1 + (2g_m + \Delta g_m)R_{CS}}\right]\right\} V_{id} \\ &+ (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\left[1 - \frac{R_{CS}(2g_m + \Delta g_m)}{1 + (2g_m + \Delta g_m)R_{CS}}\right] V_{ic} \end{aligned} \tag{11.227}$$

The differential-mode and common-to-differential-mode voltage gains can now be derived from (11.227) as

$$G_d \equiv \left.\frac{V_{od}}{V_{id}}\right|_{V_{ic}=0} = -g_m R_D - \frac{1}{2}(R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\left[1 - \frac{R_{CS}\Delta g_m}{1 + (2g_m + \Delta g_m)R_{CS}}\right] \tag{11.228}$$

and

$$G_{cd} \equiv \left.\frac{V_{od}}{V_{ic}}\right|_{V_{id}=0} = (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\left[1 - \frac{R_{CS}(2g_m + \Delta g_m)}{1 + (2g_m + \Delta g_m)R_{CS}}\right] \tag{11.229}$$

respectively. Similarly, we can derive the common-mode output voltage

$$\begin{aligned} V_{oc} &= \frac{V_{o1} + V_{o2}}{2} \\ &= \frac{1}{4}\left[R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D - (2g_m R_D + R_d\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\frac{R_{CS}\Delta g_m}{1 + (2g_m + \Delta g_m)R_{CS}}\right] V_{id} \\ &\quad - \frac{1}{2}\left[2g_m R_D + R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D - (2g_m R_D + R_D\Delta g_m + g_m\Delta R_D\right. \\ &\quad \left. + \Delta g_m\Delta R_D)\frac{R_{CS}(2g_m + \Delta g_m)}{1 + (2g_m + \Delta g_m)R_{CS}}\right] V_{ic} \end{aligned} \tag{11.230}$$

and then the common-mode and differential-to-common-mode voltage gains

$$\begin{aligned} G_c \equiv \left.\frac{V_{oc}}{V_{ic}}\right|_{V_{id}=0} = &-\frac{1}{2}\left[2g_m R_D + R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D - (2g_m R_D + R_D\Delta g_m\right. \\ &\left. + g_m\Delta R_D + \Delta g_m\Delta R_D)\frac{R_{CS}(2g_m + \Delta g_m)}{1 + (2g_m + \Delta g_m)R_{CS}}\right] \end{aligned} \tag{11.231}$$

and

$$G_{\text{dc}} \equiv \left.\frac{V_{oc}}{V_{id}}\right|_{V_{ic}=0} = \frac{1}{4}\Bigg[R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D - (2g_mR_D + R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)\frac{R_{CS}\Delta g_m}{1 + (2g_m + \Delta g_m)R_{CS}}\Bigg] \tag{11.232}$$

respectively. Finally, we can derive CMRR, DCRR, and CDRR defined in (11.210)–(11.231), making use of (11.228), (11.229), (11.231), and (11.232), as

$$\text{CMRR} = \frac{2[1 + (2g_m + \Delta g_m)R_{CS}]g_mR_D + (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)(1 + 2g_mR_{CS})}{2g_mR_D + R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D} \tag{11.233}$$

$$\text{DCRR} = \frac{2[1 + (2g_m + \Delta g_m)R_{CS}]g_mR_D + (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)(1 + 2g_mR_{CS})}{2(R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)} \tag{11.234}$$

$$\text{CDRR} = 2\frac{2[1 + (2g_m + \Delta g_m)R_{CS}]g_mR_D + (R_D\Delta g_m + g_m\Delta R_D + \Delta g_m\Delta R_D)(1 + 2g_mR_{CS})}{R_D\Delta g_m + (g_m + \Delta g_m)(1 + 2g_mR_{CS})\Delta R_D} \tag{11.235}$$

The effect of resistor mismatch and transistor mismatch alone can be obtained by setting $\Delta g_m = 0$ and $\Delta R_D = 0$, respectively. It should be noted that the amplitude and phase matching for the applied input signals for differential amplifiers may be improved by invoking the integrated design philosophy. In subsystems or systems containing differential amplifiers, in particular, or other balanced RFICs requiring input signals with certain amplitude and phase difference, in general, enhanced performance is more feasible by considering all circuits together in the design, rather than designing each individual circuit separately and then integrating them together. Integrating all circuits together in the design process (i.e., design from the subsystem or system level) will facilitate the matching of amplitude and phase of input signals for balanced circuits, which will ultimately lead to better performance.

11.5.1.3 Differential Amplifier Design. A simplified differential amplifier as shown in Figure 11.28 was used to derive simple analysis to study the amplifier's characteristics. Practical differential amplifiers are more complicated. Figure 11.36 shows a block diagram of practical differential amplifiers. As for the basic differential amplifier, the illustrated differential amplifier consists of two halves, each acting as a single-ended amplifier, coupled with each other through two coupling networks (CN1 and CN2) at the transistors' sources.

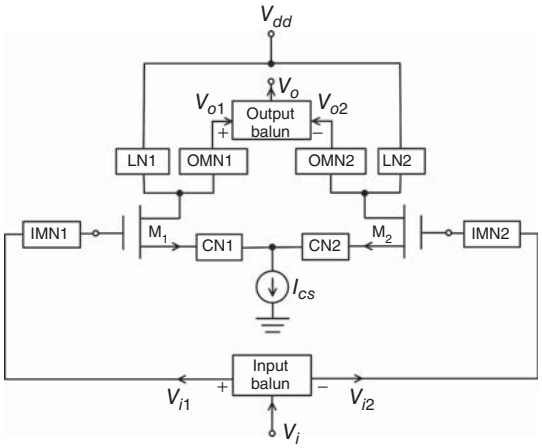


Figure 11.36. Block diagram of practical differential amplifiers. Bias circuitry is not shown. Multiple transistors can be used in each half.

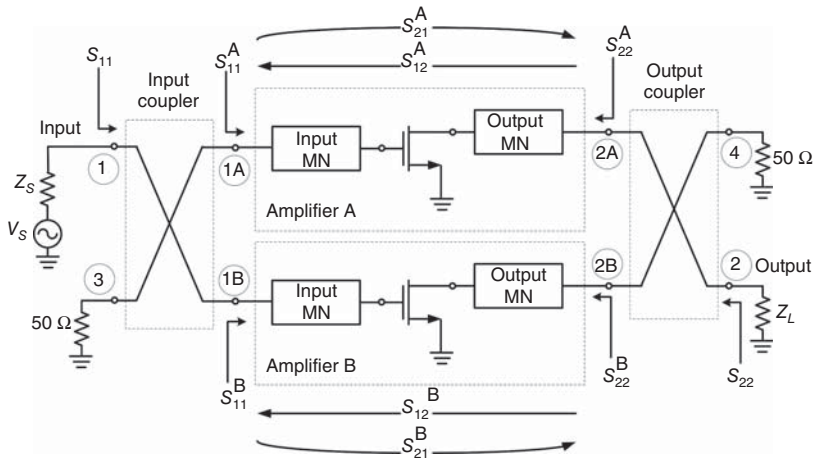


Figure 11.37. Balanced amplifier using two 90° 3-dB couplers.

These coupling networks are normally inductive degenerating networks employing (degenerating) inductors and usually identical. LN1 and LN2 are the loading networks, which may consist of linear resistors or solid-state devices such as diodes or transistors. The IMNs (IMN1 and IMN2) and OMNs (OMN1 and OMN2) are used for matching for gain and/or noise figure for the individual single-ended amplifiers. Typically, the two pairs (IMN1, OMN1) and (IMN2, OMN2) are identical. The input balun or 180° out-of-phase power divider is needed to split the input signal coming from an unbalanced device, such as a single-ended oscillator, into two equal but 180° out-of-phase signals. The output balun is used to combine the two equal but opposite-phase output signals. It is noted that the input signals (V_{i1} and V_{i2}) can come directly from a balanced device, such as a balanced antenna, and the output signals (V_{o1} and V_{o2}) can feed directly to a subsequent balanced component, such as a doubly balanced mixer, without using baluns.

It is particularly noted that a differential amplifier is essentially a combination of a single-ended amplifier and its mirror and, therefore, the design and topologies of differential amplifiers can directly implement those of single-ended amplifiers – whether they are low noise or PAs. A design approach for differential amplifiers can consist of two basic steps: (i) designing a single-ended amplifier and (ii) combining two such identical single-ended amplifiers symmetrically to form a differential amplifier. It is particularly important to make the differential amplifier symmetrical both electrically and physically. While electrical symmetry is enforced during the design stage, physical symmetry is considered in the layout process. Each of the constituent single-ended amplifiers is designed as if it was an independent amplifier for a specific design objective such as low noise or high power.

11.5.2 Ninety-Degree Balanced Amplifiers

Ninety-degree balanced amplifiers are one of the most widely used broadband amplifier topologies due to the fact that they have relatively low VSWR and flat gain with higher power handling capability as compared to unbalanced topologies. Figure 11.37 shows the block diagram of a balanced amplifier, which consists of two unbalanced amplifiers and two 90° 3-dB couplers. The couplers and amplifiers are theoretically identical, although in practice slight changes between the two couplers and amplifiers are needed for optimum performance due to design variations. The broadband matching of balanced amplifiers is not dictated by individual unbalanced amplifiers but exclusively depends on the coupler's characteristics as described in the following.

The 90° 3-dB coupler is covered in Chapter 8 and is a component that splits an input signal equally to two output ports with 90° difference. Refer to Figure 11.37 and assume that the couplers are ideal and the input signal enters (input) port 1. The input coupler splits the incoming signal at port 1 equally to ports 1A and 1B of amplifiers A and B with a 90° phase shift at 1B, respectively. The reflected signals from these amplifiers, due to mismatches at ports 1A and 1B, arrive at port 1 with the same amplitude but 180° out of phase and thus completely cancel each other. On the other hand, the signals reflecting to port 3 from ports 1A and 1B

have the same magnitude and phase, thus constructively adding together and dissipating in a 50 Ω resistor at port 3. The signals entering ports 1A and 1B are amplified by the respective amplifiers A and B via their scattering parameters S_{21}^A and S_{21}^B. Assume the two amplifiers are identical; the signal emerging from port 2B is equal to that emerging from port 2A but with 90° phase difference. Part of these signals reflects back due to mismatches and the remaining enters the output coupler. The reflected signals travel through the individual amplifiers and arrive at ports 1 and 3 with equal amplitude but 180° out of phase and in phase, respectively, and thus completely cancel and add each other at these respective ports. The signal entering the coupler from 2A divides equally to ports 2 and 4 with a 90° phase delay at 2, while that from 2B also divides equally to ports 2 and 4 but with a 90° phase delay at 4. These signals thus constructively combine at port 2 and destructively cancel at port 4. In practice, there exists a small portion of signal at port 4 due to incomplete cancelation resulting from unmatched amplifiers and couplers, which is then dissipated in the 50 Ω load resistor connected to port 4.

Mathematically, we can write the S-parameters (S_{ij}) of the balanced amplifier from those (S_{ij}^A and S_{ij}^B) of the two constituent amplifiers A and B as

$$\begin{aligned} |S_{11}| &= \frac{1}{2}|S_{11}^A - S_{11}^B| \\ |S_{21}| &= \frac{1}{2}|S_{21}^A + S_{21}^B| \\ |S_{12}| &= \frac{1}{2}|S_{12}^A + S_{12}^B| \\ |S_{22}| &= \frac{1}{2}|S_{22}^A - S_{22}^B| \end{aligned} \tag{11.236}$$

from which, we can obtain, assuming identical amplifiers A and B,

$$\begin{aligned} S_{11} &= S_{22} = 0 \\ |S_{21}| &= |S_{21}^A| = |S_{21}^B| \\ |S_{12}| &= |S_{12}^A| = |S_{12}^B| \end{aligned} \tag{11.237}$$

which indicate that the balanced amplifier is perfectly matched at both the input and output ports and its gain is equal to the gain of each individual single-ended amplifier.

It is noted that the S_{11} and S_{22} of the balanced amplifier are equal to the S_{11} and S_{22} of each of the two identical couplers, respectively, while S_{21} is the same as that of individual amplifiers. Therefore, the matching bandwidth of the balanced amplifier is limited only by that of the couplers with its input and output VSWR's depending completely on the couplers' VSWR. Its gain, however, is exclusively given by the gain of each individual amplifier and hence the bandwidth for the gain depends on the constituent amplifiers, which may be different from the bandwidth of the couplers. One can thus design the individual amplifier for desired performance, such as gain and noise figure, across a desired frequency range without paying attention to the VSWR and then connect two identical amplifiers with two couplers, whose bandwidth and VSWR match those desired for the balanced amplifier. For instance, the individual amplifier can be intentionally designed for mismatch to achieve certain gain flatness, noise figure, output power, and stability over desired frequencies. The VSWR-independent design is perhaps the main attractiveness of balanced amplifiers. Furthermore, the balanced amplifier's output power is twice of that of each single amplifier (i.e., 3 dB higher) and the balanced amplifier can still operate with reduced gain (around 6 dB less) even when one of the amplifiers fails. Moreover, the fact that the reflections due to mismatch from individual amplifiers are terminated in 50-Ω resistors helps improve the balanced amplifier's stability. Finally, a balanced amplifier facilitates connection with other components due to the inherent isolation resulting from the couplers. However, balanced amplifiers also consume large die size and DC power, two important facts that are not viewed favorably in RFICs, particularly for portable RF devices. The design procedure for balanced amplifiers is relatively simple and

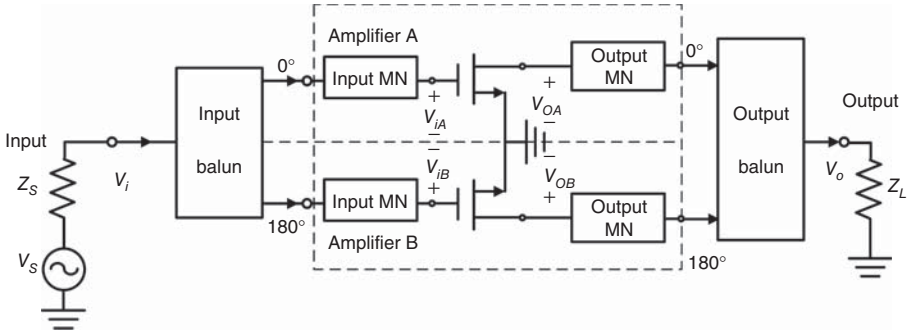


Figure 11.38. Push–pull amplifier.

consists of four steps: (i) design an (unbalanced) amplifier to achieve desired specifications such as gain and noise figure over a desired bandwidth; (ii) design a broadband 90-° 3-dB coupler covering the required bandwidth and VSWR of the balanced amplifier; (iii) connect two identical (unbalanced) amplifiers using two identical couplers to form a balanced amplifier; and (iv) perform final optimization of the entire balanced amplifier.

11.5.3 Push–Pull Amplifiers

Push–pull amplifiers are attractive for various system applications due to their advantages such as less distortion and high output power as compared to single-ended amplifiers. Figure 11.38 shows the general topology of the push–pull amplifiers consisting of two (independent) unbalanced amplifiers and two baluns (or 180° hybrids). Theoretically, the two baluns and amplifiers are respectively identical. In practice, however, slight differences between the two baluns and between amplifiers may be needed to achieve optimum performance due to design variations. In contrast to the 90° balanced amplifier, the matching of the push–pull amplifier depends on both the baluns and individual unbalanced amplifiers.

In general, a balun or 180° hybrid (discussed in Section 8.3) splits an incoming signal into two signals of equal amplitude and 180° out of phase or combines two such signals constructively into a single signal. Consider Figure 11.38 and assume the baluns are ideal, the individual amplifiers are exactly the same, and the interconnects between each of the baluns and amplifiers A and B are identical. The input balun splits the input signal equally with 180° phase difference to the input ports of amplifiers A and B. The reflected signals from these amplifiers, due to mismatches between their input ports and the output ports of the input balun, have the same amplitude but with 180° out of phase, and are thus constructively combined by the input balun and appear at the input port of the push–pull amplifier, causing undesired mismatch at the amplifier's input. The signals entering amplifiers A and B are amplified and appear at their output ports with equal magnitude and 180° phase difference. These signals are then constructively combined by the output balun and appear at the output port of the push–pull amplifier. Any reflected signals due to mismatches between the output balun and amplifiers A and B travel back to the amplifiers. These signals are combined by the input balun and appear at the push–pull amplifier's input port, resulting in further mismatch at the amplifier's input. This mismatch contribution, however, is relatively small due to the typically high reserve gain (or isolation) of each individual amplifier. The input balun thus does not cancel out the reflected signals due to each individual amplifier's mismatch with the baluns, which is contrary to the input 90° coupler used in the 90° balanced amplifier, resulting in potentially poorer input matching. This is the main disadvantage of the push–pull amplifier as compared to the 90° balanced amplifier. The output matching of the push–pull amplifier, however, depends mainly on the matching between the output balun and the load. The noncancelation of reflected signals from the input balun also results in reduced isolation between the two individual amplifiers, causing potential instability for the push–pull amplifier. Both narrow and wideband designs can be achieved for push–pull amplifiers provided that the input and output matching of the input and output baluns, respectively, and the matching between the baluns and the amplifiers are obtained properly over the desired bandwidth.

The input voltages to the transistors of amplifiers A and B can be expressed as

$$V_{iA} = -V_{iB} = V_m \cos \omega t \tag{11.238}$$

and the corresponding output voltages can be obtained via power-series expansion of the respective input voltages as

$$V_{oA} = \sum_{n=0}^{N} A_n \cos(n\omega t) \tag{11.239}$$

$$V_{oB} = \sum_{n=0}^{N} (-1)^n A_n \cos(n\omega t) \tag{11.240}$$

where A's are constant coefficients depending on the bias of MOSFETs used in the amplifiers. The output voltage of the push–pull amplifier, without the OMNs, is obtained as the difference between the individual output voltages as

$$V_o = V_{oA} - V_{oB} = 2 \sum_{n=1,3,5,\ldots}^{N} A_n \cos(n\omega t) \tag{11.241}$$

which contains only odd-order terms. Push–pull amplifiers thus inherently suppress all the spurious signals of even orders and have no DC component, leaving the third-harmonic as the principal source of distortion and signifying reduced distortion as compared to their single-ended counterparts. The (fundamental) output power is twice of that produced by single-ended counterparts. When one of the constituent amplifiers fails, the push–pull amplifier may still operate but with less output power, provided that there is no significant reflection from the failed amplifier. Furthermore, because the output of push–pull amplifiers contains no even harmonics, such amplifiers will produce more output power per transistor for a given amount of signal distortion. In addition, the fact that the output voltage has no even harmonics means that the push–pull amplifiers possess "half-wave" or "mirror" symmetry across the central axis between the individual amplifiers, implying that the bottom half of the amplifier (i.e., amplifier B), when shifted 180° along the axis, will be the mirror image of the top half (amplifier A). This results in a virtual ground along the axis that acts as an RF ground, hence eliminating the need of using external elements to connect the MOSFETs to ground. This is desirable for increased gain and bandwidth since no parasitic inductance from the MOSFETs' sources to the ground is encountered. It should be noted that these ideal results are obtained assuming two identical individual amplifiers. Otherwise, even harmonics will appear, causing degradation to circuit performance. Similar to the 90° balanced amplifiers, push–pull amplifiers also consume considerable DC power and have large die size, which are undesirable for RFICs, especially those used in portable devices. The design of push–pull amplifiers consists of four simple steps: (i) design a single-ended amplifier to achieve desired performance such as maximum gain over a design bandwidth; (ii) design a balun covering the design bandwidth; (iii) combine two identical amplifiers through two identical baluns to form a push–pull amplifier; and (iv) perform optimization of the whole push–pull amplifier. It should be noted that the matching between the input and output baluns and the individual amplifiers is critical as this affects the matching of the entire amplifier. Typical push–pull amplifier design is based on 50-Ω baluns and individual amplifiers. However, in integrated design, while the input of the input balun and the output of the output balun need to match to 50 Ω for measurement purposes or integration with other 50-Ω components, the output of the input balun and the inputs of the output balun only need to conjugately match to the input and output of the individual amplifiers, respectively. This design scheme eliminates the 50-Ω interstage matching and reduces the overall circuit size.

11.6 BROADBAND AMPLIFIERS

The analysis and design of amplifiers such as those formulated in this chapter are frequency dependent and can be used directly for amplifier design operating at a single frequency and, to some extent, over a narrow bandwidth. Many RF systems are wideband and require the use of broadband amplifiers. While the presented analysis and design are still applicable, other considerations need to be implemented in order to take into account the frequency dependence of passive and active elements making up the amplifiers. MOSFETs, for instance, have S-parameters (S_{11}, S_{12}, S_{21}, S_{22}) and noise parameters (r_n, $F_{\min}$, Γ_{So}) dependent on frequency. To achieve a particular bandwidth for specified gain and/or noise figure, certain design techniques and topologies must be employed to take into account the variations of the S-parameters and noise parameters of MOSFETs versus frequency and minimize the effects of these variations on the matching networks, as well as reduce the frequency sensitivity of the matching networks.

Commonly used techniques for broadband amplifier design employ compensated matching network, balanced topology as described in Section 11.5, distributed structure, negative feedback, and cascoded common-source topology. These broadband amplifier topologies are covered in this section.

11.6.1 Compensated Matching Networks

The gain of transistors, dictated by their S_{21} magnitude, typically reduces with frequency and the underlined principle of the compensated matching-network technique is to compensate for this gain roll-off by correspondingly varying the gain of the IMNs and OMNs. Note that this gain actually refers to the loss of the passive matching networks. Figure 11.39 illustrates this design technique for a single-transistor amplifier. As can be seen, the IMNs and OMNs are not designed to achieve good matching over the amplifier's bandwidth in order to compensate for the reduction of the transistor's gain. This technique results in poor matching for the amplifier and is thus not very practical and desirable.

To overcome the unwanted mismatch at the input and input ports, multiple transistors with interstage matching networks can be implemented. Figure 11.40 shows this design technique using two transistors. The IMNs and OMNs are designed to provide good match over a desired bandwidth while the interstage networks are designed to achieve increasing matching or increasing gain with frequency to compensate for the gain reduction of the transistors versus frequency. Although this technique produces good matching, it also

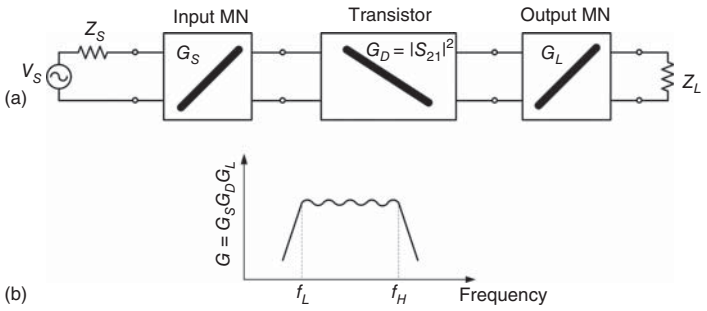


Figure 11.39. Design principle of a single-transistor broadband amplifier using compensated matching networks (MN) with block diagram (a) and overall gain (b). The straight line in each block depicts the corresponding gain variation with frequency over interested bandwidth.

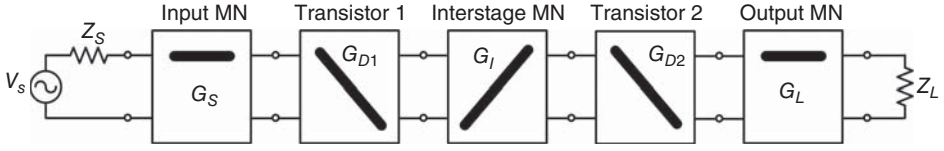


Figure 11.40. Design principle of multi-transistor broadband amplifier using compensated matching networks (MN). The straight line in the each block depicts the corresponding gain variation with frequency over an interested bandwidth.

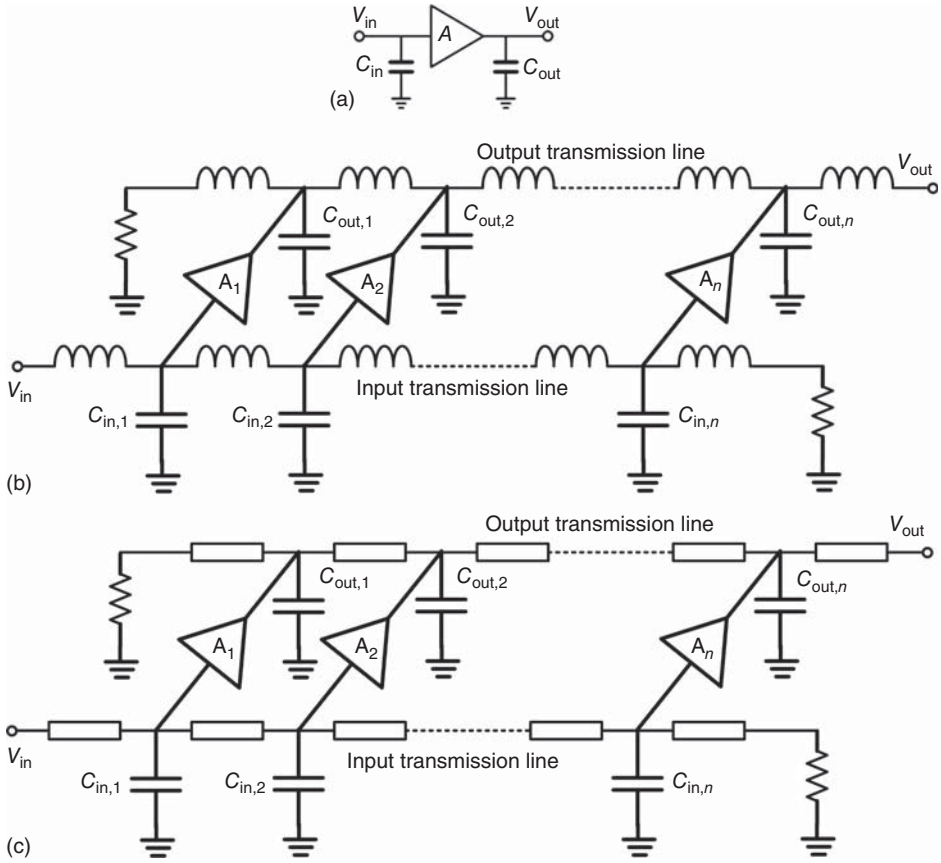


Figure 11.41. A simple generic amplifier with its input C_{in} and output C_{out} (parasitic) capacitors (a) and distributed amplifiers employing them as gain cells with inductors (b) and transmission lines (c). $C_{in,i}$ and $C_{out,i}$ $(i = 1, 2, \ldots, n)$ are the parasitic capacitances at the input and output ports of amplifier A_i, respectively.

inadvertently reduces the inherent gain of the transistors at low frequencies, thus resulting in non-optimum design.

11.6.2 Distributed Amplifiers

Distributed-circuit concept was first proposed in 1937 by Percival [4] and distributed amplifiers were reported by Ginzton et al. in 1948 [5]. Distributed amplifiers achieve ultrawide bandwidth, approaching several decades, by using the principle of transmission lines which have inherently wide operating bandwidth, typically from DC to a very high frequency above which some adverse effects, such as higher-order modes different from the normal transmission electron microscope (TEM) or quasi-TEM mode, appear. They also have excellent linearity. Distributed amplifiers, however, suffer from large size, relatively low gain and high power consumption. These drawbacks are mostly concerned for commercial CMOS/BiCMOS implementations, especially portable devices. Enhanced design, overcoming these issues, is also discussed in this section.

11.6.2.1 Principle and Analysis of Distributed Amplifiers. In order to demonstrate the principle of distributed amplifiers in general, without limiting to conventional distributed amplifiers implementing single transistors, we begin by considering a basic amplifier with two parasitic capacitors at the input and output ports as shown in Figure 11.41(a). The two parasitic capacitances are undesirable, yet they represent the capacitances actually appearing at the input and output of typical circuits including amplifiers. These capacitances, in general, come from both internal elements within the circuit itself and external elements which the circuit is connected to. For instance, the internal elements of a single-MOSFET amplifier are the MOSFET's

gate and drain–source capacitances and the external elements can be the RF on-wafer pads.[7] From a simple circuit analysis, these two parasitic capacitances form two poles that limit the bandwidth of this amplifier. Figure 11.41(b) and (c) show two amplifiers consisting of several of the basic amplifiers connected in parallel that are interconnected via inductors and transmission lines, respectively. The basic amplifiers, each functioning as a gain cell or stage, are distributed across the composite amplifier known as distributed amplifier. It is noted that identical gain cells are used herein for illustration purpose; in practice, nonidentical gain cells may need to be employed to optimize the amplifier performance. It is also particularly noted that the performance, particularly gain, of distributed amplifiers depends much on the employed gain cells, so proper choices of gain cells are important in the design of distributed amplifiers. Choices of gain cells are discussed in Section Design of Low Power Consumption and High Gain CMOS Distributed Amplifiers, which may lead to enhanced performance for distributed amplifiers.

Now we recall the principle of synthetic transmission lines discussed in Section 4.9, which states that a synthetic transmission line can be formed by combining shunt capacitors and series inductors and, depending upon values of the capacitors and inductors, and the number of these elements, it can operate over an extremely wide bandwidth. It is thus apparent that $C_{\text{in},i}$, where $i = 1, 2, \ldots, n$, and output capacitance $C_{\text{out},i}$ seen in the amplifiers of Figure 11.41(b) and (c) can be combined with proper series inductances [Figure 11.41(b)] or actual transmission lines [Figure 11.41(c)] to form two synthetic transmission lines – one at the inputs of the gain cells and another one at the outputs of the gain cells – effectively making the entire amplifier to behave as transmission lines that can operate over an extremely wide bandwidth. The resultant so-called distributed amplifier basically synthesizes two ultra-wideband (UWB) synthetic transmission lines and so its bandwidth is primarily limited by that of these transmission lines.

Now, considering Figure 11.41(b) and (c), the input capacitance $C_{\text{in},i}$ and output capacitance $C_{\text{out},i}$ either combine with the inductances [Figure 11.41(b)] or are absorbed into the actual transmission lines [Figure 11.41(c)] along the input and output, respectively, to form two synthetic transmission lines – one at the input and another one at the output. Since typical gain cells, for example, those shown in Figure 11.57, have input at the gate and output at the drain of MOSFETs, hereafter we will refer to these transmission lines as the "gate transmission line" and "drain transmission line," respectively. As these capacitances are combined with the inductances or absorbed into the real transmission lines, they no longer form the poles by themselves, thus eliminating the pole's effect on the bandwidth. The bandwidth of the distributed amplifier is now determined mainly by the cut-off frequency of the two synthetic transmission lines. To accommodate the loading input and output capacitances, the inductances or the characteristic impedances of the actual transmission lines should be properly chosen to produce desired characteristic impedances for the corresponding synthetic transmission lines – for instance, the characteristic impedances of the real transmission lines should be greater than those desired for the synthetic transmission lines, which are normally 50 Ω, for matching purpose. Moreover, low loss inductors or transmission lines are needed to obtain well-behaved synthetic transmission lines, which are critical for realizing distributed amplifiers. Mutual coupling between adjacent inductors in each transmission line and between the gate and drain transmission lines should be carefully taken into consideration to avoid possible adverse effects on the amplifier's performance and to optimize the amplifier's size. Less coupling between adjacent inductors, for instance, allows the inductors to be placed close to each other, hence reducing the amplifier's size.

The distributed amplifier operates as follows. As an input RF signal is incident at the input port of the amplifier, it travels down the gate transmission line to the first gain cell. The input of each gain cell is excited by the traveling wave and the gain cell transfers the amplified signal to its output, which then traverses the drain transmission line. If the electrical length, or equivalently the propagating velocity and (equivalent) physical length, of the gate and drain lines are equal and the phase delay across each gain cell is identical, then the signal generated at the output of the first gain cell will add coherently to the signal generated at the output of the second gain cell. The signal generation process is repeated at subsequent gain cells and, finally, the total signal at the output of the amplifier will be the sum of all the signals produced by all these

[7] In integrated circuits where the amplifier is connected directly with other circuits such as mixer, the on-wafer pads are needed only for the amplifier measurement but not in the integrated circuits.

gain cells. Any RF signal traversing the drain transmission line in opposite direction will be absorbed by the termination at the other end of the drain line. Along the gate transmission line, any signal passing the last gain cell will be dissipated in the gate line's termination. The distributed amplifier actually acts a traveling wave amplifier, a wideband amplifier topology widely known among microwave engineers.

The (power) gain of distributed amplifiers can be estimated, assuming the (synthetic) transmission lines either are lossless or have very low loss, as [6]

$$G \simeq \frac{1}{4} N^2 g_m^2 Z_g Z_d \tag{11.242}$$

where N is the number of gain cells or stages, g_m is the transconductance of each gain cell, and Z_g and Z_d are the characteristic impedances of the gate and drain transmission lines, respectively. Equation (11.242) shows that, under the ideal condition of lossless or very low loss synthetic transmission lines with given Z_g and Z_d, the gain of distributed amplifiers depends only on N and g_m and it can be increased with increasing number of gain cells.

In practice, however, the loss of inductors [Figure 11.41(b)] or actual transmission lines [Figure 11.41(c)] and hence synthetic transmission lines should not be neglected, especially at high frequencies. Inductor and transmission line's loss is particularly more pronounced in CMOS/BiCMOS process due to the use of relatively lossy silicon substrate and narrow inductors' trace and transmission lines resulting from very thin Oxide layers (typically few microns) above silicon substrate. When the loss of the (synthetic) gate and drain transmission lines is taken into account, the gain of distributed amplifiers can be derived as [7]

$$G \simeq g_m^2 Z_g Z_d \left| \frac{\overline{\gamma}_g \left(e^{-\overline{\gamma}_g N} - e^{-\overline{\gamma}_d N} \right)}{\overline{\gamma}_g^2 - \overline{\gamma}_d^2} \right|^2 \tag{11.243}$$

where Z_g and Z_d are the characteristic impedances of the gate and drain lines, respectively; $\overline{\gamma}_{g(d)} = \overline{\alpha}_{g(d)} + j\overline{\beta}_{g(od)}$ is the (total) propagation constant of the input (output) line per transmission-line section[8] with $\overline{\alpha}_{g(d)}$ and $\overline{\beta}_{g(d)}$ being the corresponding (total) attenuation constant (neper) and phase constant (radian) of each transmission-line section. This phase constant is actually the electrical length of each section. Equation (11.243) was derived based on actual transmission lines along the gates and drains. It, however, can also be used for distributed amplifiers employing inductors along the gates and drains with Z_g, Z_d, and $\overline{\gamma}_{g(d)}$ being those for inductor-based synthetic transmission lines. Assuming each gain cell is realized using identical common-source MOSFET, the characteristic impedances and propagation constants of the transmission line-based and inductor-based gate and drain transmission lines can be approximately determined using expressions derived in Section General Design Approach.

In typical distributed-amplifier design, the electrical lengths across each section of the gate and drain lines are approximately equal[9] ($\overline{\beta}_g \simeq \overline{\beta}_d$). Under this condition and the assumption of low loss transmission lines with $|\overline{\gamma}_{g(d)}| \simeq |\overline{\beta}_{g(d)}|$, the gain in (11.243) can be simplified to

$$G \simeq \frac{g_m^2 Z_g Z_d (e^{-\overline{\alpha}_g N} - e^{-\overline{\alpha}_d N})^2}{4(\overline{\alpha}_g - \overline{\alpha}_d)^2} \tag{11.244}$$

As expected, more stages, or in turn more synthetic transmission lines, result in more loss and hence limited gain, approaching zero for a very large number of stages. Consequently, the gain cannot be increased indefinitely by simply increasing the number of stages. The loss of the synthetic transmission lines affects not only the magnitude of the gain, but also its ripple. This loss thus needs to be minimized as much as possible, effectively making the synthetic transmission lines behave as well as possible. One special loss component of

[8]This propagation constant is not the propagation constant (per unit length) typically used for transmission lines.

[9]In fact, this condition needs to be satisfied; otherwise, the output signals of a gain cell and its subsequent one cannot be constructively added.

the synthetic transmission lines, whether they are formed from on-chip inductors or real transmission lines, is the resistive loss of the metal which also dissipates DC power. Reducing this loss particularly helps improve the DC power consumption of the amplifiers.

An optimum number of stages that gives maximum gain can be determined by letting the derivative of G with respect to N equal to zero. Doing so using (11.244) results in

$$\frac{dG}{dN} \simeq \frac{g_m^2 Z_g Z_d (e^{-\overline{\alpha}_g N} - e^{-\overline{\alpha}_d N})(-\overline{\alpha}_g e^{-\overline{\alpha}_g N} + \overline{\alpha}_d e^{-\overline{\alpha}_d N})}{2(\overline{\alpha}_g - \overline{\alpha}_d)^2} = 0 \tag{11.245}$$

from which, assuming $\overline{\alpha}_g \neq \overline{\alpha}_d$,

$$-\overline{\alpha}_g e^{-\overline{\alpha}_g N} + \overline{\alpha}_d e^{-\overline{\alpha}_d N} = 0 \tag{11.246}$$

Solving (11.246) for N gives

$$N_{\text{opt}} = \frac{\ln(\overline{\alpha}_d / \overline{\alpha}_g)}{\overline{\alpha}_d - \overline{\alpha}_g} \tag{11.247}$$

Typically, the losses of the gate and drain transmission lines are approximately equal and, under this case, the maximum gain using an optimum number of gain cells can be expressed as [8]

$$G_{\max} = \frac{Z_g Z_d}{16} \left(\frac{A g_m}{e \overline{\alpha}} \right)^2 \tag{11.248}$$

where $\overline{\alpha} = \overline{\alpha}_g = \overline{\alpha}_d$, $e = 2.71828$ is the *Euler* constant, and A is a constant depending on the transistor and frequency. This expression shows the direct relation between the amplifier's gain and the loss of the synthetic transmission lines. The corresponding optimum number of stages can be estimated as

$$N_{\text{opt}} \simeq \frac{1}{2\overline{\alpha}} \tag{11.249}$$

As can be seen from (11.244) and (11.247), the gain of distributed amplifiers having an optimum number of stages depends only on Z_g, Z_d, $\overline{\alpha}_g$, $\overline{\alpha}_d$, and g_m. In practice, synthetic transmission lines used as the gate and drain transmission lines have to match to the impedance of the input and output ports, respectively, which is normally 50 Ω. Consequently, the gain of distributed amplifiers can be optimized by minimizing the loss of the synthetic transmission lines (or equivalently the loss of corresponding inductors or actual transmission lines) and increasing the transconductance of each gain cell. These optimizations for the transmission lines and gain cells are done independently with each other, thus facilitating the amplifier design process.

The principle of the foregoing general distributed amplifiers is now specifically applied to conventional distributed amplifiers employing single MOSFETs as gain cells. To that end, we first consider a simplified unilateral MOSFET model as shown in Figure 11.42. Although all practical devices are bilateral, the assumption of unilateral device does not generally cause significant error in circuit design and, particularly for the

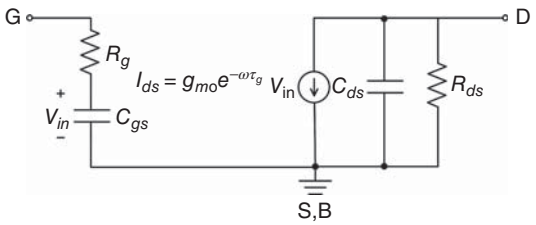


Figure 11.42. A simplified model for unilateral MOSFETs. The source (S) and bulk (B) terminals are tied together. g_{mo} is the DC transconductance, τ_g is the gate transit time, C_{gs} and C_{ds} are the gate–source and drain–source capacitance, respectively, R_g is the gate resistance, and R_{ds} is the drain–source resistance.

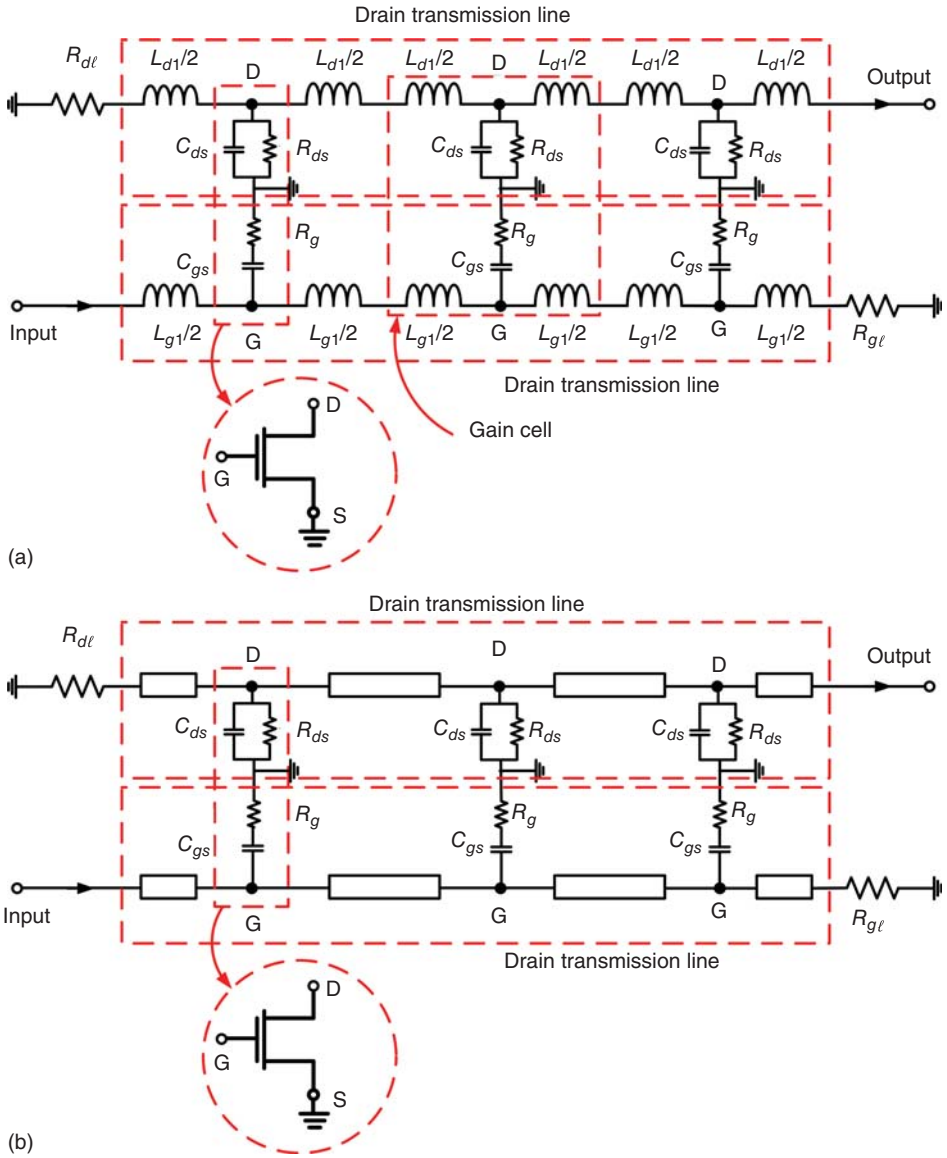


Figure 11.43. Simplified equivalent circuit of a three-stage distributed amplifier with inductors (a) and real transmission lines (b). I_{ds} of each transistor is not shown.

illustration of the distributed concept, it is well justified. This simplified model shows that the equivalent circuit of the device at the gate terminal is approximately represented by the gate–source capacitance C_{gs} and the gate resistance R_g while that at the drain terminal is approximated by the drain–source capacitance C_{ds} and the (output) drain–source resistance R_{ds}. R_g and R_{ds} for good transistors are typically small and large, respectively. It is apparent, from the principle of synthetic transmission lines, that C_{gs} and C_{ds} of MOSFETs employed in an amplifier can be combined with proper series inductances to form two synthetic transmission lines – one at the gates and another one at the drains – effectively making the entire amplifier to behave as transmission lines which have extremely wide bandwidth.

Without loss of generality, we illustrate the distributed-amplifier principle using three identical MOSFETs, whose models are shown in Figure 11.42, and inductors. Figure 11.43 shows this three-stage distributed amplifier employing inductors and actual transmission lines, neglecting the current I_{ds}. As can be seen, there are two synthetic transmission lines: the gate transmission line and the drain transmission line along the gates and drains, respectively. As in general distributed amplifiers, in practice, nonidentical transistors may need to be employed to optimize the performance.

As an input signal travels along the gate transmission line, it generates voltages across the MOSFETs' capacitors C_{gs}. These voltages in turn produce currents at the drains of the MOSFETs which then traverse the drain transmission line. If the electrical length, or equivalently the velocity and (equivalent) physical length, of the gate line is equal to that of the drain line, then the signals on the drain line add in the forward direction and arrive at the output of the distributed amplifier. Reflected signal traveling in the reverse direction along the drain line will be absorbed by the terminating resistor $R_{d\ell}$ at the drain line. Any remaining signal traveling pass the last MOSFET on the gate line will be absorbed by its terminating resistor $R_{g\ell}$.

An important property of the distributed amplifiers is their improved linearity compared to (conventional) nondistributed amplifiers. It was proved in [9] that the ratio between the output power at the fundamental frequency (carrier) and the third-order intermodulation (IM3) power is drastically improved in multistage distributed amplifiers as compared to nondistributed multistage amplifiers. We consider the transconductance of the active devices as the main source of nonlinearity and assume that the gate and drain transmission lines are matched at the input and output ports, respectively, and have equal electrical lengths at the fundamental frequency including two-tone signals (used for IM analysis) and IM3 frequencies. The carrier-to-third-order-intermodulation-power ratio, C/IM3, for the distributed amplifiers can be derived as [9]

$$\frac{C}{\text{IM3}} = \left(\frac{4g_{m1}^3}{12I_{do}^2 g_{m3}} \right)^2 e^{-2N\overline{\alpha}_d} \left\{ \frac{\sinh^3\left[\frac{N}{2}\left(\overline{\alpha}_g - \overline{\alpha}_d\right)\right] \sinh\left[\frac{1}{2}\left(3\overline{\alpha}_g - \overline{\alpha}_d\right)\right]}{\sinh^3\left[\frac{1}{2}\left(\overline{\alpha}_g - \overline{\alpha}_d\right)\right] \sinh\left[\frac{N}{2}\left(3\overline{\alpha}_g - \overline{\alpha}_d\right)\right]} \right\}^2 \tag{11.250}$$

where g_{m1} and g_{m3} are the first- and third-order small-signal transconductance coefficients of the nonlinear current–voltage (I_{ds}–V_{gs}) function. These coefficients can be obtained using the conventional I_{ds}–V_{gs} curve at certain V_{ds} or through a power-series expansion of the drain–source current I_{ds} as a function of the input voltage. I_{do} is the total current flowing through the load impedance at the drain line (output load) and is kept constant as the number of stages N is increased. Again, $\overline{\alpha}_g$ and $\overline{\alpha}_d$ denote the loss or (total) attenuation constant (neper) per section of the input and output transmission line, respectively. When $(3\overline{\alpha}_g - \overline{\alpha}_d)$ and $(\overline{\alpha}_g - \overline{\alpha}_d)$ are small, (11.250) can be simplified as

$$\frac{C}{\text{IM3}} = \left(\frac{g_{m1}^3}{3I_{do}^2 g_{m3}} \right)^2 N^4 e^{-2N\overline{\alpha}_d} \tag{11.251}$$

which shows that the ratio C/IM3 increases by $N^4 e^{-2N\overline{\alpha}_d}$ as the number of stages is increased. This increase in linearity demonstrates a significant advantage of distributed amplifiers as compared to nondistributed amplifiers whose C/IM3 is reduced as more stages are added while keeping the output power at the fundamental frequency constant.

11.6.2.2 Distributed Amplifier Design Using Inductors. On-chip inductors can be used in conjunction with MOSFETs' parasitic capacitances to realize synthetic transmission lines needed for distributed amplifiers. On-chip inductors, while lending themselves directly and conveniently into the basic distributed amplifier design approach using synthetic transmission lines, may not, in general, have values and quality factor (Q) suitable for a particular design in typical CMOS processes. The low Q is due to significant loss of highly doped silicon substrates in the gigahertz range and is particularly concerned for RFIC design. On-chip inductors may also occupy such a large size that is undesirable for RFICs. Using advanced nanometer CMOS technologies, however, results in very small inductances (in the range of several hundred pico-henries) needed for the synthetic transmission lines in CMOS distributed amplifiers. These small inductances facilitate not only the design, but also the Q-enhancement of on-chip inductors. Additionally, the availability of thick top metal layers in advanced CMOS processes also leads to further Q improvement for on-chip inductors. Altogether, these make it possible for distributed amplifiers implementing on-chip inductors in advanced CMOS processes to achieve simultaneously good performance and small die area.

The inductance of on-chip inductors can be calculated from the standard lossless equation for characteristic impedance as

$$L = CZ_o^2 \tag{11.252}$$

where C stands for either C_{gs} or C_{ds} corresponding to the gate or drain transmission line, respectively, and Z_o is the characteristic impedance of these (assumed lossless) transmission lines. The corresponding cut-off frequency of the gate or drain transmission line is derived approximately in Eqs. (4.241) and (4.242) as

$$f_c = \frac{1}{\pi\sqrt{LC}} \tag{11.253}$$

or

$$f_c = \frac{1}{\pi C Z_o} \tag{11.254}$$

for given or desired Z_o. This cut-off frequency basically dictates the bandwidth of distributed amplifiers which, according to (11.254), decreases as the (parasitic) capacitance C increases. Typically, Z_o is fixed for matching purpose and so C represents the remaining parameter controlling the amplifier's bandwidth. As the gate width of a MOSFET is increased, its capacitance increases while its transconductance, and hence the gain of corresponding distributed amplifiers, increases. Equation (11.254) thus can be used for possible trade-off between the gain and bandwidth in distributed amplifier design. Typically, same characteristic impedance is used for both the gate and drain lines. Furthermore, as discussed earlier, both these lines need to have equal (equivalent) physical length which, together with the same velocity resulting from equal characteristic impedance, produces the same phase delay across these lines. As for any RFIC design, actual on-chip inductors with accurate models need to be used in the final analysis and optimization of the amplifier to obtain accurate performance.

Design Example of Distributed Amplifier Using Inductors

Figure 11.44 shows the schematic and photograph of a CMOS distributed amplifier using on-chip inductors [10]. Four stages of 0.25-μm MOSFETs, with each having a width of 160 μm, are used in the design, considering trade-off between gain, bandwidth and chip area. The characteristic impedance of the gate and drain transmission lines formed by the on-chip inductors and MOSFETs' parasitic capacitors at the input and output ports and their corresponding terminating resistors R_{gt} and R_{dt} are 50 Ω. To obtain the desired

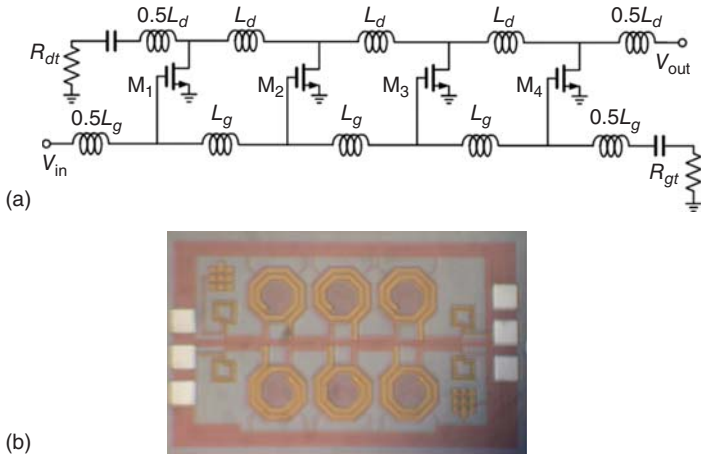


Figure 11.44. Schematic (a) and photograph (b) of a four-stage CMOS distributed amplifier with on-chip inductors. The die area is 1.2×0.8 mm^2. The capacitors at the terminating resistors are DC blocking capacitors. (After Guan, Jin, and Nguyen [10]. © The Institution of Engineering and Technology.)

bandwidth and satisfy the input and output matching conditions, the inductances (L_g and L_d) are estimated as 1 and 0.85 nH, respectively, after calculations and optimization, in order to accommodate the parasitic capacitances of the employed transistors. However, in order to make the design simple, L_d is also chosen to be equal to L_g (1 nH). Calculations show that using this increased value for L_d boosts the gain slightly (about 0.9 dB), while causing only a small degradation at the output matching, which is still acceptable for the design purposes.

The six on-chip inductors in the central area, as can be seen in Figure 11.44(b), are realized using two-turn octagonal spirals with patterned ground shield (PGS) discussed in Section 3.4.5. Each inductor has a diameter of 220 μm and a trace width of 30 μm. The first and second top-most metal layers are used for the spiral strip and under-path, respectively, while the lowest metal layer is used for the PGS. These inductors are designed to have a resonant frequency of 16.6 GHz, which is higher than the specified upper operating frequency of 11 GHz of the designed distributed amplifier. Their calculated Q peaks about 12 at around 8 GHz.

The use of on-chip PGS inductors results in high Q and reduced (silicon) substrate coupling among adjacent inductors, enabling high circuit performance and small die size, respectively. The reduced substrate coupling is particularly attractive for distributed amplifiers that typically employ many inductors. PGS inductors, however, also cause several problems due to unwanted parasitic capacitances produced by the PGS. First, the PGS introduces additional parasitic capacitances to the inductors, resulting in lower resonant frequencies than those of inductors without PGS, as discussed in Section 3.4.5. This may lead to serious problems in distributed amplifier design (or virtually any UWB RFIC design), whose bandwidth is so wide that its upper frequency is near or beyond the PGS inductors' resonant frequencies. Second, to maintain constant characteristic impedance for the synthetic transmission lines, a higher inductance for the PGS inductors is needed to off-set the increase in parasitic capacitance due to PGS. The increased inductance and parasitic capacitance, however, inevitably reduce the bandwidth according to Eq. (11.253). Third, the characteristic impedance of the synthetic transmission lines decreases at high frequencies due to the increase and decrease of the capacitance and inductance, respectively, hence affecting both the input and output matching at the upper operating frequencies. Finally, the inductance and Q of PGS inductors vary more than those of inductors without PGS from DC to the resonant frequency of the PGS inductors. This increased variation in inductance adds another challenge to the implementation of PGS inductors for CMOS distributed amplifier design. The remaining four inductors, as can be seen near the input and output ports in Figure 11.44(b), are typical on-chip spiral inductors with inductance of 500 pH. One turn without PGS is used for these inductors since it is relatively easy to obtain high Q for such small inductance.

Figure 11.45 shows the effect of inductance on the performance of the distributed amplifier using ideal inductors for the six PGS inductors. The results show that a change of inductance from 0.8 to 1.0 nH only has a slight effect on the amplifier's gain and input matching across 1–15 GHz. As expected, the decrease in inductance lowers the gain and slightly improves the bandwidth. These resultant gain and bandwidth due to change of inductance, however, are tolerable in typical amplifier design.

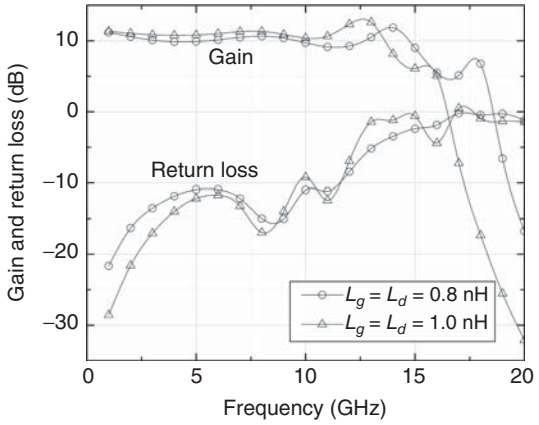


Figure 11.45. Simulated gain and input return loss of the distributed amplifier with different values for ideal inductors.

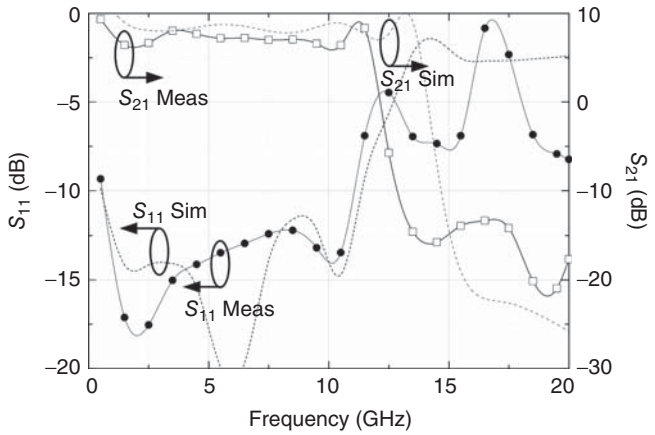


Figure 11.46. Simulated and measured input reflection coefficient (S_{11}) and power gain (S_{21}). (After Guan, Jin, and Nguyen [10]. © The Institution of Engineering and Technology.)

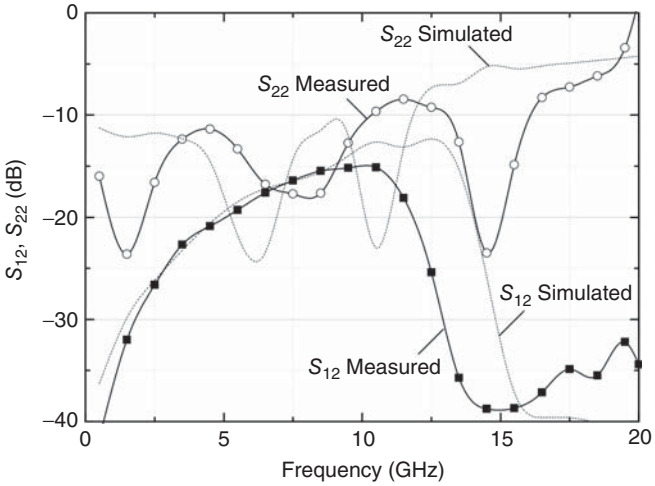


Figure 11.47. Simulated and measured output reflection coefficient (S_{22}) and reverse isolation (S_{12}).

The distributed amplifier was fabricated in a 0.25-μm CMOS process. To take advantage of the inherent isolation between PGS inductors produced by the PGS, adjacent PGS inductors are placed as close as 30 μm, which is the same as the width of the trace in the inductors. Figures 11.46 and 11.47 show the measured and calculated *S*-parameters. The amplifier exhibits a measured gain of around 7 dB with little fluctuation, input and output return losses higher than 10 dB, and reverse isolation below 15 dB from DC-11 GHz. These results agree reasonably well with those calculated, except the bandwidth of the measured gain (S_{21}) is 2 GHz lower than that of the simulated one. This discrepancy may be caused by unwanted parasitic capacitances in the layout, which were not taken into account in the simulation. The simulated and measured S_{21} have a small peak before the cut-off frequency, which is due to a pair of poles formed by some inductors and capacitors in the circuit located just before the cut-off frequency of the synthetic transmission lines. Figure 11.48 shows the noise figure of this distributed amplifier. It has a measured noise figure of 4.1–6.1 dB from 0.5 to 11 GHz. The measured power consumption is 78 mW using a voltage supply of 2 V. The entire passive structures on the input and output lines, including three octagonal PGS inductors and two rectangular non-PGS inductors, were modeled as two six-port structures and calculated using the electromagnetic (EM)-simulator IE3D. The *S*-parameters were then imported into the Agilent Advanced Design System (ADS) and simulated along with other circuit elements for the entire amplifier.

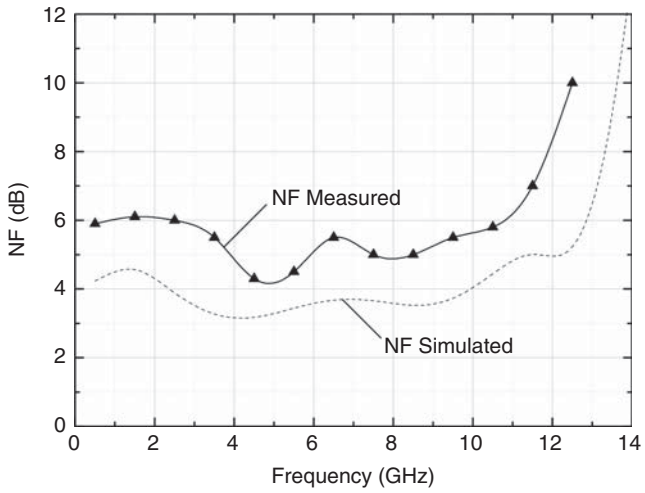


Figure 11.48. Simulated and measured noise figure (NF). (After Guan, Jin, and Nguyen [10]. © The Institution of Engineering and Technology.)

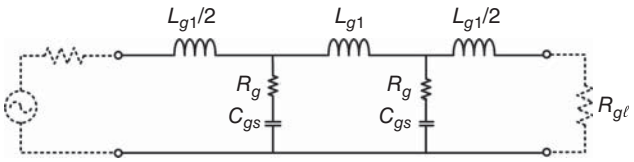


Figure 11.49. Schematic of the gate transmission line.

11.6.2.3 Distributed Amplifier Design Using Transmission Lines

General Design Approach

Besides inductors, transmission lines are commonly used in distributed amplifiers. Using transmission lines may alleviate the low Q of on-chip inductors but may also results in larger chip size. The use of transmission lines can be viewed under two different but related design principles. One is using the transmission line to absorb the (parasitic) gate–source (C_{gs}) or drain–source (C_{ds}) capacitances of MOSFETs into the transmission line, effectively turning the distributed amplifier's gate or drain network into a physical transmission line. Another is to consider the transmission line itself as an inductor and combine it with C_{gs} or C_{ds} to create a synthetic transmission line at the gate or drain, respectively. In the first consideration, C_{gs} or C_{ds} becomes part of the capacitance making up the transmission line. On the other hand, in the second category, the capacitance elements of the transmission line must be incorporated into the device's capacitance C_{gs} or C_{ds}.

To simplify the design illustration, we use two identical common-source MOSFETs as gain cells and consider the schematics of the gate and drain transmission lines separately. For the convenience of illustration, without loss of generality, we also assume that the length of each of the actual transmission lines used in the amplifier is less than a quarter-wavelength. Figure 11.49 shows the lumped-element schematic of the gate transmission line and Figure 11.50 shows the gate transmission line consisting of real transmission lines and parasitic capacitances C_{gs} [Figure 11.50(a)] and its transformation into equivalent lumped-element networks [Figure 11.50(b)–(d)]. In Figure 11.50(b), each of the transmission lines is replaced with its π equivalent circuit, assuming the transmission lines are lossless and less than a quarter-wavelength long. Following the indicated transformation and using the π equivalent-circuit model for a lossless transmission line whose length is less than a quarter-wavelength long given in Problem 4.33, we can write

$$C_g = C_{gs} + C'_{tg} + C_{tg} \tag{11.255}$$

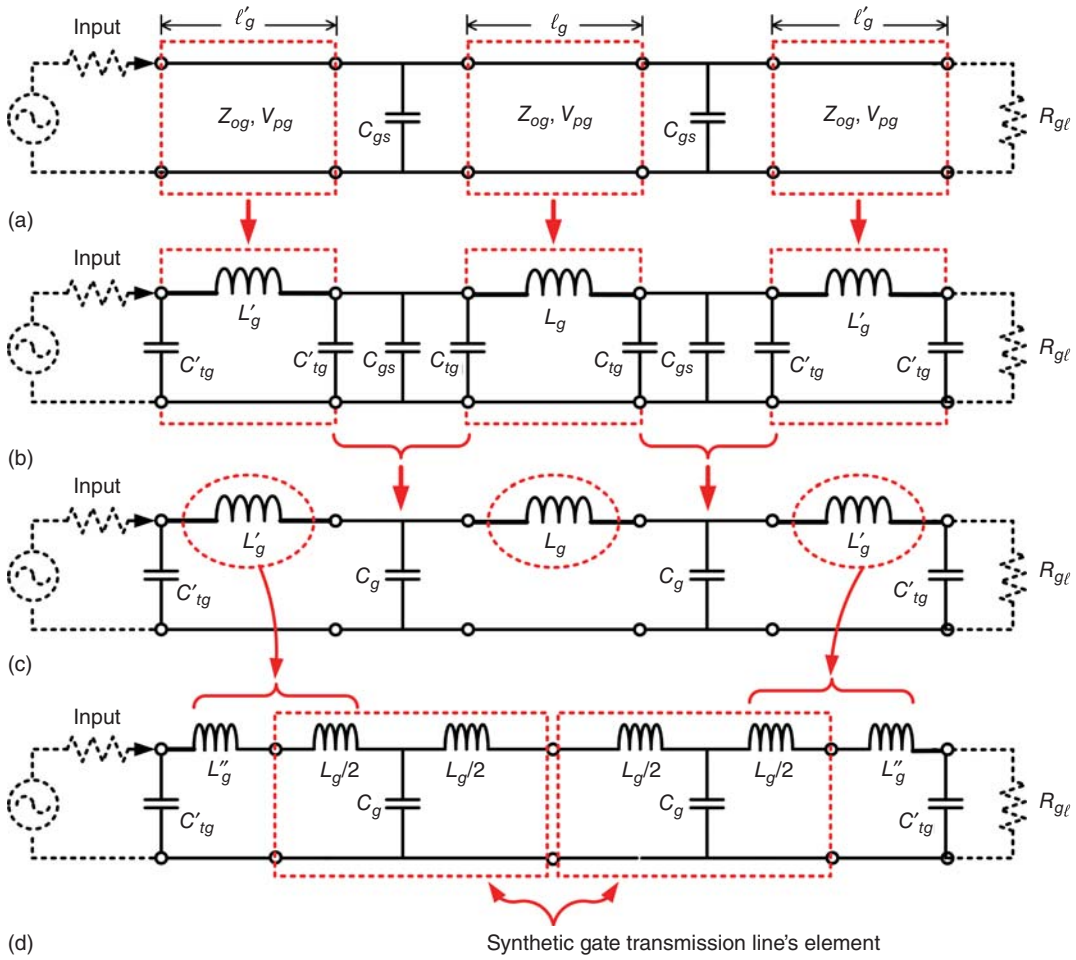


Figure 11.50. Gate transmission line formed by C_{gs} and actual transmission lines with characteristic impedance Z_{og}, phase velocity v_{pg} and lengths ℓ_d and ℓ'_d (a) and its equivalent lumped-element circuits (b)–(d). The gate resistance r_g is neglected for simplicity.

$$C_{tg} = \frac{1}{\omega_{cg} Z_{og}} \tan\left(\frac{\omega_{cg} \ell_g}{2 v_{pg}}\right) \tag{11.256}$$

$$C'_{tg} = \frac{1}{\omega_{cg} Z_{og}} \tan\left(\frac{\omega_{cg} \ell'_g}{2 v_{pg}}\right) \tag{11.257}$$

$$L_g = \frac{Z_{og}}{\omega_{cg}} \sin\left(\frac{\omega_{cg} \ell_g}{v_{pg}}\right) \tag{11.258}$$

$$L'_g = \frac{Z_{og}}{\omega_{cg}} \sin\left(\frac{\omega_{cg} \ell'_g}{v_{pg}}\right) \tag{11.259}$$

$$L'_g = L''_g + \frac{L_g}{2} \tag{11.260}$$

and the characteristic impedance, phase velocity and cut-off frequency of the resultant (synthetic) gate transmission line, as shown in Figure 11.50(d), can be obtained as

$$Z_g = \sqrt{\frac{L_g}{C_g}} \tag{11.261}$$

$$v_g = \sqrt{L_g C_g} \tag{11.262}$$

and

$$\omega_{cg} = \frac{2}{\sqrt{L_g C_g}} \tag{11.263}$$

respectively.

Similarly, the (synthetic) drain transmission line can be formed as shown in Figure 11.51 with characteristic impedance, phase velocity and cut-off frequency of

$$Z_d = \sqrt{\frac{L_d}{C_d}} \tag{11.264}$$

$$v_d = \sqrt{L_d C_d} \tag{11.265}$$

and

$$\omega_{cd} = \frac{2}{\sqrt{L_d C_d}} \tag{11.266}$$

respectively, where

$$C_d = C_{ds} + C'_{td} + C_{td} \tag{11.267}$$

$$C_{td} = \frac{1}{\omega_{cd} Z_{od}} \tan\left(\frac{\omega_{cd} \ell_d}{2 v_{pd}}\right) \tag{11.268}$$

$$C'_{td} = \frac{1}{\omega_{cd} Z_{od}} \tan\left(\frac{\omega_{cd} \ell'_d}{2 v_{pd}}\right) \tag{11.269}$$

$$L_d = \frac{Z_{od}}{\omega_{cd}} \sin\left(\frac{\omega_{cd} \ell_d}{v_{pd}}\right) \tag{11.270}$$

$$L'_d = \frac{Z_{od}}{\omega_{cd}} \sin\left(\frac{\omega_{cd} \ell'_d}{v_{pd}}\right) \tag{11.271}$$

$$L'_d = L''_d + \frac{L_d}{2} \tag{11.272}$$

Although the foregoing design illustration is valid, the derived equations (11.255)–(11.272) are only suitable when the actual transmission lines in each section are less than a quarter-wavelength. In practice, the lengths of the actual transmission lines may be longer than a quarter-wavelength. The design equations for any transmission-line length, however, are simple, following the standard equations for the characteristic impedance and phase velocity of lossless transmission lines. Let L_{go}, L_{do}, and C_{go}, C_{do} be the inductances and capacitances per unit length of the actual transmission lines along the gates and drains, respectively, and

ℓ_g and ℓ_d be the corresponding lengths of these transmission lines in each section. Assuming the capacitances C_{gs} and C_{ds} are distributed equally along the respective transmission lines in each section, the resultant per-unit-length capacitances of the (synthetic) gate and drain transmission lines can be approximately expressed as

$$C_{gt} \simeq C_{go} + \frac{C_{gs}}{\ell_g} \tag{11.273}$$

$$C_{dt} \simeq C_{do} + \frac{C_{ds}}{\ell_d} \tag{11.274}$$

The characteristic impedances, propagation constants (per unit length), phase velocities and cut-off frequencies of the gate and drain transmission lines, assuming lossless, can then be obtained approximately as

$$Z_g \simeq \sqrt{\frac{L_{go}}{C_{go} + C_{gs}/\ell_g}} \tag{11.275}$$

$$\gamma_g = j\beta_g \simeq j\omega\sqrt{L_{go}\left(C_{go} + \frac{C_{gs}}{\ell_g}\right)} \tag{11.276}$$

$$v_g \simeq \sqrt{L_{go}C_{gt}} \tag{11.277}$$

$$\omega_{cg} = \frac{2}{\sqrt{L_{go}C_{gt}}} \tag{11.278}$$

and

$$Z_d \simeq \sqrt{\frac{L_{do}}{C_{do} + C_{ds}/\ell_d}} \tag{11.279}$$

$$\gamma_d = j\beta_d \simeq j\omega\sqrt{L_{do}\left(C_{do} + \frac{C_{ds}}{\ell_d}\right)} \tag{11.280}$$

$$v_d \simeq \sqrt{L_{do}C_{dt}} \tag{11.281}$$

$$\omega_{cd} = \frac{2}{\sqrt{L_{do}C_{dt}}} \tag{11.282}$$

respectively. Equations (11.273)–(11.282) can be used for lossless inductor-based gate and drain transmission lines with L_{go} and L_{do} approximately considered the on-chip inductances in each section along the gate and drain lines, respectively, C_{go} and C_{do} neglected, and ℓ_g and ℓ_d equal to 1.

It should be noted that the characteristic impedance $Z_{g(d)}$, propagation constant $\gamma_{g(d)}$, phase velocity $v_{g(d)}$ and cut-off frequency $\omega_{cg(d)}$ of the (synthetic) gate (drain) transmission line are what concerned in the design – not the characteristic impedance $Z_{og(d)}$ and phase velocity $v_{pg(d)}$ of the actual transmission line used in Figure 11.50 or 11.51 itself. Typically, it is desired to have the same characteristic impedance, phase velocity and cut-off frequency for the gate and drain transmission lines. Particularly, Z_g and Z_d are normally chosen as 50 Ω to attain matching with other external 50-Ω components connecting to the distributed amplifier. Since $C_{ds} < C_{gs}$ for typical MOSFETs, in order to maintain the same characteristic impedance for both the gate and drain transmission lines, Z_{og}, Z_{od}, ℓ_g, and ℓ_d of the corresponding actual transmission

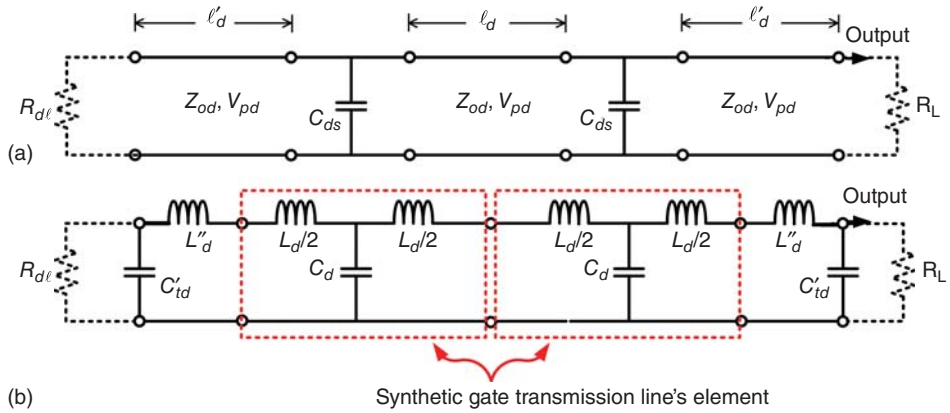


Figure 11.51. Drain transmission line formed by C_{ds} and actual transmission lines with characteristic impedance Z_{od}, phase velocity v_{pd} and lengths ℓ_d and ℓ'_d (a) and its equivalent lumped-element circuit (b). The output resistance R_{ds} is neglected for simplicity.

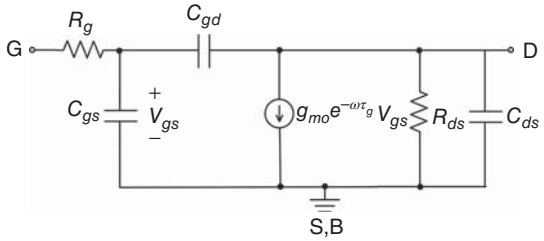


Figure 11.52. A simplified model of bilateral MOSFETs. The source (S) and bulk (B) terminals are tied together.

lines must be chosen properly – for instance, using a shorter transmission line at the gate than that at the drain to achieve equal (total) per-unit-length gate capacitance C_{gt} and drain capacitance C_{dt}. Furthermore, the cut-off equations (11.263), (11.266), (11.278), and (11.282) are useful for trade-off study between gain and bandwidth of distributed amplifiers considering fixed characteristic impedance as mentioned earlier for (11.254). Additionally, it is noted that, since C_{gs} and C_{ds} are function of bias voltages, the phase velocities and phase constants (and hence total phases) of the gate and drain transmission lines can be changed through bias voltages.

The foregoing design procedure assumes a very simple unilateral MOSFET model mainly for illustration purpose. More accurate MOSFET models should be used for the final analysis and optimization of distributed amplifiers. As an example, Figure 11.53 shows the gate and drain transmission lines using a simple MOSFET bilateral model shown in Figure 11.52. Note that C_{gd} is still omitted in Figure 11.53, assuming the MOSFET functioning close to unilateral, to facilitate the formation of the gate and drain transmission lines. This capacitance, however, should be included in the final analysis of the entire amplifier. The added drain–source resistances (R_{ds}) cause frequency-dependent loss along the drain transmission line in addition to that inherently incurred on the actual lossy transmission line.

Assuming R_g, C_{gt}, C_{gs} and R_{ds}, C_{ds} of the MOSFET are uniformly distributed across the length of the actual gate (ℓ_g) and drain (ℓ_d) transmission line, respectively, we can draw the equivalent circuits of an infinitely short section ($\Delta\ell \ll \lambda$) of the gate and drain transmission lines together with the MOSFET's equivalent circuit as shown in Figure 11.54. The impedance per unit length of the gate transmission line is obtained as

$$Z_{gu} = R_{gt} + j\omega L_{gt} \tag{11.283}$$

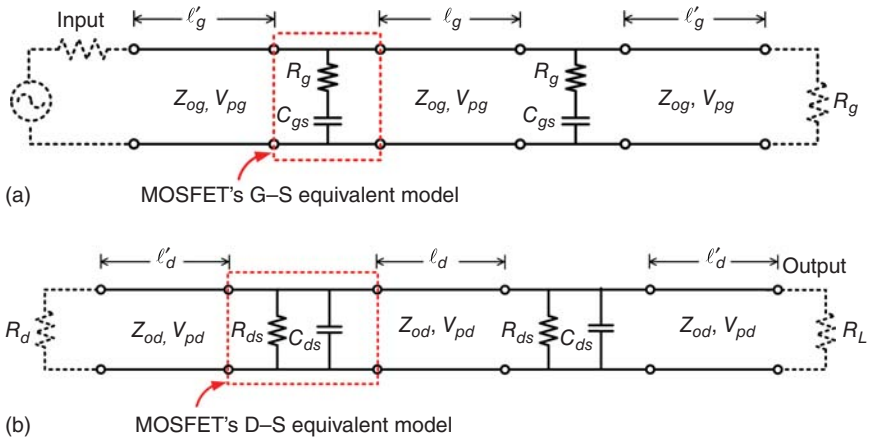


Figure 11.53. Gate (a) and drain (b) transmission lines using the simple bilateral MOSFET model of Figure 11.52 with g_m omitted. The transmission lines shown are actual transmission lines used in the amplifier.

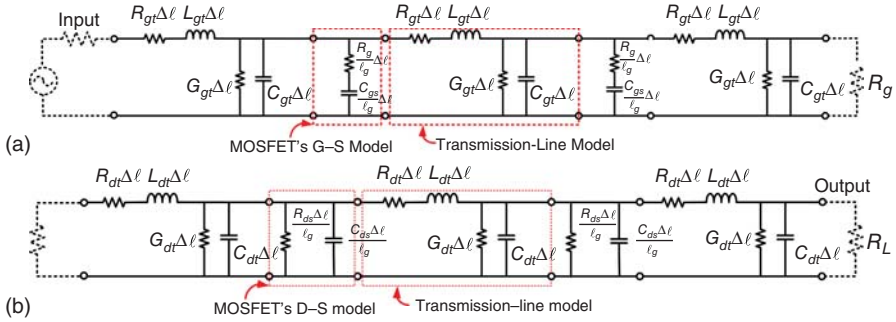


Figure 11.54. Equivalent circuits of infinitely short sections ($\Delta\ell \ll \lambda$) of (synthetic) gate (a) and drain (b) transmission lines with the MOSFET's gate resistance R_g and (output) drain–source resistance R_{ds} included. The actual gate and drain transmission lines consist of many of these corresponding sections connected in cascade. For good MOSFETs, R_g is typically small and R_{ds} is typically in the order of several hundred Ohms.

The admittance per unit length of the gate transmission line can be derived as

$$Y_{gu} = G_{gt} + j\omega C_{gt} + \frac{1}{\frac{R_g}{\ell_g} - \frac{j}{\omega C_{gs}/\ell_g}} \tag{11.284}$$

or

$$Y_{gu} = G_{gt} + \frac{R_g \ell_g \omega^2 C_{gs}^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} + j\omega \left(C_{gt} + \frac{C_{gs} \ell_g^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} \right) \tag{11.285}$$

after some manipulations. Similarly, the impedance per unit length of the drain transmission line as shown in Figure 11.54 is obtained as

$$Z_{du} = R_{dt} + j\omega L_{dt} \tag{11.286}$$

while the admittance per unit length of the drain transmission line is derived as

$$Y_{du} = G_{dt} + \frac{\ell_d}{R_{ds}} + j\omega \left(C_{dt} + \frac{C_{ds}}{\ell_d} \right) \tag{11.287}$$

The characteristic impedance and propagation constant (per unit length) of the gate and drain lines can now be obtained as

$$Z_g = \sqrt{\frac{Z_{gu}}{Y_{gu}}} = \sqrt{\frac{R_{gt} + j\omega L_{gt}}{G_{gt} + \frac{R_g \ell_g \omega^2 C_{gs}^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} + j\omega \left(C_{gt} + \frac{C_{gs} \ell_g^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} \right)}} \tag{11.288}$$

$$\gamma_g = \alpha_g + j\beta_g = \sqrt{Z_{gu} Y_{gu}} = \sqrt{(R_{gt} + j\omega L_{gt}) \left[G_{gt} + \frac{R_g \ell_g \omega^2 C_{gs}^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} + j\omega \left(C_{gt} + \frac{C_{gs} \ell_g^2}{\ell_g^2 + R_g^2 \omega^2 C_{gs}^2} \right) \right]} \tag{11.289}$$

$$Z_d = \sqrt{\frac{Z_{du}}{Y_{du}}} = \sqrt{\frac{R_{dt} + j\omega L_{dt}}{G_{dt} + \frac{\ell_d}{R_{ds}} + j\omega \left(C_{dt} + \frac{C_{ds}}{\ell_d} \right)}} \tag{11.290}$$

$$\gamma_d = \alpha_d + j\beta_d = \sqrt{Z_{du} Y_{du}} = \sqrt{(R_{dt} + j\omega L_{dt}) \left[G_{dt} + \frac{\ell_d}{R_{ds}} + j\omega \left(C_{dt} + \frac{C_{ds}}{\ell_d} \right) \right]} \tag{11.291}$$

Assume the actual transmission line used in the gate line is low loss and R_g of the MOSFETs is small such that the following conditions are met:

$$\begin{aligned} R_{gt} &\ll \omega L_{gt} \\ G_{gt} &\ll \omega C_{gt} \\ \overline{R}_g &\ll \frac{1}{\omega \overline{C}_{gs}} \end{aligned} \tag{11.292}$$

where $\overline{R}_g \equiv R_g/\ell_g$ and $\overline{C}_{gs} \equiv C_{gs}/\ell_g$ represent the gate resistance and gate–source capacitance per unit length of the gate line, respectively, we can approximate the gate line's characteristic impedance and propagation constant using binomial series as

$$Z_g \simeq \sqrt{\frac{L_{gt}}{C_{gt} + C_{gs}}} \tag{11.293}$$

and

$$\gamma_g = j\beta_g \simeq j\omega \sqrt{L_{gt}(C_{gt} + C_{gs})} \tag{11.294}$$

respectively, which are similar to (11.275) and (11.276). Similarly, when the actual transmission line used in the drain line has small loss and R_{ds} of the MOSFETs is large such that

$$\begin{aligned} R_{dt} &\ll \omega L_{dt} \\ G_{dt} &\ll \omega C_{dt} \\ \overline{R}_{ds} &\gg \frac{1}{\omega \overline{C}_{ds}} \end{aligned} \tag{11.295}$$

where $\overline{R}_{ds} \equiv R_{ds}/\ell_d$ and $\overline{C}_{ds} \equiv C_{ds}/\ell_d$ represent the drain–source resistance and drain–source capacitance per unit length of the drain line, respectively, we can derive approximate expressions for the drain line's characteristic impedance and propagation constant using binomial series as

$$Z_d \simeq \sqrt{\frac{L_{dt}}{C_{dt} + \frac{C_{ds}}{\ell_d}}} \tag{11.296}$$

and

$$\gamma_d = j\beta_d \simeq j\omega\sqrt{L_{dt}\left(C_{dt} + \frac{C_{ds}}{l_d}\right)} \tag{11.297}$$

respectively. These equations are the same as (11.279) and (11.280).

For distributed amplifiers employing on-chip inductors, the inductors also have loss which results in complex characteristic impedances and propagation constants for the gate and drain transmission lines. Assume the lossy inductor can be modeled as a combination of series resistor $R_{gt(dt)}$ and inductor $L_{gt(dt)}$ and shunt conductor $G_{gt(dt)}$ and capacitor $C_{gt(dt)}$ as for the transmission lines in Figure 11.54 with $\Delta\ell = 1$, where $R_{gt(dt)}$ represents the loss due to metal strip, $L_{gt(dt)}$ is the inductance of the on-chip inductor, $G_{gt(dt)}$ represents the loss due to dielectric layers and silicon substrate, and $C_{gt(dt)}$ represents the parasitic capacitor caused by the dielectric layers and silicon substrate, the characteristic impedances and propagation constants of the gate and drain transmission lines can also be approximately determined using (11.293)–(11.294) and (11.296)–(11.297), provided that ℓ_g and ℓ_d are let to be equal to 1.

Design Example of a CMOS Distributed Amplifier based on Transmission Lines

Figure 11.55 shows the schematic and photograph of a distributed amplifier with six NMOS transistors on a 0.25-μm CMOS process [11]. The employed transmission lines are finite-ground coplanar waveguide (CPW). They are realized on the topmost metal layer to achieve low loss due its thickest metallization and farthest distance from the lossy silicon substrate. The CPW connecting the gates and that connecting the drains are periodically loaded with the NMOS gate–source and drain–source capacitances, respectively. Together, they form synthetic transmission lines at the gates (gate line) and drains (drain line) of the NMOS transistors. The gate and drain lines have characteristic impedance of 50 Ω and are terminated with 50-Ω resistors. The characteristic impedance of the CPW is chosen as 75 Ω to accommodate the gate–source and drain–source parasitic capacitances of the MOSFETs. Bias circuits are formed by on-chip inductors and capacitors. The on-chip capacitors C, as seen in Figure 11.55(a), block the DC from the supplied bias voltages to the 50-Ω terminating resistors. On-chip inductors L, acting as RF chokes, feed the DC voltage and current to the MOSFETs, while stopping the RF signals from going through it. The finite-ground CPW is used to allow easy connection of the sources of the NMOS transistors to the ground while maintaining a compact size for the amplifier. CPW also facilitates wider central strip for high characteristic impedance as compared to its microstrip line counterpart, thus lowering the conductor loss. Low loss is needed to maintain well-behave synthetic transmission lines, which are critical for realizing good distributed amplifiers as mentioned previously. The fabricated distributed amplifier has a measured gain around 6 dB and a return loss of more than 13 dB across 3.1–10.6 GHz. It is driven from a 3.3-V power supply and consumes 68 mW.

11.6.2.4 Enhanced Design for CMOS Distributed Amplifiers. The design of distributed amplifiers is faced with three basic problems: high power consumption, relatively low gain and large die size. These issues are concerned for commercial CMOS applications – particularly for portable wireless devices. In this section, we will address two specially designed distributed amplifiers: one lending to not only low power consumption but also high gain and another one resulting in miniature die. These distributed amplifiers not only address specific designs enabling usage of distributed amplifiers in RF systems, but also pave the way for other possible designs leading to low power consumption, high gain and miniaturization for CMOS distributed amplifiers.

Design of Low Power-Consumption and High Gain CMOS Distributed Amplifiers

A major drawback of distributed amplifiers is their large DC power consumption, severely limiting their usage in wireless portable devices. This is due to the fact that several parallel transistors, with each transistor draining current from the source, are needed to form the required artificial transmission lines and to achieve a reasonable gain on a (standard) 50-Ω load. Typical distributed amplifiers employ at least three or more gain stages. The most crucial element in distributed amplifiers is the gain cell, which dictates the performance of

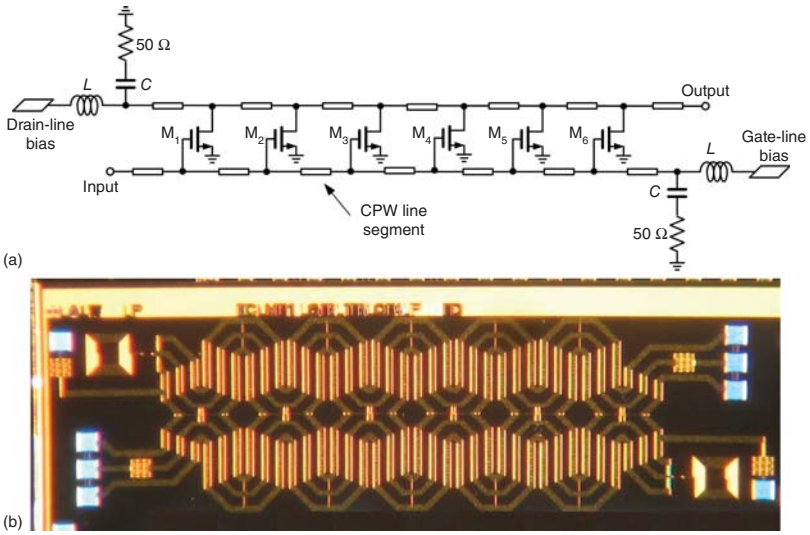


Figure 11.55. Schematic (a) and photograph (b) of a 0.25-μm CMOS distributed amplifier on CPW. The chip area is 0.8×2.6 mm^2.

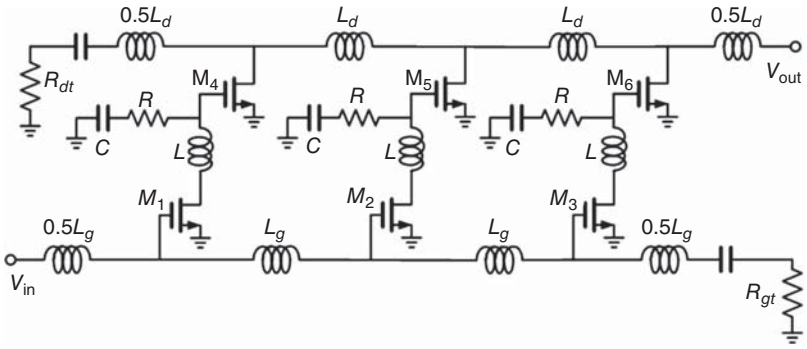


Figure 11.56. Simplified schematic of the new low power distributed amplifier.

the entire amplifiers such as gain, noise figure, linearity, power consumption. Therefore, proper design of gain cell including its topology to achieve certain design goal, such as optimum gain or noise figure, should be used. A CMOS distributed amplifier employing particular gain cells that produce very low power consumption is presented in this section. The gain cell and hence amplifier also provides substantially enhanced transconductance and hence gain over ultrawide bandwidths. While most of the design approaches for distributed amplifiers seek to improve only the gain cell used in the amplifiers through its transconductance, considering both the gain cell's transconductance and power consumption provides more optimum design results, particularly for portable wireless devices.

Circuit Analysis. A major challenge in designing a low power-consumption distributed amplifier is the trade-off between power consumption and gain. In order to lower the power consumption, each transistor has to be biased at a very low overdrive voltage. This, however, leads to insufficient gain for the whole amplifier. Figure 11.56 shows a distributed amplifier topology that has low power consumption with decent gain. This amplifier employs multiple gain cells, each consisting of two cascade common-source transistors with peaking inductor instead of a single common-source transistor used in conventional distributed amplifiers. As mentioned earlier, gain-cell configurations affect substantially the performance, particularly gain, of distributed amplifiers. Particularly, the employed gain-cell configuration improves the gain significantly with similar power consumption or similar gain with substantially reduced power consumption, over an extremely wideband, as compared to the conventional distributed amplifiers.

As previously discussed, the gain of a distributed amplifier can be increased by simply increasing the number of stages until it reaches a maximum number, as given in (11.247). This increased gain is obtained at the cost of a larger die area and higher power consumption, which are not very desirable, particular for commercial CMOS-based portable devices. Apparently, the transconductance of each gain cell is the most important issue for the gain as can be seen in (11.244). It should be noted that the intrinsic gain of each gain cell, defined as $g_m R_{\text{out}}$ where g_m and R_{out} are the transconductance and output resistance of the gain cell, respectively, is no longer important, because each gain cell in the distributed amplifier is driving a small resistive load, which is typically 50-Ω. Furthermore, these gain cells cannot take advantages of their large output impedance for gain enhancement because of the small resistive load at output. The capability to provide sufficiently large current at the output of the gain cells will determine the gain of the distributed amplifier.

Figure 11.57(a)–(c) show several possible gain-cell configurations used in CMOS distributed amplifiers [12]. Figure 11.57(a) is a common-source stage, which is considered the most basic and conventional gain cell and has been widely used in distributed amplifiers. It provides a decent gain and very large bandwidth. Figure 11.57(b) is a cascode structure used to enhance reverse isolation. This structure does not provide significantly higher g_m than a single common-source transistor and thus does not have considerable gain advantage over that in Figure 11.57(a). In Figure 11.57(c), a current reuse gain cell is formed by NMOS and PMOS transistors. This configuration is supposed to have higher transconductance at proper bias voltages and transistor dimensions. However, because of the mobility difference of NMOS and PMOS devices, the bias and dimensions for this structure are very difficult to be determined. Figure 11.57(d) shows a cascade common-source gain cell. Two common-source transistors are connected with each other through a peaking inductor. A resistor is connected to the gate of the second transistor. Either passive or active inductors can be used. Active inductors, however, suffer possible variation of inductance and increase in noise over extremely wide bandwidths. This particular structure enhances the transconductance significantly and hence the amplifier's gain.

Figure 11.58(a) shows a small-signal equivalent circuit of the cascade common-source gain cell in Figure 11.57(d). C_{gsi}, C_{dbi}, and C_{gdi} $(i=1, 2)$ are the respective gate-to-source, drain-to-bulk, and gate-to-drain capacitances of the NMOS transistor M_1 $(i=1)$ and M_2 $(i=2)$ in each gain cell. g_{m1}, g_{m2} and r_{O1}, r_{O2} are the transconductances and output resistances of M_1 and M_2, respectively. R and L are the resistance and inductance used in each gain cell, as shown in Figure 11.57(d), respectively. Typically, r_{O1} and r_{O2} are relatively large and therefore can be neglected. Neglecting r_{O1} and r_{O2} and combining C_{gd1}, C_{gd2} with C_{gs1}, C_{gs2} and C_{db1}, C_{db2} to form C'_{gs} and C'_{db}, which are the Miller equivalent capacitance observed at the gate and drain of the first transistor, the small-signal model can be simplified as shown in Figure 11.58(b). Since the gain of distributed amplifiers depends only on the transconductance of each gain cell, this transconductance is needed for the analysis of the amplifier's gain and is thus determined, using the small signal model in Figure 11.58 as follows.

The current I_{out} in Figure 11.58(c) can be expressed as

$$I_{\text{out}} = g_{m2} V_{gs2} \tag{11.298}$$

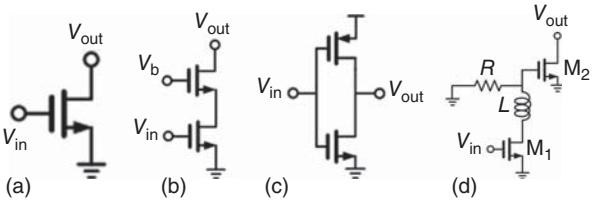


Figure 11.57. Possible gain cell configurations used in CMOS distributed amplifiers. (a) Traditional common-source gain cell. (b) Cascode gain cell. (c) Current-reuse gain cell. (d) Cascade common-source gain cell. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

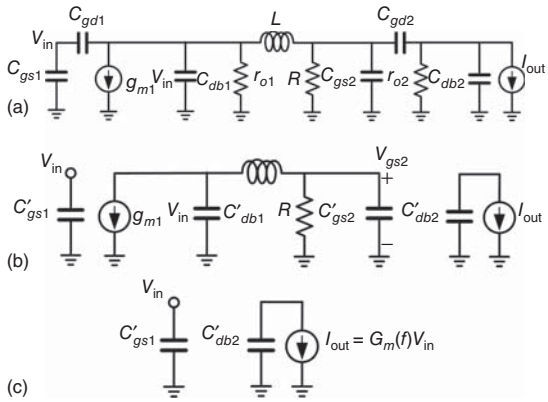


Figure 11.58. (a) Small-signal model of the cascade gain cell. (b) Simplified small-signal model. (c) Equivalent-circuit model for the gain and bandwidth analysis. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

where V_{gs2}, as seen in Figure11.58(b), is the gate–source voltage of M_2 and can be obtained from Figure 11.58(b) as

$$V_{gs2} = I_L\left(R//\frac{1}{sC'_{gs2}}\right) = g_{m1}V_{gs1}\frac{\frac{1}{sC'_{db1}}}{\frac{1}{sC'_{db1}} + sL + R//\frac{1}{sC'_{gs2}}} \cdot \left(R//\frac{1}{sC'_{gs2}}\right) \tag{11.299}$$

Substituting (11.299) into (11.297) gives

$$I_{\text{out}} = g_{m2}V_{gs2} = g_{m1}g_{m2}V_{gs1}\frac{\frac{1}{sC'_{db2}}}{\frac{1}{sC'_{db2}} + sL + R//\frac{1}{sC'_{gs2}}} \cdot \left(R//\frac{1}{sC'_{gs2}}\right) \tag{11.300}$$

Utilizing

$$V_{gs1} = V_{\text{in}} \tag{11.301}$$

we can then derive the transconductance from (11.300) as

$$G_m = \frac{\partial I_{\text{out}}}{\partial V_{\text{in}}} = g_{m1}g_{m2}\frac{R//\frac{1}{sC'_{gs2}}}{\frac{1}{sC'_{db1}} + sL + R//\frac{1}{sC'_{gs2}}} \cdot \frac{1}{sC'_{db1}} \tag{11.302}$$

which becomes

$$\begin{aligned} G_m &= g_{m1}g_{m2}\frac{\frac{R\cdot\frac{1}{sC'_{gs2}}}{R+\frac{1}{sC'_{gs2}}}}{1 + s^2LsC'_{db1} + \left(\frac{R\cdot\frac{1}{sC'_{gs2}}}{R+\frac{1}{sC'_{gs2}}}\right)\cdot sC'_{db1}} \\ &= g_{m1}g_{m2}\frac{R}{\left(1 + s^2LC'_{db1} + \frac{R}{sC'_{gs2}R+1}\cdot sC'_{db1}\right)\cdot(ssC'_{gs2}R + 1)} \\ &= g_{m1}g_{m2}\frac{R}{LC'_{db1}sC'_{gs2}Rs^3 + LC'_{db1}s^2 + RC'_{db1}s + RsC'_{gs2}s + 1} \end{aligned} \tag{11.303}$$

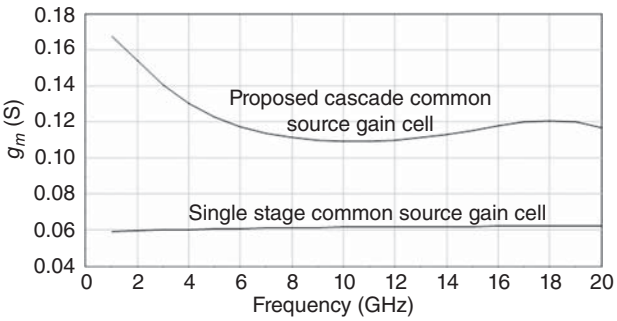


Figure 11.59. Calculated transconductances of the cascade common-source gain cell and a single common-source gain cell. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

Usually, $C'_{db1} \ll C'_{gs2}$, and so (11.303) can be further simplified as

$$\begin{aligned} G_m &\approx g_{m1}g_{m2}\frac{R}{LC'_{db1}C'_{gs2}Rs^3 + LC'_{db1}s^2 + RC'_{gs2}s + 1} \\ &= g_{m1}g_{m2}\frac{R}{(LC'_{db1}s^2 + 1)(RC'_{gs2}s + 1)} \end{aligned} \tag{11.304}$$

As can been from Eq. (11.304), there exists three poles in the gain response. One pole is formed by the input capacitor of the upper transistor M_2. This pole normally dominates the low frequency response due to usually large gate-to-source capacitance of MOSFETs. The other two complex conjugate poles are created by some inductance and output capacitance of the lower transistor M_1. In practice, these two poles are located on the left *s*-plane instead of exactly on the complex axis because of the loss incurred in real circuits. The presence of these two poles will boost up the transconductance of the gain cell at frequency

$$f_c = \frac{1}{\sqrt{LC'_{db1}}} \tag{11.305}$$

which can also be considered the cut-off frequency of the gain cell's transconductance.

By properly choosing the inductance and resistance in the cascade common-source gain cell, a nearly constant transconductance can be obtained between the first-pole frequency and the cut-off frequency f_c in (11.305). Unlike a single-transistor gain cell, such as that in Figure 11.57(a), whose transconductance remains fairly constant over a particular frequency range, the cascade common-source gain cell exhibits more frequency variation for its transconductance. Considering this and the fact that the gain of distributed amplifiers is proportional to the transconductance of each gain cell, the cut-off frequency of a distributed amplifier employing the cascade common-source gain cell is determined by the cut-off frequency of each gain cell's transconductance, instead of that of the synthetic transmission lines at the input and output of the amplifier, provided that the former is lower than the latter. Figure 11.59 shows the calculated transconductance of the cascade common-source gain cell and that of a single common-source transistor.

Comparing with the single common-source transistor, the cascade common-source gain cell provides a significantly higher transconductance, which is expected to lead to a significantly higher gain. Exploiting this resultant unique capability of providing high transconductance, the cascade common-source gain cells can be biased at a very low voltage to consume a very small amount of power, while still providing enough transconductance to obtain a decent gain for the entire amplifier.

The distributed amplifier topology employing cascade common-source gain cells also facilitates the enhancement of stability. The cascade structure used in each gain cell stage results in higher reverse isolation as compared to a single-transistor cell, while maintaining good input and output matching, which may help improve the stability of the distributed amplifier. The inductors used in the gain cells may resonate with the input gate capacitors of some transistors, leading to possible oscillation if negative resistance is

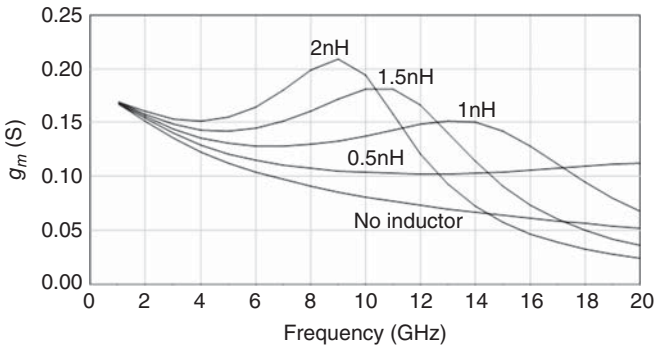


Figure 11.60. Calculated frequency response of the transconductance of the proposed gain cell for different values of inductance. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

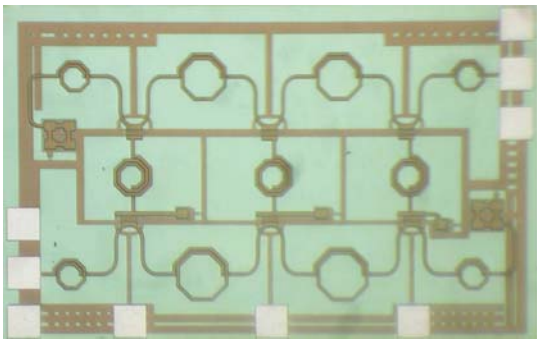


Figure 11.61. Microphotograph of the distributed amplifier employing cascade common-source gain cells. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

generated. However, since all of the transistors are grounded at the source, as long as the generated negative resistance is lower than the resistance of the resistor used in each gain cell, the circuit will remain stable. For instance, for the transistors used in the designed distributed amplifier, the induced negative resistance is lower than 80 Ω, and so the circuit is stable when an 80-Ω resistor is used in each gain cell. Simulated results of the designed distributed amplifier show that its stability factor K is greater than 1 across DC to 18 GHz, demonstrating its expected unconditional stability.

Design Example. A distributed amplifier with cascade common-source gain cells is designed and fabricated using a 0.18-μm CMOS process. The gate width of each transistor is optimized for gain and power consumption. The inductance L in each gain cell determines the bandwidth of the distributed amplifier. The calculated frequency response with different values of inductance is shown in Figure 11.60. One nanohenry is chosen for the inductance to obtain a reasonably flat gain over the desired frequency range of 3.1–10.6 GHz.

On-chip inductors are used for the inductive elements of the synthetic transmission lines in the distributed amplifier. Very small inductance values are used in order to obtain high Q resulting from less eddy current loss in the silicon substrate. The full-wave EM simulator IE3D is used for simulations of all passive components including bends, interconnects, and spiral inductors. The two synthetic transmission lines are terminated with 50-Ω resistors and AC-coupling capacitors to minimize reflections.

The designed distributed amplifier uses three stages to ensure a decent gain, reasonable die area and small power consumption. The amplifier has a chip size of 1.6×0.9 mm^2. Figure 11.61 shows its microphotograph.

The fabricated CMOS distributed amplifier was measured using on-wafer probes. Bias-tees and DC biasing probes were used to bias the circuit. Figure 11.62 shows the calculation and measurement results for the power gain and input return loss. The amplifier exhibits a measured gain of around 10 dB and input return lossless than 20 dB across 3.1–10.6 GHz, which agrees reasonable with those calculated. The power consumption was measured at 19.6 mW with the entire circuit biased at an extremely low voltage of 0.7 V. Figure 11.63

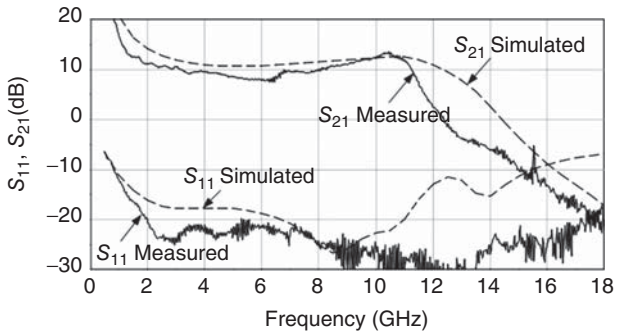


Figure 11.62. Measured and simulated power gain (S_{21}) and input return loss (S_{11}) in low power consumption mode (measured $P_{dc} = 19.6$ mW). (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

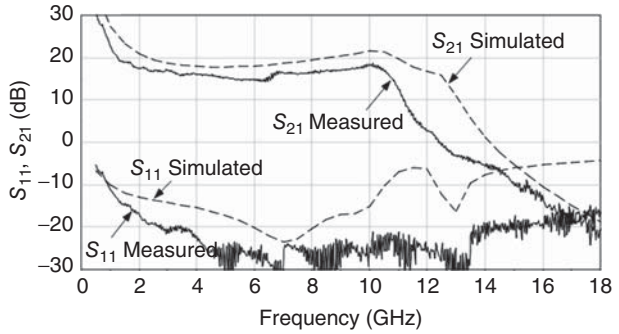


Figure 11.63. Measured and simulated power gain and input return loss in high gain mode (measured $P_{dc} = 100$ mW). (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

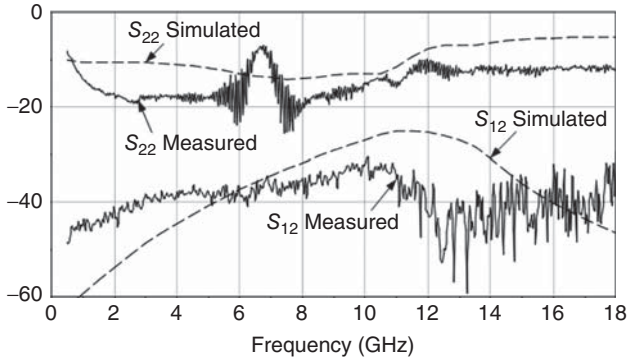


Figure 11.64. Measured and simulated output return loss (S_{22}) and isolation (S_{12}). (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

shows the calculated and measured gain and input return loss at a high gain mode. As can be seen, the distributed amplifier demonstrates a high gain of around 16 dB and return loss below 20 dB, measured across 3.1–10.6-GHz range, with a power consumption of 100 mW. The measured and calculated results also agree fairly well. A very high gain near-DC is observed from both Figures 11.62 and 11.63. This is due to the L, R, and C network that appears between the two transistors in each gain cell, as expected from the simulations. The measured gain shows a lower cut-off frequency than that of simulation. This may be attributed to the parasitics resulting from the layout that were not taken into account in the simulations.

Figure 11.64 shows the measured and simulated results for the isolation and return loss at the output. Both calculated and measured results show that there is only a very small difference between the high gain mode and low power mode. It is also noted that the use of the cascade gain cells results in high isolation of more than 30 dB for the distributed amplifier. The noise observed in the measured S_{11} and S_{22} data shown

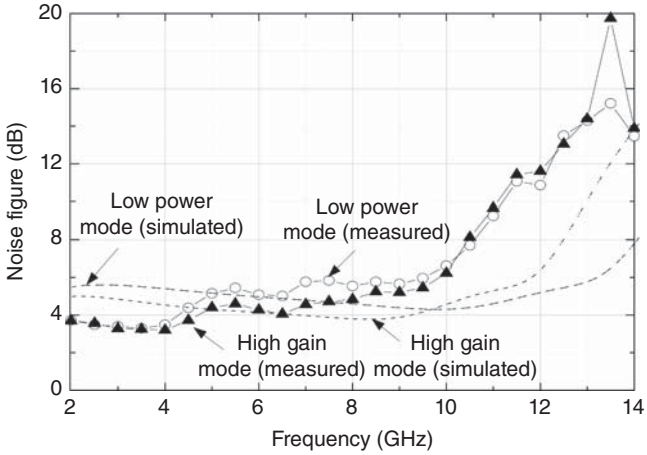


Figure 11.65. Measured and simulated noise figure in high gain and low power modes. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

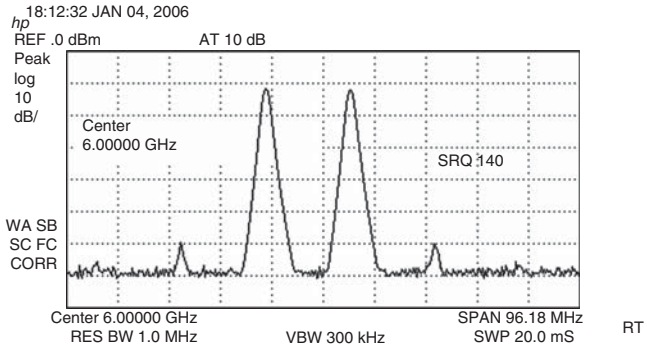


Figure 11.66. Output spectrum of the distributed amplifier with input tones at 5.99 and 6.01 GHz. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

in Figures 11.62–11.64 were due to calibration of the vector network analyzer. Nevertheless, multiple measurements for different chips have been performed and, despite this noise, the results are very consistent. The abnormal peak in the measured S_{22} near 6.5 GHz seen in Figure 11.64 may be due to some unexpected parasitics at the output port of the fabricated amplifier.

The noise figure of the distributed amplifier at both high gain and low power modes has also been measured. The measured and simulated results are plotted in Figure 11.65. The noise figure in the low power mode is slightly higher than that in the high gain mode both in simulation and measurement. Since sufficient gain is provided by the lower transistor in each gain cell, the distributed amplifier with cascade gain cell configuration does not give higher noise figure than that of traditional distributed amplifiers.

The input third-order intercept point (IIP3) was also measured for the distributed amplifier. Two input tones with 10 MHz adjacent to each other were combined by a power combiner and input into the amplifier. In the output spectrum, two fundamental tones and two third-order intermodulation (IM3) products can be observed. Figure 11.66 shows the output spectrum when IIP3 was measured at high gain mode and 6 GHz. Two fundamental tones are located at 5.99 and 6.01 GHz with a power of −12 dBm and two IM3 products are located at 5.97 and 6.03 GHz with a power of −61 dBm. The IIP3 of the distributed amplifier was measurement at both high gain and low power mode from 2 to 10 GHz. The results, plotted in Figure 11.67, show the IIP3 of 0–2 dBm for high gain mode and −9.5 to −3.5 dBm for low power mode.

Design of Miniature CMOS Distributed Amplifiers

As seen in Equations (11.242)–(11.244), the gain of distributed amplifiers depends primarily on the transconductance of the gain cells, the characteristic impedances and losses of the synthetic transmission

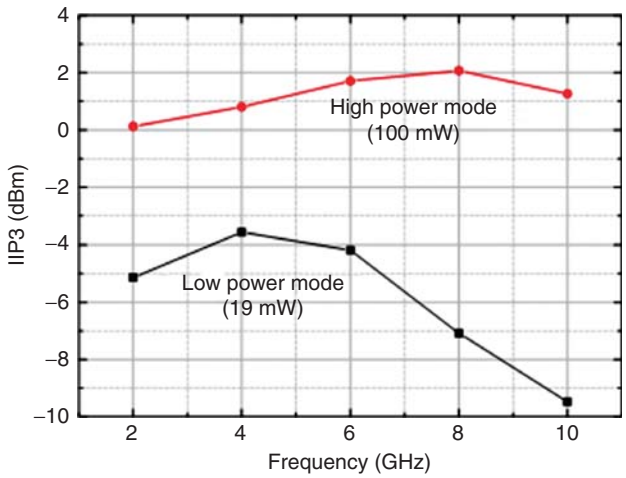


Figure 11.67. Measured IIP3 of the distributed amplifier in both high gain and low power modes. (After Guan and Nguyen [12]. Reprinted with permission of IEEE.)

lines (or in turn the on-chip inductors or actual transmission lines), and the number of stages. High characteristic impedance is desirable for gain enhancement and ease in absorption of the parasitic capacitances of the gain cells. However, large characteristic impedances are not typically feasible in CMOS technologies due to (unrealizable) extremely narrow line widths resulting from very thin Oxide layers acting as dielectrics for the transmission lines. For instance, the maximum characteristic impedance for microstrip lines in most six-metal CMOS technologies does not typically exceed 100 Ω. CPW structures allow greater characteristic impedances due to their more flexible geometry utilizing both gap and strip width; but they have larger size even with finite ground planes. It is expected that certain high characteristic impedances needed to meet particular design requirements may not be realized even with more advanced CMOS structures involving many metal layers. At the same time, relatively high losses associated with transmission lines implemented in CMOS also degrade the amplifier gain as seen from (11.243) and (11.244). The challenge of loss reduction could be mildly addressed by effectively shielding the transmission lines from the silicon substrate. However, the problem of loss in transmission lines is dependent upon CMOS technology and cannot be entirely controlled by the RFIC designers. The amplifier gain can always be improved to offset the use of lower-than-desired characteristic impedances and the transmission-line loss by increasing the number of stages up to a limited number. This, however, will result in larger size and more power consumption which are indeed undesirable, particularly for commercial wireless portable devices.

With regard to the size of the distributed amplifiers, the size of on-chip inductors or transmission lines is of significant concern to CMOS RFIC designers since it mainly determines the amplifier's size. This size, therefore, might render the distributed amplifiers practically useless for potential integration with other circuits in CMOS devices, as it increases the unit cost of implementation. Miniaturized versions of on-chip inductors or transmission lines (or equivalently synthetic transmission lines) can be explored to solve this problem, but unless they prove that the dimensions of the distributed amplifiers are compatible with the inductor-based analog broadband circuits such as feedback amplifiers, they would not be preferred by the RFIC designers, chip makers, or industry. That implies size reduction of 90% or more from the distributed amplifiers employing conventional transmission lines or inductors.

Typical distributed amplifiers employing conventional transmission lines or inductors are relatively large. For instance, the chip area for most distributed amplifiers that employ on-chip spiral inductors is still in the range of a few millimeters square, rendering the circuit too expensive for CMOS implementation. It is particularly recognized that the size reduction for distributed amplifiers is directly related to miniaturization of on-chip inductors or transmission lines. The following two sections describe two miniaturized CMOS distributed amplifier topologies that utilize miniature passive elements to obtain significant size compression.

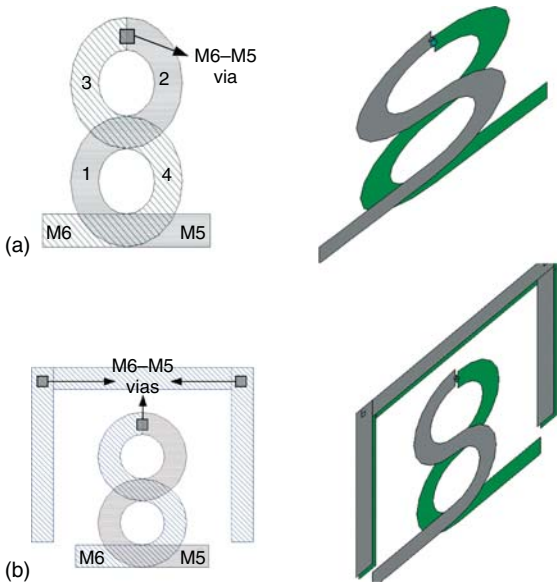


Figure 11.68. (a) Octagonal multilayer inductor and its 3D view and (b) pseudo-CPW with its 3D view. (After Chirala, Guan, and Nguyen [13]. © The Institution of Engineering and Technology.)

A Miniature Distributed Amplifier. This section describes a miniaturized distributed amplifier implemented using both on-chip inductors and transmission lines realized on multiple metal layers [13]. It is important to recognize that, while reducing the size of inductors and transmission lines, the loss of these structures should also be minimized as much as possible as it affects the gain – not only its magnitude but also its ripple, as can be seen from (11.244). Losses in CMOS based passive structures which include inductors and transmission lines are due to the conductivity of silicon substrate and oxide layers as well as the conductivity (or equivalently ohmic resistance) of the metals. The loss due to the silicon substrate can be removed by shielding it from the passive structures. An effective way that largely removes the (electrical) loss is using a solid or pattern ground shield as discussed in Section 3.4.5.

Multilayer Inductors and Transmission Lines. Figure 11.68 shows the layouts of the multilayer inductor and transmission line employed in the distributed amplifier. Both structures have a curvilinear "8" shaped pattern. The inductor, as shown in Figure 11.68(a), is implemented using a single conducting strip having two segments on two different metal layers (M5 and M6), connected through the via-hole. The transmission line is implemented using two conducting strips. One conducting strip (signal strip) consists of two segments on two different metal layers (M5 and M6) connected through via. Another conducting strip has two separate strips on M5 and M6 connected through vias. The transmission line is essentially a two-conductor transmission line and can be viewed as a nonuniform pseudo-CPW in which the characteristic impedance changes along the location of the signal line due to varying distance between it and the other strip. The primary advantage of this pseudo-CPW is that it can generate very high characteristic impedance due to relatively large capacitance per unit length resulting from large spacing between the conducting strips, while the loss is relatively low due to concentration of fields within the low loss oxide layers surrounding the strips rather than the more lossy silicon substrate. Table 11.1 shows the calculated parameters of the pseudo-CPW. As expected, the distance of 25 μm results in moderately large characteristic impedance, as is evident from Table 11.1. Even greater characteristic impedance could have been obtained if the distance was increased even further but that would enormously increase the size of the structure and possibly transform the transmission line into inductors.

For the inductor, the loss of the structure is also minimized possibly due to the strong negative coupling within the structure itself. The total inductance presented by the structure can be obtained approximately as

$$L_{\text{tot}} \simeq L_{\text{slf1}} + L_{\text{slf2}} + L_{\text{slf3}} + L_{\text{slf4}} - 2(m_{1,4} + m_{2,3}) \tag{11.306}$$

TABLE 11.1. Transmission Line Parameters of the Pseudo-CPW for Different Spacing between the Transmission Tine's Two Conducting Strips

Spacing (μm)	Frequency (GHz)	R_o (Ω)	λ (mm)	α (dB/mm)
10	1	60.6	49.1	0.15
	3	65.5	32.3	0.17
	6	74.2	25.4	0.18
	11	92.6	15.1	0.22
	20	99.2	10.8	0.25
16	1	81.1	31.4	0.19
	3	90.4	18.8	0.20
	6	102.3	10.4	0.20
	11	115.3	8.5	0.22
	20	124.8	4.4	0.24
25	1	110.6	25.7	0.20
	3	117.9	18.8	0.23
	6	130.3	11.3	0.23
	11	142.6	6.1	0.26
	20	179.2	3.4	–0.27

R_o is the real part of the characteristic impedance Z_o, λ is the wavelength, and α is the attenuation constant. The width of the signal strip is 12 μm and that of the other strip is 15 μm. The spacing is referenced to the narrowest distance between the two strips.

where $L_{\text{slf}x}$, $x = 1–4$, denotes the self inductance of each segment while $m_{i,j}$, $i, j = 1–4$ and $i \neq j$, denotes the mutual coupling between different segments. The presence of such strong negative coupling within the structure tends to lower the net inductance, which is acceptable and in fact mandated by the matching conditions for the particular distributed amplifier circuit discussed in this section.

Design and Performance. The miniature distributed amplifier was designed on a TSMC 0.18-μm CMOS process [14]. For the pseudo-CPW, both strips are implemented on M6 and M5 metal layers with the second strip spaced equidistant from each segment of the signal strip so that the overall symmetry of the structure is maintained. The design and optimization of the pseudo-CPW was performed using the EM simulator IE3D. Several iterations were performed to arrive at the best structure which yielded the required performance. It is worth mentioning here that, when the structure is configured as an inductor by removing its second strip as in Figure 11.68(a), the segment length has to be increased in order to obtain half of the inductance associated with the corresponding transmission line. The requirement for the inductors of 0.5-nH inductance was dictated by the input and output matching conditions for the designed distributed amplifier. When implemented as a transmission line as shown in Figure 11.68(b) with the widths of the signal and other strips being 12 and 15 μm, respectively, and the distance between these strips being kept at a moderately large distance of 25 μm, the characteristic impedance and loss of the structure are 180 Ω and 0.27 dB/mm at 20 GHz, respectively. The corresponding structure is 200×180 μm in size. When implemented as an inductor, the size was similar to the transmission line but the inductance was drastically reduced. The low loss and high characteristic impedance associated with the transmission-line structure enables using only three stages for the distributed amplifier to achieve a decent gain across an extremely wide bandwidth. Further, only a single transistor element is utilized as the gain cell. If a cascade common-source gain cell structure, as described in Figure 11.57(d), was employed, an even better gain, gain flatness and bandwidth could be achieved. The Cadence layout of the miniature distributed amplifier is shown in Figure 11.69.

Post-layout simulations were performed on the designed distributed amplifier. The resulting *S*-parameters and noise figure of the amplifier was calculated by importing the *S*-parameters of the passive components calculated from IE3D into ADS. The results are shown in Figure 11.70. The power gain is about 8 dB with 0.2-dB ripple in the entire frequency spectrum of DC-20 GHz. The input and output return losses stay well below 10 dB up to 17 GHz while the noise figure is less than 5 dB. The structure occupies just 1.05×0.37 mm^2 of chip area, demonstrating its miniaturization.

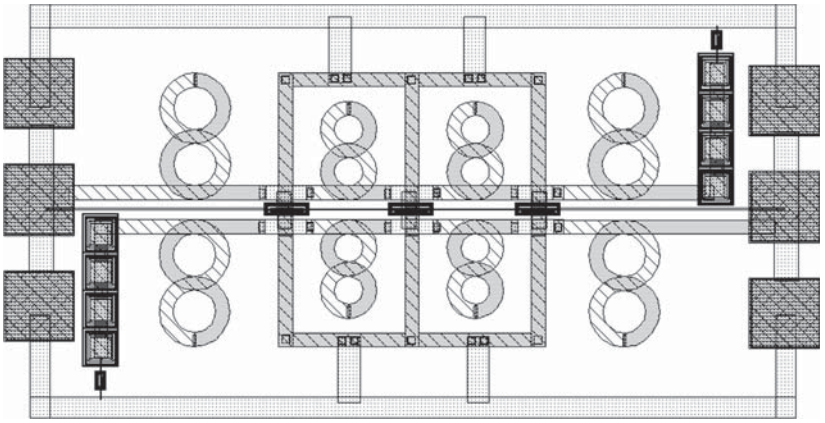


Figure 11.69. Layout of the distributed amplifier employing transmission lines and inductors. (After Chirala, Guan, and Nguyen [13]. © The Institution of Engineering and Technology.)

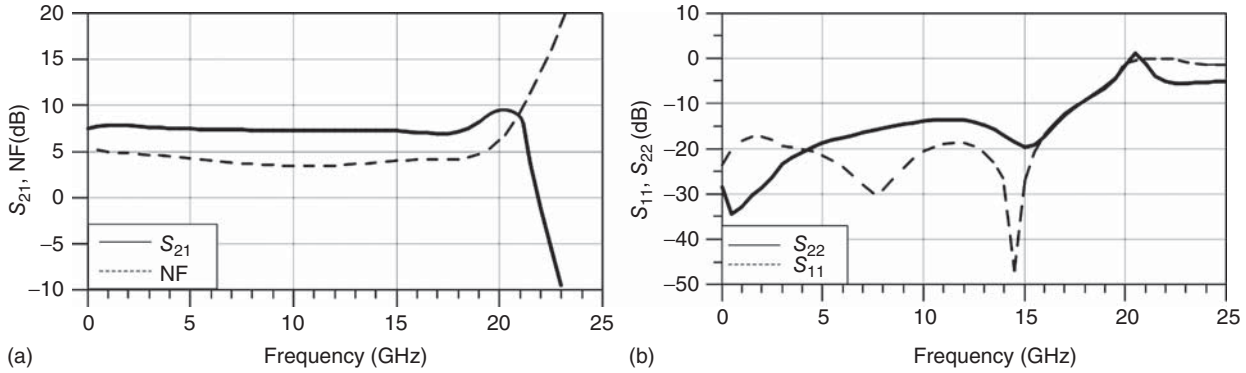


Figure 11.70. Post-layout simulation results of the miniature distributed amplifier showing (a) power gain (S_{21}) and noise figure and (b) input (S_{11}) and output (S_{22}) return loss. (After Chirala, Guan, and Nguyen [13]. © The Institution of Engineering and Technology.)

In spite of achieving compact dimensions as compared to conventional distributed amplifier topologies, the designed distributed amplifier is still 1-mm long resulting from the large size consumed by the pseudo-CPW. In the following section, we describe another design for distributed amplifiers that produce extremely compact circuit suitable for CMOS implementation. This ultrasmall distributed amplifier demonstrates the benefits of employing a distributed topology in practical CMOS devices, specifically as a LNA in a receiver front-end.

An Ultracompact Distributed Amplifier. This section discusses the design and implementation of an ultracompact distributed amplifier employing vertically integrated multilayer on-chip inductors [15]. Multilayer on-chip inductors can be configured to make them suitable for RF distributed structures. The amplifier has an extremely small size meeting the size constraint imposed by CMOS devices for wireless low cost applications.

The basic idea behind the design of the ultracompact distributed amplifier lies in the realization that, in order to achieve minimal chip area, the size of the on-chip inductors used to realize synthetic transmission lines needs to be miniaturized substantially. Hence, the design challenge is to implement inductor topologies that can limit the chip area consumption while simultaneously providing high quality factor and self resonant frequency. To simultaneously satisfy these constraints, a few observations need to be made. First, obtaining a higher quality factor inductor is not a big challenge for smaller inductor values in most CMOS technologies. In fact, obtaining a higher self-resonant frequency is more crucial for broadband amplifier design applications, which is usually the case for distributed amplifiers. Second, the inductor design needs to be application specific

as the inductors tailored to suit a distributed circuit such as voltage-controlled oscillator (VCO) may not satisfy all the requirements of another distributed circuit such as amplifier, because distributed amplifier design may require very good matching at ports, while a distributed VCO requires equalization of group delays. Another issue one needs to bear in mind is the fact that miniature inductors themselves are not difficult to design, but the key challenge lies in studying and taking into account the impact of closely packed inductor structures needed for achieving miniature circuits.

The design of the ultracompact distributed amplifier therefore involves studying the means by which the vertical multilayer inductor structures can be effectively integrated and analyzing the impact of those integrated closely packed inductors on the overall circuit, so that efficient miniaturized structure can be made possible. This is accomplished by treating all the inductors along the drains and gates of transistors, which constitute the drain and gate transmission lines, as unique multiport inductor structures and analyzing the impact of mutual coupling within that structure.

Integration of Vertical Multilayer Inductors. As mentioned earlier, while miniaturization of inductors can be generally accomplished by using multilayer for inductors, the issues of integrating those inductors efficiently for distributed amplifiers or other RFICs to conserve chip area play the most crucial role in the design. To illustrate this, we consider the vertical multilayer inductor structures implemented on a six-metal CMOS process as shown in Figures 11.71. The smaller inductor depicted in Figure 11.71(a) uses only the top two metal layers – M6 and M5 – while the longer one shown in Figure 11.71(b) uses M6, M5, and M4. Both the structures take advantage of the tightly spaced CMOS metal stack, which increases the mutual inductance between the upper and lower traces without occupying significant area. Furthermore, the thickest top-most metal M6 has lower resistivity, which enables reduction of loss associated with each inductor. The integrated behavior of these inductors raises the obvious issue of strong negative coupling between adjacent inductors when placed close to one another. Coupling between any two wires or segments of metal strongly depends on the distance between them. Assuming that the pitch of an inductor's turn (d) is varied for different inductor lengths (ℓ), mutual inductance drops inversely with respect to the distance separating the two metals as shown in Figure 11.72 [16]. This makes a strong case for having a minimum pitch when strong positive coupling is desired and maximum values for avoiding negative coupling. In circuit design implementing integrated inductors, however, a minimum pitch is chosen to integrate the inductors in order to arrive at an optimum dimension. Figure 11.73(a) shows the resulting multilayer inductor structure on a six-metal-layer CMOS process. It is apparent that the structure increases negative mutual coupling between successive inductors. However, increase in negative mutual coupling is not necessarily negative for wideband circuit's overall performance. The impact of negatively coupled inductors is suggested to be desirable in one of the earliest studies conducted on distributed amplifiers in 1948 [5]. It is claimed that negatively coupled inductors tend to linearize the amplifier phase shift. In its simplified form, the phase shift associated with the gain transfer

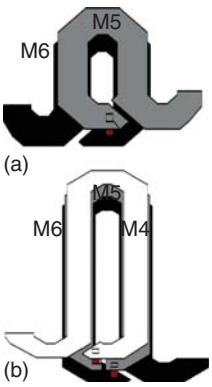


Figure 11.71. Inductors employing two (a) and three (b) vertically integrated metal layers. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

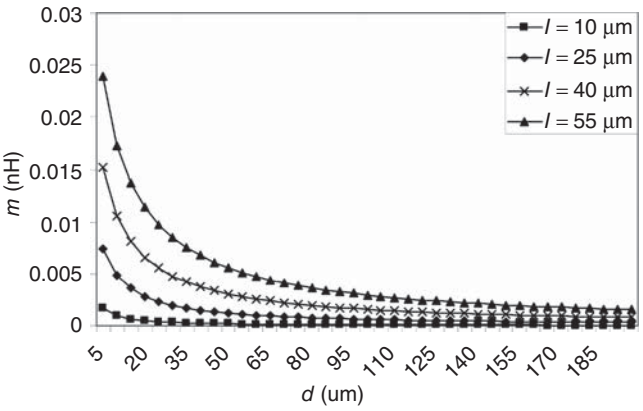


Figure 11.72. Variation of mutual inductance with pitch for different metal lengths. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

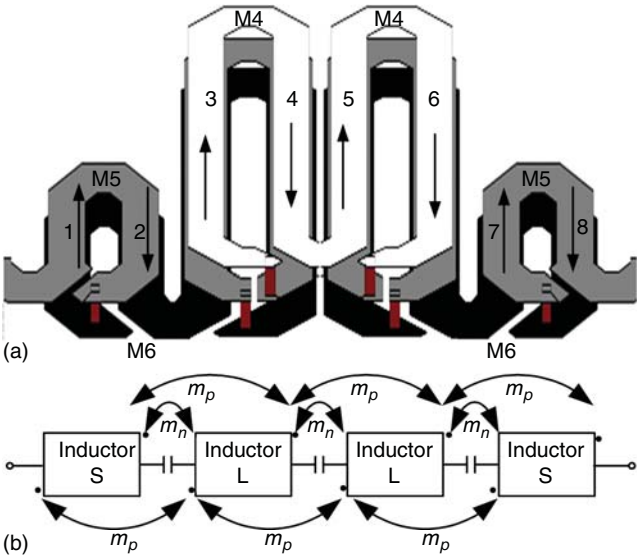


Figure 11.73. (a) Integrated multilayer inductor structure and (b) schematic representation of the above structure. Vias are used to connect different metal layers in (a). (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

function is given by:

$$\phi = 2N\tan^{-1}\left(\frac{mf/f_o}{\sqrt{m^2-(f/f_o)^2}}\right) \tag{11.307}$$

where m is the mutual coupling coefficient, N is the number of cascading gain stages, and f_o is the resonant frequency at the drain (or gate) node of the nodal inductance and the capacitance associated with that node. Equation (11.307) shows that, once the mutual coupling coefficient is greater than 1 (as is the case for negative coupling) and the resonant frequencies facilitated by smaller drain–gate segment inductances which are much higher than the operating frequency, the phase shift will be a linear function of the frequency of operation. Having a linear phase shift results in very low distortion of the signal as the group delay is related to the phase through a derivative. This property is very attractive for several time-domain applications such as UWB based on pulsed signals.

Another important observation is made through the total inductance of the integrated inductor segment of Figure 11.73. Using the approach in [17], we can derive the total inductance as

$$L_T = 2(L_{sm} + L_{lo}) - 2(m_{23} + m_{45} + m_{67}) + 4(m_{13} + m_{35} + m_{57}) + (2L_{6t} + L_{4t}) \tag{11.308}$$

where L_{sm} and L_{lo} are the overall inductances of the small and large inductors indicated by blocks S and L in the schematic of Figure 11.73(b), respectively, including their internal mutual coupling parameters; m_{ij} $(i, j = 1–8, i \neq j)$ denotes the coupling between the ith and jth segments as shown in Figure 11.73, and L_{kt} shows the inductance of the turn segment in the kth layer, owing to both self inductance and mutual inductance due to other inductor segments. The expression shows that, owing to the extremely compact dimensions, there is also a first-order positive mutual coupling between alternate segments, which significantly mitigates the impact of negative coupling. The impact of the first-order positive coupling between alternate segments can be estimated by

$$m_{ij} = 2\ell \left\{ \ln \left[\frac{\ell}{d} + \sqrt{1 + \left(\frac{\ell}{d}\right)^2} \right] - \sqrt{1 + \left(\frac{d}{\ell}\right)^2} + \frac{d}{\ell} \right\} \tag{11.309}$$

wherein ℓ and d represent the length of the metal strip and its pitch, respectively. Equation (11.308) shows how the overall inductance is not adversely affected by the strong negative mutual coupling between adjacent inductors. It is estimated that the increase in the overall inductance is about 8% if the positive mutual inductance is 52% of the value of negative coupling. The negative mutual coupling is extremely strong owing to the smaller pitch as shown in Figure 11.72. For shorter lengths, having a smaller pitch presents the best opportunity to increase mutual inductance, which explains why the multilayer inductor structure offers a larger tight positive mutual inductance between the CMOS stacked layers.

To further study the properties of this integrated inductor segment, the structure was laid out and characterized in Jazz CA18HR 0.18-μm RF/Mixed signal process [18]. The cumulative properties of the multilayered inductor segment are found from the measured and simulated S-parameters of the inductor segment shown in Figure 11.74. It is apparent that the inductor segment exhibits wideband characteristics. This can be explained by the schematic model in Figure 11.73(b) which shows how the positive mutual inductance m_p, negative mutual inductance m_n, and the series capacitance from successive segments contribute to an enhancement of the bandwidth by increasing the overall self resonant frequency of the integrated segment. While the series capacitance serves to lower the individual capacitance of each inductance, the net inductance will be slightly increased by the mutual coupling. Hence, it can be construed that the prime effects of extreme miniaturization using vertically integrated multilayer inductors on a broadband circuit performance are not necessarily

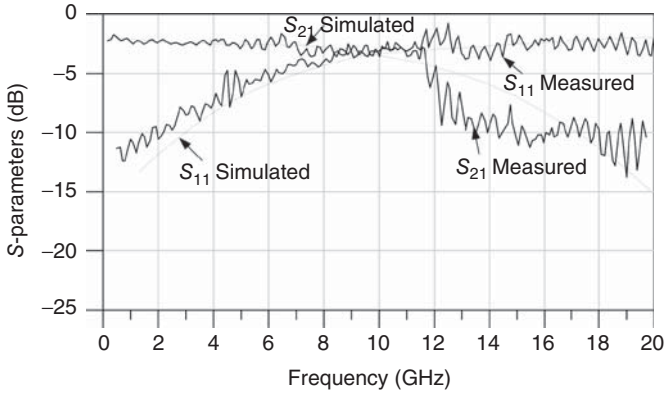


Figure 11.74. Simulated and measured S-parameters of the integrated inductor segment. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

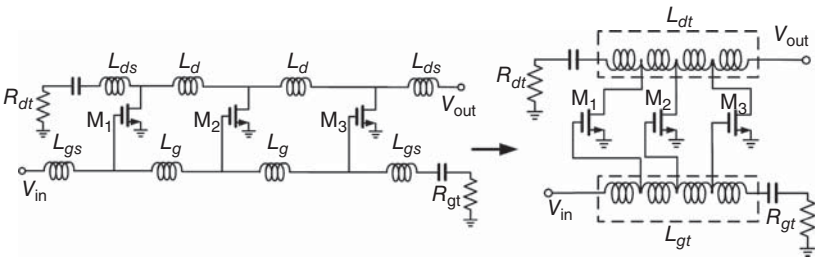


Figure 11.75. Schematic of the ultracompact distributed amplifier with corresponding five-port inductor segments used in the analysis. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

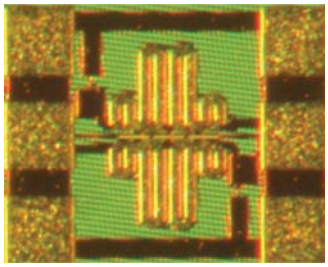


Figure 11.76. Die photograph of the fabricated ultracompact distributed amplifier (288 μm × 291 μm without RF pads). (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

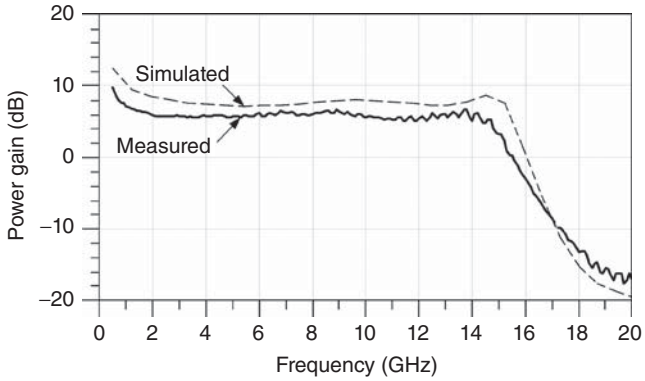


Figure 11.77. Simulated and measured power gain. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

negative to some extent and any of the impending losses can be reduced by utilizing multilayered inductors in the top metal layers and careful optimization of their design parameters.

Design and Performance. The design of the ultracompact distributed amplifier was done using Jazz CA18HR 0.18-μm RF/Mixed signal process [18]. It requires optimization of transistors' width for maximum gain and lower power consumption requirements. The requirements were met for an aspect ratio of 136/0.18 (W/L). The gate inductance resonates with transistor's input parasitic capacitance influencing the noise spectrum and, as such, small inductances of 0.85 and 0.4 nH are used to meet the bandwidth and noise requirements. The entire inductor segment was then jointly optimized as a five-port network in IE3D to take into account both the negative and positive mutual coupling associated with adjacent multilayer inductors. Figure 11.75 shows the schematic of the distributed amplifier treating the inductors as a single five-port so that drain and gate inductors could be jointly optimized. The amplifier was further tested for stability to check for unwanted oscillations, and calculations proved that it remains unconditionally stable up to 20 GHz.

Figure 11.76 shows a photograph of the distributed amplifier die. The fabricated amplifier was measured using on-wafer probes and bias tees. In order to minimize reflection and enhance isolation, 50-Ω on-chip

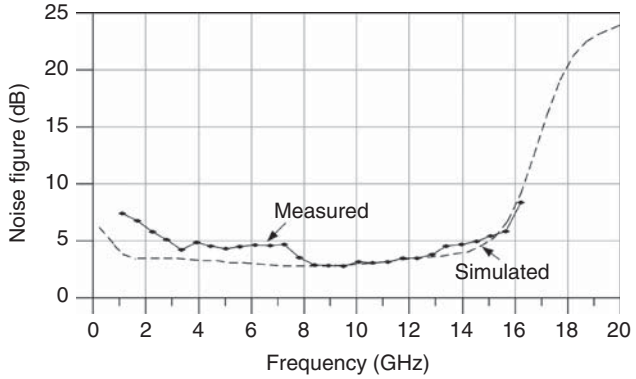


Figure 11.78. Simulated and measured noise figure. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

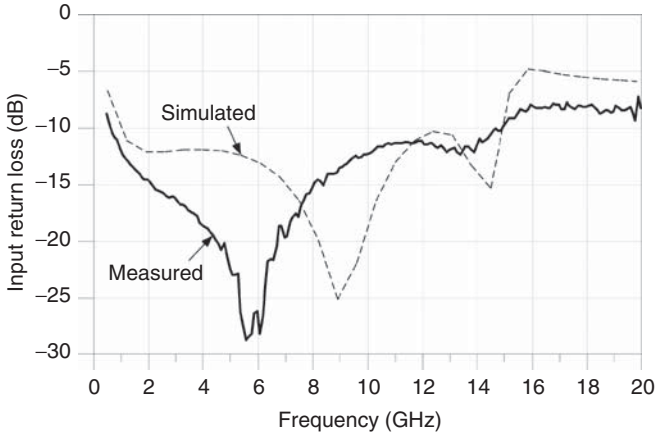


Figure 11.79. Simulated and measured input return loss. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

resistors were used along with 10-pF DC blocking capacitors at the isolation terminals of the gate and drain transmission lines. Figure 11.77 shows the measured power gain of the amplifier. The amplifier exhibits around 6-dB flat gain throughout the entire 3.1–10.6 GHz range, which follows very well the simulated results, albeit with 1.5-dB lower gain than expected. This difference is primarily attributed to the discrepancies in via-modeling as well as parasitics that could not be de-embedded. Noise measurement results are shown in Figure 11.78, which indicate a noise figure of 2.7 dB from 8 to 10 GHz and less than 5 dB over a bandwidth of 14 GHz, which reasonably matches the calculated results. The input and output return losses, as shown in Figures 11.79 and 11.80, remain well below 12 and 15 dB, respectively. Simulations indicate that the amplifier shows a third-order input intermodulation product (IIP3) of about +13.2 dBm at 14 GHz, signifying that the amplifier maintains an exceptional linearity even at the end of the operating spectrum. The circuit draws a 12 mA from a 1.8-V power supply.

It must be noted that the design of the ultracompact distributed amplifier is merely used to demonstrate the applicability of the vertically integrated multilayer inductor structures to the realization of an extremely small distributed amplifier, or any other components, suitable for practical RFIC, and hence a simple single common-source transistor gain cell was used, which inevitably does not show a large gain. Nevertheless, larger gain can be easily achieved using higher-gain cells such as the cascade common-source gain cells shown in Figure 11.57(d), while still achieving an ultracompact size due to the fact that the size of the inductors dictate the overall amplifier's size. However, judging on the overall aspects of size, matching, power consumption, linearity, and noise figure, the designed distributed amplifier still compares favorably to the performance of other distributed amplifiers, especially noting that it merely occupies a maximum of 10% of the chip area reported by the next smallest counterpart. This amplifier is therefore pursued as a high linearity, low noise,

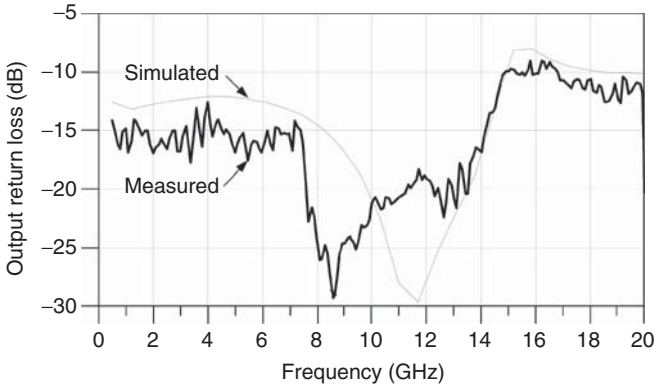


Figure 11.80. Simulated and measured output return loss. (After Chirala, Guan, and Nguyen [15]. Reprinted with permission of IEEE.)

broadband amplifier that can be integrated in realistic low cost CMOS wideband receiver front-end. Furthermore, the discussed integrated inductor structure can also be used to attain extremely small size for other RFICs such as distributed VCO or mixer.

11.6.3 Feedback Amplifiers

The feedback amplifier was invented in 1927 by Black [19]. Feedback is a classical yet conventional approach for wideband amplifier design. The topology of the shunt feedback amplifiers considered here is based on "negative" resistive feedback as opposed to "positive" feedback employed for oscillators. The negative feedback technique has been widely used for RFIC amplifiers, particularly wideband amplifiers, because of several advantages including simplicity, wide bandwidth, small size, low power consumption, stability of gain against changes in transistors' parameters, and reduced signal distortion. Typical feedback amplifiers, however, have low gain, poor reverse isolation, poor linearity, and potential oscillation due to the feedback. The potential oscillation is particularly troublesome at high frequencies, especially for broadband design. Additionally, the noise caused by the resistive feedback network degrades the overall amplifier noise figure. Nevertheless, with careful design, RFIC feedback amplifiers can be realized with decent gain, low noise figure, good isolation, and high stability.

11.6.3.1 Analysis. Figure 11.81(a) shows a general simple schematic of a shunt feedback amplifier and its equivalence using a CMOS transistor. The input impedance of the amplifier as shown in Figure 11.81(b) can be derived as

$$Z_{\text{in}} = \frac{v_{\text{in}}}{i_{\text{in}}} = \frac{v_{\text{in}}}{\frac{v_{\text{in}} - v_{\text{out}}}{R_f}} = \frac{R_f}{1 - \frac{v_{\text{out}}}{v_{\text{in}}}} = \frac{R_f}{1 + |A|} \tag{11.310}$$

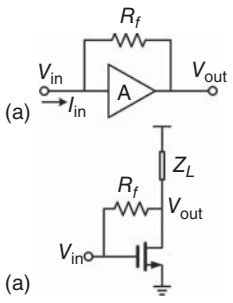


Figure 11.81. (a) General schematic of a shunt feedback amplifier and (b) with a CMOS transistor. Z_L represents the load impedance.

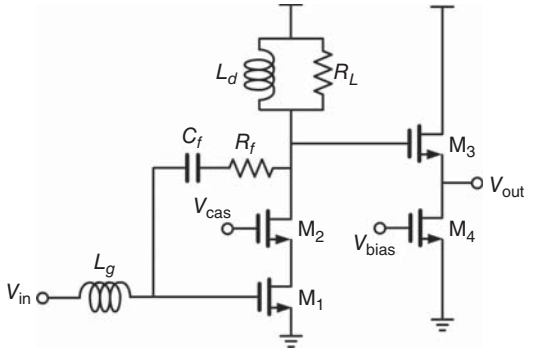


Figure 11.82. Schematic of a shunt feedback amplifier.

where $A = v_{out}/v_{in}$ represents the voltage gain and R_f is the feedback resistance. Equation (11.310) shows that, with properly chosen values for A and R_f, Z_{in} can be theoretically made to be equal to almost any impedance needed for input matching – typically 50 Ω. In practice, however, the problem is more complicated because of the inherent relations between different amplifier parameters such as gain and feedback resistance. These relations are difficult to derive accurately, especially at RF frequencies and for submicron MOSFETs. The problem is much more pronounced for complex amplifier topologies represented by Block A in Figure 11.81(a). Specifically, since the amplifier gain is a function of the feedback resistance, load impedance, and frequency, Eq. (11.310) can only give an approximated value for Z_{in} which is valid only at low frequencies. For extremely wide bandwidths, such as the 3.1–10.6 GHz range used for UWB applications, it is challenging to achieve decent gain and noise figure across the entire bandwidths using the simple shunt feedback technique.

Figure 11.82 shows the schematic of a CMOS amplifier implementing the shunt feedback technique. The topology, as can be seen, is very simple. This simplicity results in small size, making it ideally suited for low cost CMOS applications. The amplifier consists of two stages. The first stage has two transistors connected in cascode with a resistive shunt feedback. In this stage, R_f is the feedback resistor, C_f is the blocking capacitor and L_g is the inductor connecting the gate of the transistor M_1 to the input of the amplifier for input matching purpose. This stage also contains a loading element formed by the resistor R_L in parallel with the inductor L_d. It is particularly noted that this load element in the feedback amplifier is important because the amplifier's input impedance also depends on it. The second stage is a conventional source follower, which primarily contributes to the output matching and provides a sufficient current to drive the (output) amplifier load, which is typically 50 Ω. The analysis for the important parameters of the feedback amplifier including the input and output matching, gain and noise figure is presented as follows.

Input Matching

As mentioned earlier, the input matching of the shunt feedback amplifier is primarily determined by the first stage of the amplifier. As such, we analyze the amplifier's input matching by considering only the first stage – specifically its small-signal model as shown in Figure 11.83. In the figure, C_L is the load capacitance representing the sum of the total parasitic capacitance at the output node of the first stage and the input capacitance of the second stage; L_d and R_L constitute the parallel load at the drain of the transistor M_2 seen in Figure 11.82; g_m is the transconductance of the transistor M_1 in Figure 11.82. The gate–drain capacitance C_{gd1} of M_1 and the small-signal resistance r_{ds12} between the source of M_1 and the drain of M_2 are not considered in the small-signal model as the cascode configuration of the first stage almost eliminates the impact of these elements. Moreover, C_f is also neglected in the small-signal analysis since it functions only as a blocking capacitor. Furthermore, for simplicity, we let Z_L represent the parallel RLC network consisting of R_L, L_d, and C_L as the load impedance. Figure 11.83(b) shows the resultant simplified small-signal model used for the input-matching analysis.

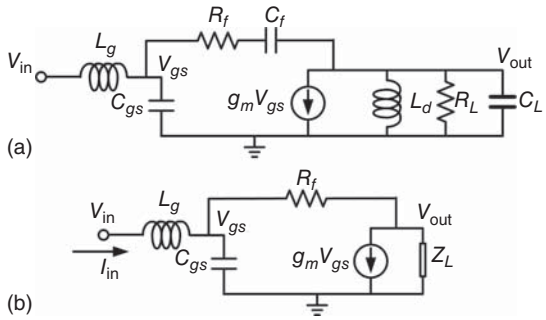


Figure 11.83. (a) Small-signal model of the shunt feedback amplifier and (b) its simplification used in the input-matching analysis.

Applying Kirchoff current law (KCL) to Figure 11.83(b), we can write

$$\frac{v_{gs} - v_{\text{out}}}{R_f} + jwC_{gs}v_{gs} = i_{\text{in}} \tag{11.311}$$

and

$$g_m v_{gs} + \frac{v_{\text{out}}}{Z_L} = \frac{v_{gs} - v_{\text{out}}}{R_f} \tag{11.312}$$

Solving for v_{out} in (11.312) and substituting into (11.311) gives

$$v_{\text{out}} = \frac{Z_L(1 - Z_f g_m)}{Z_L + Z_f} v_{gs} \tag{11.313}$$

Substituting v_{out} from (11.313) into (11.311), we get

$$\left[1 - \frac{Z_L\left(1 - R_f g_m\right)}{Z_L + R_f}\right] v_{gs} + jwC_{gs}R_f v_{gs} = i_{\text{in}}R_f \tag{11.314}$$

or

$$\frac{1 + Z_L g_m}{Z_L + R_f} v_{gs} + jwC_{gs}v_{gs} = i_{\text{in}} \tag{11.315}$$

from which, we obtain

$$\frac{i_{\text{in}}}{v_{gs}} \equiv Y'_{\text{in}} = jwC_{gs} + \frac{1 + Z_L g_m}{Z_L + R_f} \tag{11.316}$$

where Y'_{in} is the admittance looking into the gate of M_1. It is possible to choose the load network to produce large Z_L along with proper bias for the transistor to obtain $Z_L g_m \gg 1$ across the frequencies of interest and, under this assumption, (11.316) can be simplified as

$$Y'_{\text{in}} \simeq j\omega C_{gs} + \frac{Z_L g_m}{Z_L + R_f} = jwC_{gs} + \frac{1}{\frac{1}{g_m} + \frac{R_f/g_m}{Z_L}} \tag{11.317}$$

Substituting Z_L obtained from Figure 11.83(a) as

$$\frac{1}{Z_L} = \frac{1}{R_L} + j\omega C_L + \frac{1}{j\omega L_d} \tag{11.318}$$

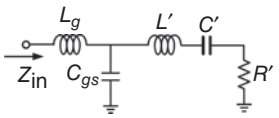


Figure 11.84. Equivalent circuit of the shunt feedback amplifier's input network.

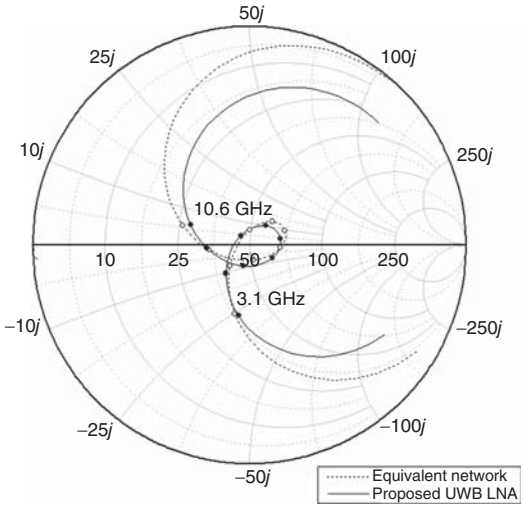


Figure 11.85. Calculated input reflection coefficients of the shunt feedback amplifier and its equivalent input network.

into (11.317) gives

$$Y'_{\text{in}} \simeq jwC_{gs} + \frac{1}{\frac{1}{g_m} + \frac{R_f}{g_m}\left(\frac{1}{R_L} + j\omega C_L + \frac{1}{j\omega L_d}\right)} \tag{11.319}$$

which becomes, after letting $q = \frac{R_f}{g_m}$

$$Y'_{\text{in}} \simeq jwC_{gs} + \frac{1}{\frac{1}{g_m}\left(\frac{R_f}{R_L} + 1\right) + j\omega q C_L + \frac{1}{j\omega\frac{L_d}{q}}} \tag{11.320}$$

Equation (11.320) shows that the first part of Y'_{in} is contributed by the parasitic capacitance at the gate of M_1, while its second part is formed by the load impedance coupled through the feedback resistor R_f. This part of impedance can be described as the impedance of a serial RLC network, consisting of resistance $R' = \frac{1}{g_m}\left(\frac{R_f}{R_L} + 1\right)$, capacitance $C' = \frac{L_d}{k}$, and inductance $L' = qC_L$, which represents the equivalent impedance seen at the input port through the feedback resistor. Using this equivalent impedance, an equivalent circuit for the input network of the shunt feedback amplifier can be derived as shown in Figure 11.84. Now letting $R_f \simeq R_L$ leads to R' depending only on g_m. Under this condition and with a proper value for g_m obtained with specific bias voltages, R' can then be made to approximately equal to a real impedance such as 50 Ω. Consequently, the input network behaves just like a lumped-elements band-pass filter and hence wideband matching can be easily achieved with proper design.

To validate the derived input equivalent circuit, we show in Figure 11.85 the calculated input reflection coefficients of the shunt feedback amplifier shown in Figure 11.82 and the equivalent input network shown in Figure 11.84 versus frequency from 1 to 20 GHz. The transistors used in the calculations are MOSFETs available in a 0.18-μm CMOS process and the values of other circuit elements are $L_g = 1.0$ nH and $R_f = R_L = 400\,\Omega$. The results show a very good match between those calculated from the input equivalent circuit and the shunt feedback amplifier across an extremely wide frequency range from 3.1 to 10.6 GHz. Good input matching

is also obtained over this frequency range. Furthermore, the input reflection coefficients also have similar pattern across 1–20 GHz. The deviation of the results between the two circuits gradually increases below 3 GHz and beyond 10 GHz due to the assumption of $Z_L g_m \gg 1$ made previously that is no long valid at those frequencies for the considered load and MOSFET.

Gain

The gain can also be derived approximately using the small-signal model shown in Figure 11.83(b). From that model, we can write, making use of the input impedance $Z_{\text{in}} = v_{\text{in}}/i_{\text{in}}$:

$$v_{\text{in}} = v_{gs} + jwL_g i_{\text{in}} = v_{gs} + jwL_g \frac{v_{\text{in}}}{Z_{\text{in}}} \tag{11.321}$$

from which, we get

$$v_{gs} = \frac{Z_{\text{in}} - jwL_g}{Z_{\text{in}}} v_{\text{in}} \tag{11.322}$$

Substituting (11.322) into (11.313), we obtain

$$v_{\text{out}} = \frac{Z_L(1 - R_f g_m)(Z_{\text{in}} - jwL_g)}{Z_{\text{in}}(Z_L + R_f)} v_{\text{in}} \tag{11.323}$$

The (voltage) gain of the shunt feedback amplifier can now be obtained from (11.323) as

$$G = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{Z_L(1 - R_f g_m)(Z_{\text{in}} - jwL_g)}{Z_{\text{in}}(Z_L + R_f)} \tag{11.324}$$

From Figure 11.83(b), we can write making use of Eq. (11.316):

$$Z_{\text{in}} - jwL_g = \frac{1}{Y'_{\text{in}}} = \frac{1}{j\omega C_{gs} + \frac{Z_L g_m + 1}{Z_L + R_f}} \tag{11.325}$$

The gain of the shunt feedback amplifier in (11.324) can then be rewritten using (11.325) as

$$\begin{aligned} G &= \frac{1}{Z_{\text{in}}} \cdot \frac{Z_L(1 - R_f g_m)}{Z_L + R_f} \cdot \frac{1}{j\omega C_{gs} + \frac{Z_L g_m + 1}{Z_L + R_f}} \\ &= \frac{1}{Z_{in}} \cdot \frac{Z_L(1 - R_f g_m)}{j\omega C_{gs}(Z_L + R_f) + (Z_L g_m + 1)} \end{aligned} \tag{11.326}$$

By choosing proper values for R_f, Z_L, and g_m (through proper bias voltages), we can achieve $R_f g_m \gg 1$ and $Z_L g_m \gg 1$. Under these conditions, we can approximate the gain in (11.326) as

$$\begin{aligned} G &\simeq -\frac{1}{Z_{\text{in}}} \cdot \frac{Z_L R_f g_m}{j\omega C_{gs}(Z_L + R_f) + Z_L g_m} \\ &= -\frac{R_f}{Z_{\text{in}}} \cdot \frac{1}{1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right)} \end{aligned} \tag{11.327}$$

where

$$\omega_T = 2\pi f_T \simeq \frac{g_m}{C_{gs}} \tag{11.328}$$

is the (radian) cut-off frequency (or frequency of unity current gain) of the MOSFET. As can be seen in Figure 11.83(a), under the resonance condition between L_d and C_L, R_f/Z_L becomes R_f/R_L. As R_f and R_L are comparable, R_f/Z_L is approximately equal to 1 and $j\frac{\omega}{\omega_T}\left(1+\frac{R_f}{Z_L}\right)$ can then be neglected when $\omega \ll \frac{\omega_T}{2}$. Consequently, the gain of the shunt feedback amplifier from (11.327) can be simply estimated as

$$G \simeq -\frac{R_f}{Z_{\text{in}}} \tag{11.329}$$

when the amplifier's operating frequency is much lower than half of the MOSFET's cut-off frequency. Since Z_{in} is normally fixed (typically 50 Ω), the gain of the shunt feedback amplifier is mainly determined by the feedback resistance R_f. It should be noted that, in (11.327), $j\frac{\omega}{\omega_T}\left(1+\frac{R_f}{Z_L}\right)$ reflects how the amplifier's operating frequency with respect to the cut-off frequency of the employed MOSFET affects the gain of the amplifier and how the RLC parallel network, as seen in Figure 11.83(a), helps boost the gain. Our design example of the shunt feedback amplifier operating from 3 to 10.6 GHz, to be described later, uses 0.18-μm CMOS transistors with f_T around 70 GHz. For this amplifier, the gain obtained from (11.329) can be used as a good initial approximation.

Noise Figure

The noise of the shunt feedback amplifier is dominated by two parts: the noise of the feedback resistor R_f and the noise from the transistor M_1. Figure 11.86 shows a small-signal model used for the noise analysis of the shunt feedback amplifier, which consists of the (voltage) noise source e_{n,R_f} representing the thermal noise of the feedback resistor and the transistor's induced gate noise, due to the coupling of the fluctuating channel charge into the gate terminal, and channel thermal noise, due to the carrier thermal agitation in the channel, represented by the currents i_{ng} and i_{nd}, respectively. The power spectrum density (PSD) of e_{n,R_f} can be expressed as

$$\overline{e_{n,R_f}^2} = 4kTR_f\Delta f \tag{11.330}$$

where k is the Boltzman constant, T is the absolute temperature in kelvin, Δf is the bandwidth over which the noise is measured or calculated. The PSD of the induced gate noise of the transistor M_1 is given by

$$\overline{i_{ng}^2} = 4kT\delta g_g\Delta f \tag{11.331}$$

where δ is the coefficient of the induced gate noise and approximately around 4/3 for long-channel transistors and g_g is the equivalent shunt gate conductance given as

$$g_g = \frac{\omega^2 C_{gs}^2}{5g_{d0}} \tag{11.332}$$

where g_{d0} is the channel conductance at $V_{ds} = 0$ V. Similarly, the PSD of the channel thermal noise of the transistor M_1 can be expressed as

$$\overline{i_{nd}^2} = 4kT\gamma g_{d0}\Delta f \tag{11.333}$$

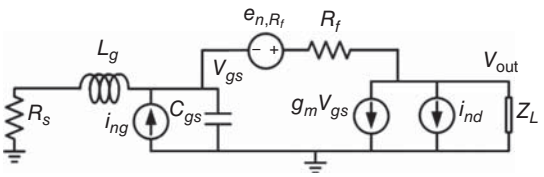


Figure 11.86. Small-signal model of the shunt feedback amplifier used for the noise analysis.

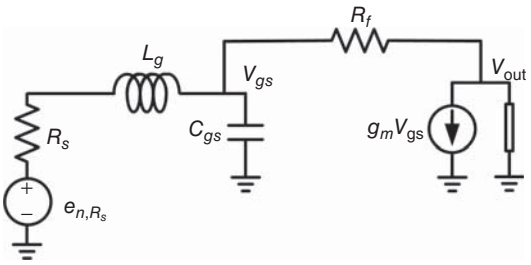


Figure 11.87. Small-signal model for the analysis of the noise contribution of the source resistor.

where γ is the coefficient associate with the transistor channel – around 2/3 for long-channel transistors and as large as 2.5 for short-channel transistors. The induced gate noise is correlated with the channel thermal noise with a correlation coefficient c defined as

$$c = \frac{\overline{i_{ng} \cdot i_{nd}^*}}{\sqrt{\overline{i_{ng}^2 \cdot i_{nd}^2}}} \tag{11.334}$$

For long-channel devices, $c \simeq -j0.395$.

The noise figure of the shunt feedback amplifier is contributed by the output noise PSD's of the source and feedback resistors. Figure 11.87 shows the small-signal model used for the analysis of the noise contribution from the source resistor. Applying KCL to Figure 11.87 gives

$$\frac{e_{n,R_s} - v_{gs}}{R_s + j\omega L_g} = j\omega C_{gs} v_{gs} + \frac{v_{gs} - v_{\text{out}}}{R_f} \tag{11.335}$$

and

$$\frac{v_{gs} - v_{\text{out}}}{R_f} = g_m v_{gs} + \frac{v_{\text{out}}}{Z_L} \tag{11.336}$$

We obtain from (11.336)

$$v_{gs} = \frac{Z_L + R_f}{Z_L(1 - g_m R_f)} v_{\text{out}} \tag{11.337}$$

Substituting v_{gs} from (11.337) into (11.335) and solving for v_{out}, we get

$$v_{\text{out}} = -\frac{g_m}{\frac{1}{Z_L} + \frac{1}{R_f}} \cdot \frac{1}{1 + (R_s + jwL_g)\left(\frac{1}{R_f} + jwC_{gs}\right)} e_{n,Rs} \tag{11.338}$$

The output noise PSD contributed by the source resistor can now be derived from (11.338) and the (rms) noise voltage per hertz from (11.88) as

$$S_{n,R_s} = v_{\text{out}}^2 = \frac{4kTR_s}{\frac{1}{g_m^2}\left|\frac{1}{Z_L} + \frac{1}{R_f}\right|^2 |1 + t|^2} \tag{11.339}$$

where t is defined as

$$t = (R_s + jwL_g)\left(\frac{1}{R_f} + jwC_{gs}\right) \tag{11.340}$$

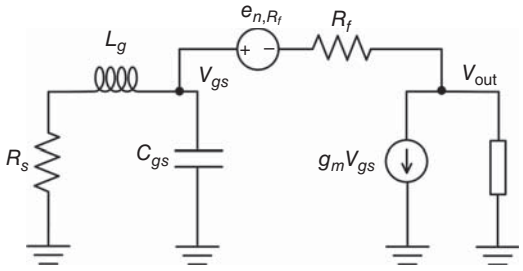


Figure 11.88. Small-signal model for the analysis of the noise contribution of the feedback resistance.

Similarly, Figure 11.88 shows the small-signal model used for determining the output noise PSD contributed by the feedback resistor R_f. We apply KCL to Figure 11.88 and obtain

$$\frac{e_{n,R_f} + v_{gs} - v_{\text{out}}}{R_f} = g_m v_{gs} + \frac{v_{\text{out}}}{Z_L} = -\frac{v_{gs}}{(R_s + j\omega L_g)//\frac{1}{j\omega C_{gs}}} \tag{11.341}$$

Eliminating v_{gs} in (11.341) leads to

$$e_{n,R_f} = \left(1 + \frac{R_f}{Z_L} \cdot \frac{j\omega C_{gs} + \frac{1}{R_s + j\omega L_g}}{g_m + j\omega C_{gs} + \frac{1}{R_s + j\omega L_g}}\right) v_{\text{out}} \tag{11.342}$$

Defining

$$p = j\frac{\omega}{\omega_T} + \frac{1}{g_m(R_s + j\omega L_g)} \tag{11.343}$$

and solving for v_{out} from (11.342), we obtain

$$v_{\text{out}} = \frac{e_{n,R_f}}{1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p}} \tag{11.344}$$

The output noise PSD contributed by the feedback resistor can now be derived using (11.344) and (11.88) as

$$S_{n,R_f} = v_{\text{out}}^2 = \frac{4kTR_f}{\left|1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p}\right|^2} \tag{11.345}$$

Since the transistor is normally biased to produce large transconductance g_m needed for gain and ω/ω_T is typically small, p can be considered small. Furthermore, since Z_L and R_f are comparable within the operating bandwidth as discussed earlier, S_{n,R_f} can be approximated as $4kTR_f$ with very little fluctuation within the operating frequency range. It is noted that the frequency response of S_{n,R_f} is mainly caused by those of p and Z_L representing the RLC parallel network in Figure 11.83(a).

The noise contribution of the feedback resistor to the overall noise factor or noise figure of the shunt feedback amplifier can be derived from

$$F_{R_f} = \frac{S_{n,R_f}}{S_{n,R_s}} \tag{11.346}$$

where F_{R_f} represents the noise factor due to R_f. Substituting (11.339) and (11.345) into (11.346), we get

$$F_{R_f} = \frac{R_f}{R_s} \cdot \frac{1}{g_m^2} \left| \frac{1}{Z_L} + \frac{1}{R_f} \right|^2 \cdot \frac{|1+t|^2}{\left| 1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p} \right|^2} \tag{11.347}$$

A simplified expression for the noise factor F_{R_f} can be derived noting that the output noise PSD contributed by the source resistor can also be obtained from the gain expression in (11.327). Referring to Figure 11.87, the output voltage can be written as

$$v_{\text{out}} = G v_{\text{in}} = G \frac{Z_{\text{in}}}{R_s + Z_{\text{in}}} e_{n,R_s} \tag{11.348}$$

Substituting G from (11.327) into (11.348) yields

$$v_{\text{out}} = -\frac{R_f}{R_s + Z_{\text{in}}} \cdot \frac{1}{1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right)} e_{n,R_s} \tag{11.349}$$

The output noise PSD contributed by the source can then be obtained from (11.349) and (11.88) as

$$S_{n,R_s} = 4kTR_s \left| \frac{R_f}{Z_{\text{in}} + R_s} \cdot \frac{1}{1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right)} \right|^2 \tag{11.350}$$

The noise factor F_{R_f} can thus be obtained from (11.345) and (11.350) as

$$F_{R_f} = \frac{S_{n,R_f}}{S_{n,R_s}} = \frac{|Z_{\text{in}} + R_s|^2}{R_f R_s} \cdot \frac{\left| 1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right) \right|^2}{\left| 1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p} \right|^2} \tag{11.351}$$

When the amplifier is matched at the input, $Z_{\text{in}} \simeq R_s$, and this noise factor becomes

$$F_{R_f} = 4\frac{R_s}{R_f} \cdot \frac{\left| 1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right) \right|^2}{\left| 1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p} \right|^2} \tag{11.352}$$

which can be further simplified using $G \simeq -\frac{R_f}{Z_{\text{in}}}$ from (11.329) as

$$F_{R_f} = \frac{4}{|G|} \cdot \frac{\left| 1 + j\frac{\omega}{\omega_T}\left(1 + \frac{R_f}{Z_L}\right) \right|}{\left| 1 + \frac{R_f}{Z_L} \cdot \frac{p}{1+p} \right|^2} \simeq \frac{4}{|G|} \tag{11.353}$$

This noise-factor expression shows that the contribution of the feedback resistor to the noise figure of the shunt feedback amplifier does not increase as the feedback resistance increases. On the contrary, the output noise is lower as the feedback resistance increases. It seems that a larger feedback resistance can lead to both higher gain and lower noise figure. However, because of the trade-off between gain and bandwidth, this resistance cannot be selected too large to avoid reduction in the bandwidth.

The output noise contributed by the induced gate noise is now derived. Applying KCL to the small-signal model shown in Figure 11.86, considering only the existence of the induced gate noise current source i_{ng}, we can write

$$i_{ng} = \frac{v_{gs}}{Z_g} + \frac{v_{gs} - v_{\text{out}}}{R_f} \tag{11.354}$$

and

$$\frac{v_{gs} - v_{\text{out}}}{R_f} = g_m v_{gs} + \frac{v_{\text{out}}}{Z_L} \tag{11.355}$$

where

$$Z_g = (R_s + j\omega L_g) // \left(\frac{1}{j\omega C_{gs}}\right) \tag{11.356}$$

Equation (11.354) becomes, after eliminating v_{gs} using (11.355)

$$i_{ng} = v_{\text{out}} \left[\frac{R_f + Z_L}{Z_L Z_g \left(1 - R_f g_m\right)} + \frac{1 + Z_L g_m}{Z_L (1 - R_f g_m)}\right] \tag{11.357}$$

which can be simplified, under the assumption of $Z_L g_m \gg 1$ as previously made, to

$$i_{ng} \simeq \left[\frac{R_f + Z_L + Z_g}{Z_L Z_g R_f g_m} - \frac{1}{R_f}\right] v_{\text{out}} \tag{11.358}$$

Substituting (11.356) into (11.358) and making use of (11.343), we get

$$i_{ng} \simeq \left(\frac{p}{Z_L} + \frac{1+p}{R_f}\right) v_{\text{out}} \tag{11.359}$$

The output noise PSD due to the induced gate noise of the transistor M_1 is obtained, using (11.331) and (11.359), as

$$S_{ng} = v_{\text{out}}^2 = \frac{4kT\delta g_g}{\left|\frac{p}{Z_L} + \frac{1+p}{R_f}\right|^2} \tag{11.360}$$

We now consider the output noise contributed by the channel thermal noise. Applying KCL to the small-signal model shown in Figure 11.86, considering only the channel thermal noise represented by the current noise source i_{nd}, results in

$$\frac{v_{gs} - v_{\text{out}}}{Z_f} = j\omega C_{gs} v_{gs} + \frac{v_{gs}}{R + j\omega L_g} \tag{11.361}$$

and

$$\frac{v_{\text{out}}}{Z_L} + i_{nd} + g_m v_{gs} = \frac{v_{gs} - v_{\text{out}}}{Z_L} \tag{11.362}$$

Eliminating v_{gs} in (11.362) using (11.361) yields

$$v_{\text{out}} = -\frac{1}{\frac{1}{Z_L} + \frac{Z_g g_m + 1}{R_f + Z_g}} i_{nd} \tag{11.363}$$

Substituting (11.356) into (11.363) and making use of (11.343) give

$$v_{\text{out}} = -\frac{i_{nd}}{\frac{1}{Z_L} + \frac{1+p}{R_f p+1/g_m}} \tag{11.364}$$

from which, the output noise PSD contributed by the channel thermal noise can be derived upon using (11.343) as

$$S_{nd} = v_{\text{out}}^2 = \frac{4kT\gamma g_{d0}}{\left|\frac{1}{Z_L} + \frac{1+p}{R_f p+1/g_m}\right|^2} \tag{11.365}$$

The induced gate noise and channel thermal noise of the transistor are correlated as mentioned earlier. This correlation causes another noise to the amplifier, which can be obtained from

$$v_{\text{out,total noise}} = v_{\text{out},ng} + v_{\text{out},nd} = ai_{ng} + bi_{nd} \tag{11.366}$$

where $v_{\text{out,total noise}}$ is the total output noise due to the sum of the gate induced noise and channel thermal noise, $v_{\text{out},ng}$ and $v_{\text{out},nd}$ are the output noise due to gate induced noise and channel thermal noise, respectively, and a and b are the noise impedances relating the output noise voltages to current noise sources in the transistor M_1. These noise impedances are obtained from (11.359) and (11.364) as

$$a = \frac{1}{\frac{p}{Z_L} + \frac{1+p}{R_f}} \tag{11.367}$$

$$b = -\frac{1}{\frac{1}{Z_L} + \frac{1+p}{R_f p+1/g_m}} \tag{11.368}$$

respectively.

The PSD of the output noise due to the sum of the gate induced noise and channel thermal noise can be derived using (11.365) as

$$\begin{aligned} S_{n,\text{total}} &= \overline{v_{\text{out,total noise}}^2} = (v_{\text{out},ng} + v_{\text{out},nd})(v_{\text{out},ng} + v_{\text{out},nd})^* \\ &= (ai_{ng} + bi_{nd})(ai_{ng} + bi_{nd})^* \\ &= (ai_{ng} + bi_{nd})(a^* i_{ng}^* + b^* i_{nd}^*) \\ &= |a|^2\overline{i_{ng}^2} + |b|^2\overline{i_{nd}^2} + ai_{ng}b^* i_{nd}^* + a^* i_{ng}^* bi_{nd} \\ &= |a|^2\overline{i_{ng}^2} + |b|^2\overline{i_{nd}^2} + 2\text{Re}(ab^* i_{ng} i_{nd}^*) \\ &= \overline{i_{nd}^2}\left[|a|^2\frac{\overline{i_{ng}^2}}{\overline{i_{nd}^2}} + |b|^2 + 2\text{Re}\left(ab^* c\sqrt{\frac{\overline{i_{ng}^2}}{\overline{i_{nd}^2}}}\right)\right] \end{aligned} \tag{11.369}$$

where * indicates the conjugate and c is the correlation coefficient given in (11.334). In (11.369), the first two parts are the output noise PSD's due to the gate induced noise and channel noise, and the last term is caused by the correlation of these noises. This correlated noise is written specifically from (11.369) as

$$S_{ndg,c} = 2\overline{i_{nd}^2}\text{Re}\left[ab^* c\sqrt{\frac{\overline{i_{ng}^2}}{\overline{i_{nd}^2}}}\right] \tag{11.370}$$

$\sqrt{\frac{\overline{i_{ng}^2}}{\overline{i_{nd}^2}}}$ can be obtained from (11.331) to (11.333)as

$$\sqrt{\frac{\overline{i_{ng}^2}}{\overline{i_{nd}^2}}} = \sqrt{\frac{\delta\omega^2 C_{gs}^2}{5\gamma g_{d0}^2}} = \frac{\omega C_{gs}}{g_{d0}}\sqrt{\frac{\delta}{5\gamma}} \tag{11.371}$$

Substituting (11.371) into (11.370), we can write the output noise PSD contributed by the correlation of the gate induced noise and channel noise as

$$S_{ndg,c} = 8kTwC_{gs}\sqrt{\frac{\delta\gamma}{5}}\mathrm{Re}\left[\frac{c}{\left(\frac{1}{Z_L}+\frac{1+p}{R_f p+1/g_m}\right)\left(\frac{p}{Z_L}+\frac{1+p}{R_f}\right)^*}\right] \tag{11.372}$$

Finally, the total noise figure of the shunt feedback amplifier can be derived from

$$F = \frac{S_{n,\text{total}}}{S_{n,R_s}} = 1 + \frac{S_{n,R_f}}{S_{n,R_s}} + \frac{S_{ng}}{S_{n,R_s}} + \frac{S_{nd}}{S_{n,R_s}} + \frac{S_{ndg,c}}{S_{n,R_s}} \tag{11.373}$$

which becomes, after substituting the individual PSD's from (11.339), (11.345), (11.360), (11.365), and (11.372),

$$\begin{aligned} F = {} & 1 + \frac{R_f}{R_s}\cdot\frac{1}{g_m^2}\left|\frac{1}{Z_L}+\frac{1}{R_f}\right|^2 \cdot \frac{|1+t|^2}{\left|1+\frac{R_f}{Z_L}\cdot\frac{p}{1+p}\right|^2} \\ & + \frac{\delta}{g_{d0}R_s}\left(\frac{\omega}{\omega_T}\right)^2 \frac{\left|1+\frac{R_f}{Z_L}\right|^2}{\left|1+p+p\frac{R_f}{Z_L}\right|^2}|1+t|^2 \\ & + \frac{1}{g_m^2}\cdot\frac{\gamma g_{d0}}{R_s}\cdot\frac{\left|1+\frac{Z_L}{R_f}\right|^2}{\left|1+\frac{1+p}{p+1/g_m R_f}\cdot\frac{Z_L}{R_f}\right|^2}|1+t|^2 \\ & + \frac{2wC_{gs}}{R_s}\sqrt{\frac{\delta\gamma}{5}}\mathrm{Re}\left[\frac{c}{\left[\left(\frac{1}{Z_L}+\frac{1+p}{R_f p+1/g_m}\right)\left(\frac{p}{Z_L}+\frac{1+p}{R_f}\right)^*\right]}\right]\left|\frac{1}{Z_L}+\frac{1}{R_f}\right|^2|1+t|^2 \end{aligned} \tag{11.374}$$

Output Matching

The second stage of the shunt feedback amplifier provides the output matching and current to drive a load – typically 50-Ω – and, as can be seen in Figure 11.82, it is just a source follower. The output matching of the amplifier can thus be approximated based on the matching of the second stage.

The voltage gain for a source follower in CMOS can be derived as [20]

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{R_L g_m}{1+R_L g_m} \tag{11.375}$$

where R_L is the load resistance, typically 50-Ω, and g_m is the transconductance of the transistor M_3. Since its output impedance is roughly equal to $1/g_m$, when it is matched to R_L, we have

$$R_L g_m \simeq 1 \tag{11.376}$$

Consequently, this condition leads to matching at the output of the amplifier. Under this matching condition, the voltage gain of the second stage from (11.375) drops to

$$\frac{v_{\text{out}}}{v_{\text{in}}} \simeq \frac{1}{2} \tag{11.377}$$

or −6 dB, which actually degrades the overall gain of the amplifier.

11.6.3.2 Design Example. A shunt feedback amplifier, whose schematic is shown in Figure 11.82, operating from 3.1 to 10.6 GHz was designed and fabricated in a 0.18-μm CMOS process [18]. The gate width of the transistor M_1 was chosen as 160 μm for optimum gain, noise figure and power consumption. The resistance of the feedback resistor R_f was selected as 400 Ω considering trade-off between gain and bandwidth. To achieve both matching and gain over the ultrawide bandwidth of 3.1–10.6 GHz, the load impedance was formed with an RLC parallel network. For the gain, bandwidth and noise consideration, R_L was set equal to R_f as 400 Ω. The inductance L_d was chosen as 4.5 nH to achieve resonance with the load capacitance of the first stage, which is the parasitic capacitance at the drain of the transistor M_2. The input matching of the amplifier is completed with a choice of 1.0 nH for the gate inductor L_g. Since the Q of the gate inductor is always critical due to its large impact on the overall noise performance of the amplifier, a high Q inductor was designed and its S-parameters calculated using the EM simulator IE3D were imported into Cadence SpectreRF [21] for the simulation of the entire amplifier. The cascode transistor M_2 was chosen to be identical with M_1 in dimensions. Its gate was biased at V_{dd}, which is 1.8 V. The source follower is formed by the transistors M_3 and M_4 having gate width of 80 μm for compromise between gain and output matching. The g_m of these transistors was biased by the DC biasing voltage at the gate of M_4, which is around 0.9 V. The entire amplifier uses two inductors, the gate inductor L_g and the load inductor L_d. L_g with an inductance of 1.0 nH could be replaced with a bonding wire. For this design, however, both L_g and L_d were implemented as on chip inductors. The entire shunt feedback amplifier consumes a chip area of 0.75×0.6 mm^2 including RF and DC pads.

The designed shunt feedback amplifier was simulated using Cadence SpectreRF. The parasitic effects were extracted from the layout and taken into account in the post-layout simulation. The measurement of the amplifier was done with on-wafer RF probes. Transistor M_1 was biased externally using a bias-Tee at the input of the amplifier. The V_{dd} and the gate bias for M_4 were fed using two on-wafer DC probes.

Figure 11.89 shows the input reflection coefficient in both schematic and post-layout simulations. In the schematic simulation, the input reflection coefficient is lower than −10 dB from 3 to 11 GHz. The post-layout simulation, however, shows that this bandwidth is reduced to 3–10 GHz. The bandwidth reduction is mainly

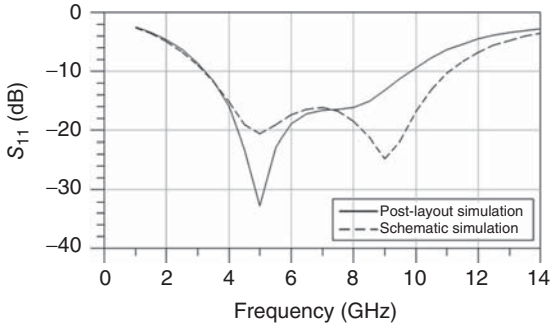


Figure 11.89. Post-layout simulation results of the input reflection coefficient (S_{11}) of the shunt feedback amplifier.

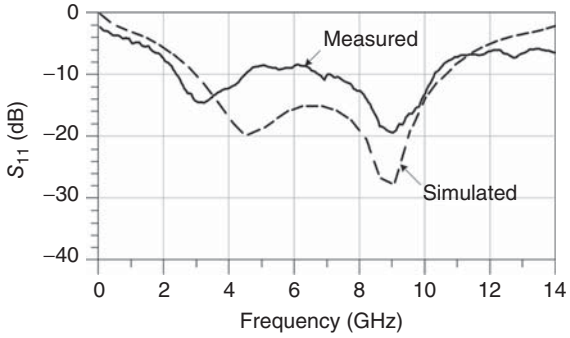


Figure 11.90. Measured and calculated results of the input reflection coefficient of the shunt feedback amplifier.

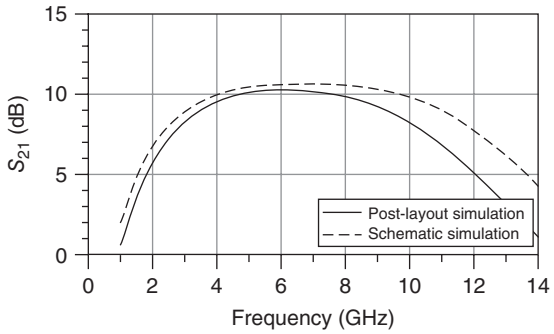


Figure 11.91. Post-layout simulation results of the gain (S_{21}) of the shunt feedback amplifier.

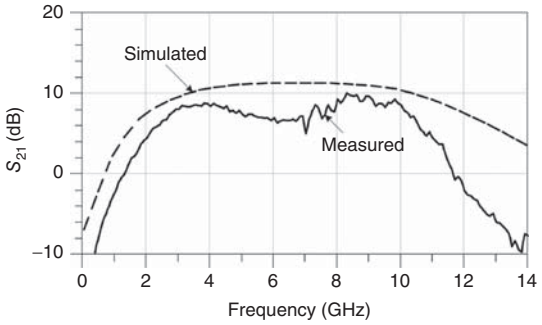


Figure 11.92. Measured and calculated results of the gain of the shunt feedback amplifier.

due to the parasitic capacitance of the input pad at the gate of M_1. One approach to alleviate this problem is slightly changing L_g and R_f to cancel the effects of the parasitic capacitance. Nevertheless, the input reflection coefficient in the post-layout simulation results is still below −9 dB over the 3–11 GHz bandwidth, indicating an acceptable input matching across the operating frequency range. The simulated and measured input reflection coefficients are plotted in Figure 11.90. The measurement results, while showing similar trend with those simulated, have some discrepancy with the calculated performance – while the input reflection coefficient is still lower than −10 dB across 2.5–10.3 GHz, that between 5 and 6.3 GHz degrades to around −8 dB. This discrepancy is mainly due to the parasitic capacitance of the input RF pads and that at the gate of M_1, which is typically expected in CMOS RFIC design.

The gain of the designed feedback amplifier is plotted in Figures 11.91 and 11.92. The amplifier exhibits a gain of 9.5–11 dB from 3 to 11 GHz in the schematic simulation, 8–10.5 dB from 3 to 10 GHz in the post-layout simulation, and 8–10 dB between 3 and 10 GHz in the measurement. It drains only 5-mA current from a 1.8-V voltage supply for the amplifier core of M_1 and M_2. It is noted that both the gain and

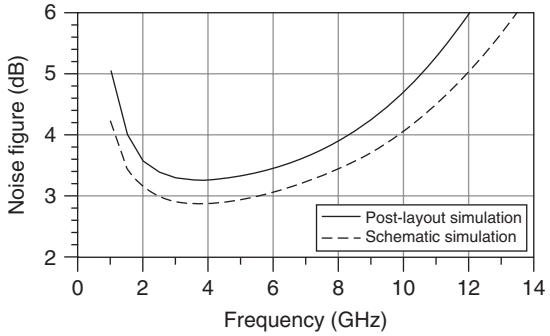


Figure 11.93. Post-layout simulation results for the noise figure of the shunt feedback amplifier.

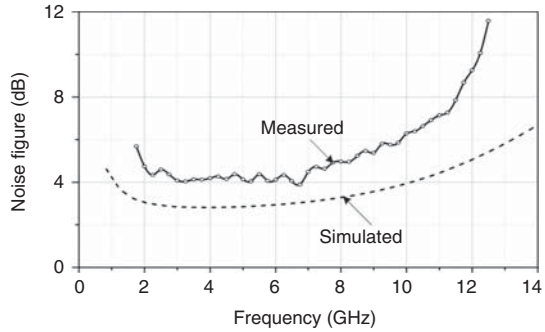


Figure 11.94. Measured and calculated results of the noise figure of the shunt feedback amplifier.

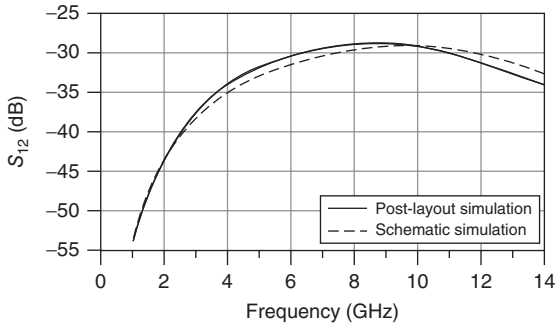


Figure 11.95. Post-layout and schematic simulation results of the isolation (S_{12}) of the shunt feedback amplifier.

bandwidth could be improved by increasing the biasing voltage at the gate of M_1 but at an expense of the power consumption. The measurement also shows that the amplifier provides a gain of 9–11 dB from 3 to 12 GHz with a current consumption of 12 mA.

Figures 11.93 and 11.94 show the noise performance of the feedback amplifier. In the schematic simulation result, the noise figure drops to as low as 2.7 dB at around 3.5 GHz, keeps going up as frequency goes higher, and eventually passes 4 dB at 10 GHz. The post-layout simulation measurement results indicate roughly 0.5 and 1–2 dB higher than those obtained from the schematic simulation, respectively. Similar to gain, the noise performance can also be improved at the cost of the power consumption. When draining a current of 12 mA from M_1, the amplifier shows measured noise figure of only 3.5–5.5 dB across 3–10 GHz.

The (reverse) isolation is an important part of the feedback amplifiers, as poor isolation may lead to oscillation. The potentially poor isolation and tendency of oscillation (or stability) are the two main drawbacks of typical feedback amplifiers as mentioned earlier. With careful design, however, good isolation and high stability can be achieved as demonstrated in the designed shunt feedback amplifier. Figures 11.95 and 11.96

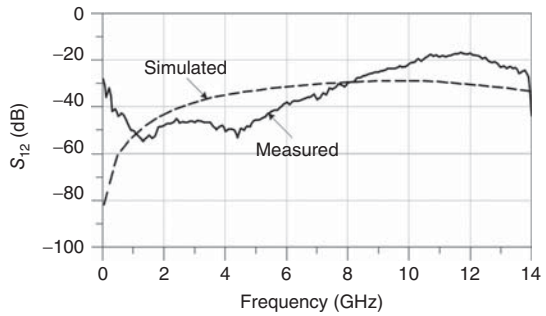


Figure 11.96. Measured and calculated results of the isolation of the shunt feedback amplifier.

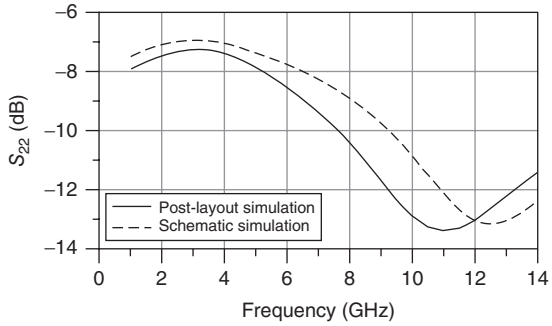


Figure 11.97. Post-layout simulation results of the output reflection coefficient (S_{22}) of the shunt feedback amplifier.

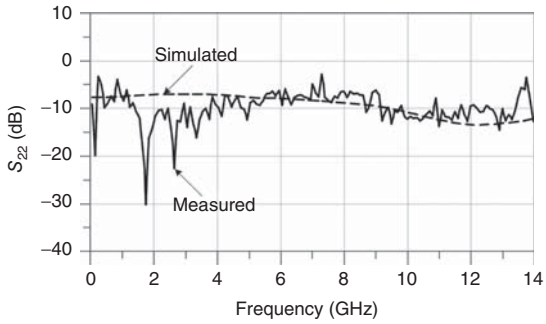


Figure 11.98. Measured and calculated results of the output reflection coefficient of the shunt feedback amplifier.

show the isolation of the amplifier. As can be seen, high isolation from 30 to 50 dB is observed in both the schematic and post-layout simulations and between 19 and 55 dB in the measurement across 1–11 GHz, suggesting that the shunt feedback technique does not degrade the amplifier's isolation if the design was carefully made.

The output matching of the shunt feedback amplifier is shown in Figures 11.97 and 11.98. Both the schematic and post-layout simulation results show the output reflection coefficient lower than −7 dB. Although some fluctuations are observed in the measurement results, it still matches reasonably well with the calculated performance, which is lower than −8 dB from DC-14 GHz. As the output matching is mainly determined by the bias current in the transistors M_3 and M_4, 5 mA was chosen for trade-off between the gain and output matching.

Stability is a very important issue for amplifier designs, especially for feedback amplifiers. The calculated stability factor K of the shunt feedback amplifier is plotted in Figure 11.99. Both the schematic and post-layout

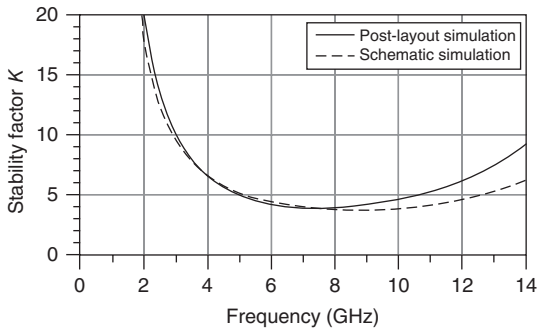


Figure 11.99. Stability factor K of the shunt feedback amplifier.

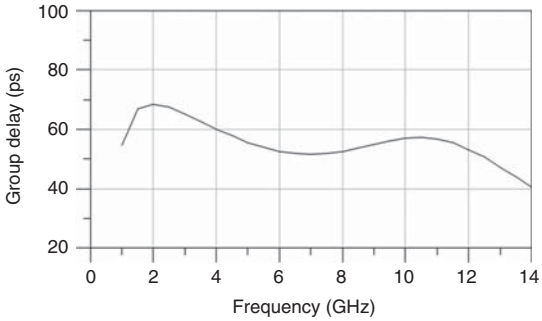


Figure 11.100. Group delay of the shunt feedback amplifier.

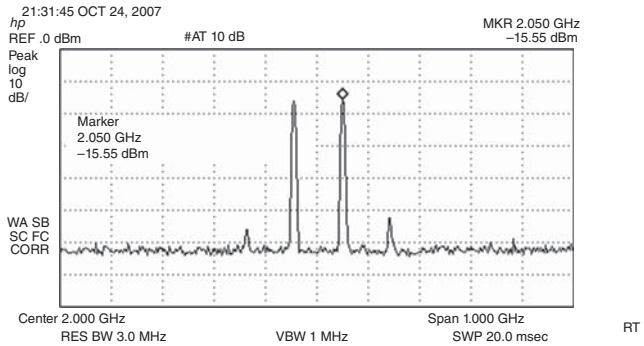


Figure 11.101. Spectrum of the output of the shunt feedback amplifier in a two-tone test.

simulation results show the stability factor $K > 3.8$ from DC-14 GHz, demonstrating that highly stable feedback amplifiers can be realized with careful design.

The group delay of the designed shunt feedback amplifier was also calculated from the simulated gain (S_{21}) results. Figure 11.100 shows a group delay of 55 ps with only ±10-ps fluctuation from 3 to 12 GHz.

Linearity is another important specification for amplifier. To measure the input third-order intercept point (IIP3) of the designed shunt feedback amplifier, a two-tone test was performed. Two fundamental tones with 100 MHz adjacent to each other were input to the amplifier through a power combiner. Figure 11.101 shows the output spectrum of the amplifier with two input tones at 1.95 and 2.05 GHz. The measurement was made at nine frequency points between 2 and 10 GHz. Figure 11.102 shows the measured IIP3 of the amplifier. As can be seen, the designed amplifier has a good linearity with the IIP3 varying from −2.1 to −11.4 dBm across 3–11 GHz.

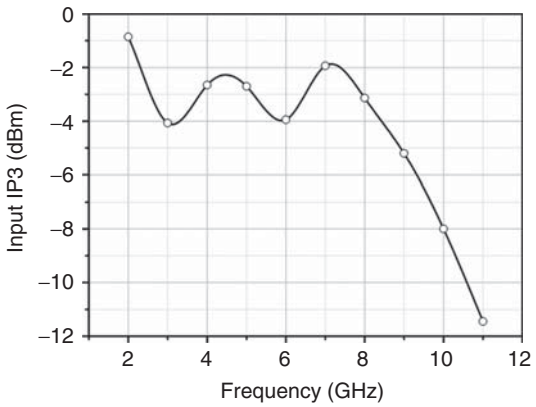


Figure 11.102. Measured input third-order intercept point (IIP3) of the shunt feedback amplifier.

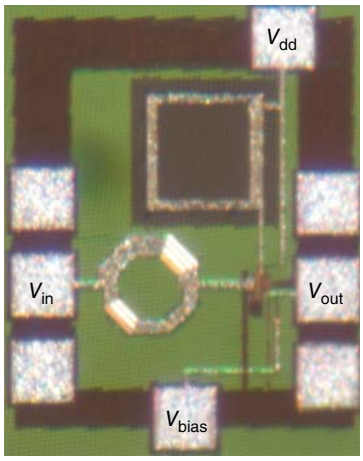


Figure 11.103. Microphotograph of the shunt feedback amplifier.

Figure 11.103 shows a microphotograph of the designed 0.18-μm CMOS shunt feedback amplifier. The die size of the amplifier is 0.75×0.6 mm^2 including all RF and DC pads, or 0.55×0.4 mm^2 without pads, signifying its miniaturization suitable for low cost CMOS applications.

11.6.4 Cascoded Common-Source Amplifiers

Cascoded common-source inductively degenerated amplifier topologies are often used for both narrow and wideband amplifiers to achieve high gain and low noise figure. Conventional cascoded amplifiers are based on narrow-band design. The bandwidth of cascoded amplifiers, however, can be extended significantly by incorporating proper wideband filter structures into their IMNs, resulting in not only good match and high gain, but also noise figure close to the device minimum noise figure, across wide frequency ranges. Such wideband cascoded amplifiers, however, have larger size than conventional cascoded narrow-band amplifiers due to the large die sizes typically occupied by the on-chip inductors used in the filters.

11.6.4.1 Analysis. Figure 11.104 shows the cascoded common-source inductively degenerated amplifier with a wideband IMN. The topology is based on the cascoded inductively degenerated common-source LNA [2, 22] and the analysis follow these.

For the narrow-band amplifier, that is, the amplifier shown in Figure 11.104 without the wideband matching network, the cascoded configuration formed by transistors M_1 and M_2 reduces the Miller-effect and improves the amplifier's (reverse) isolation as well as its frequency response. M_1 functions as the amplifying transistor

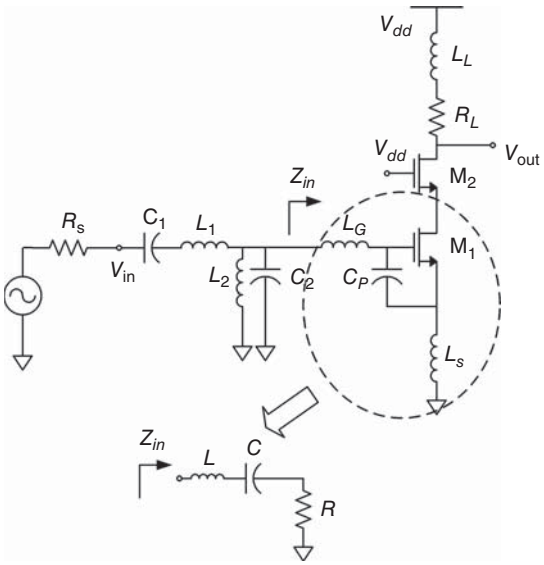


Figure 11.104. Wideband cascoded common-source inductively degenerated amplifier.

providing primarily the gain for the amplifier. The enhanced isolation enables the effects of M_2, R_L and L_L to the amplifier's input impedance to be neglected and hence simplifying the analysis. The amplifier has a series inductor, L_G, at the gate of M_1 and a capacitor, C_P, in parallel with C_{gs} of M_1, making the amplifier design more flexible. The input impedance Z_{in} looking into the gate of M_1 with inductive source degeneration, through L_G, as shown in Figure 11.104, can be written as

$$Z_{\text{in}} = \frac{1}{j\omega(C_{gs} + C_P)} + j\omega(L_s + L_G) + \omega_T L_s \tag{11.378}$$

where $\omega_T = g_m/(C_{gs} + C_P) = g_m/C_T$ is the cut-off frequency of M_1. Equation (11.378) shows that this impedance can be modeled as a series RLC network, as seen in Figure 11.104, where $R = \omega_T L_s$, $L = L_G + L_s$, and $C = C_{gs} + C_P$. In the design of the narrow-band amplifier, the real part of Z_{in} is chosen to be equal to the source resistance R_s while its reactive part is forced to resonate with nearly optimal noise figure at operating frequency.

In the wideband cascoded common-source inductively degenerated amplifier shown in Figure 11.104, the bandwidth of the narrow-band inductively degenerated cascoded amplifier is extended by using a third-order Chebyshev band-pass filter as a wideband IMN.[10] This band-pass filter is realized by a series resonator, consisting of inductor L_1 and capacitor C_1, a parallel resonator, consisting of inductor L_2 and capacitor C_2, and the series resonator RLC formed by the input impedance Z_{in} described earlier. The band-pass filter is shown

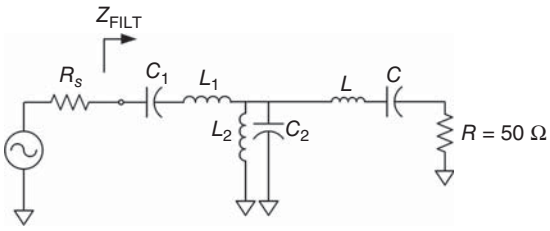


Figure 11.105. Third-order Chebyshev band-pass filter.

[10]A third-order Chebyshev band-pass filter is used as an example. Lower or higher order band-pass filter, either Chebyshev, Butterworth or other kinds of filters, can be employed.

TABLE 11.2. Component Values of the Third-Order Chebyshev Band-Pass Filter

L_1 (nH)	C_1 (pF)	L_2 (nH)	C_2 (pF)	L (nH)	C (pH)
1.65	0.47	1.64	0.47	1.65	0.47

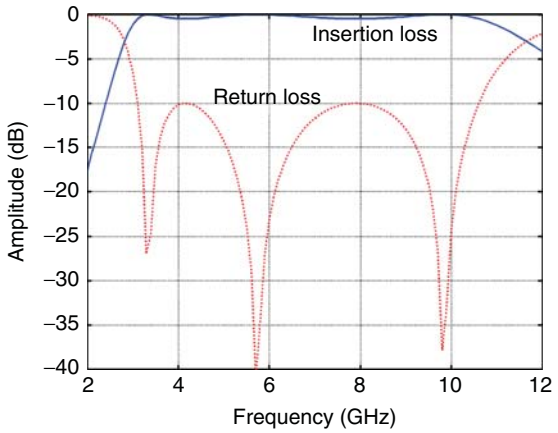


Figure 11.106. Calculated insertion loss and return loss of the third-order Chebyshev band-pass filter.

explicitly in Figure 11.105 with R_s and R representing the source and load impedances of the filter, respectively. The elements of the band-pass filter (L_1, C_1, L_2, C_2, L, and C) can be determined, using the band-pass filter analysis in Section 8.5.4, for given pass-band frequency range, pass-band ripple and source and load impedances. Table 11.2 shows the calculated values of these elements for pass-band frequencies from 3.1 to 10.6 GHz, pass-band ripple of 0.5 dB, and input and output impedances of 50 Ω. The simulated return loss and insertion loss of the band-pass filter are shown in Figure 11.106 demonstrating the desired pass-band from 3.1 to 10.6 GHz.

The shunt-peaking topology employing the series inductor L_L and resistor R_L as the load, as seen in Figure 11.104, facilitates flat gain over the designed frequency range. The value of L_L should be large enough to provide a large gain at the upper frequency. This value, however, must also be sufficiently small so that the resonant frequency of the resonator formed by L_L and C_{out}, where C_{out} is the total capacitance between the drain of M_2 and the ground, is much higher than the highest operating frequency. The value of R_L is such that the zero-frequency $\omega_Z = R_L/L_L$ is close to the lower frequency edge of the band to improve the gain at lower frequencies. R_L, however, is limited by an upper value above which the voltage drop is such large that reduces the voltage V_{dd} supplied to the drain of M_2.

Gain

As a desired design criterion for filters and can be seen in Figure 11.106, the Chebyshev band-pass filter or the input network of the cascoded common-source inductively degenerated amplifier should have as high return loss and low insertion loss as possible, leading to a valid assumption of unity in the pass-band for the filter's transfer function. The input impedance looking into the filter in Figure 11.105 or into the amplifier's input port in Figure 11.104 can therefore be considered R_s over the pass-band. The current flowing into the transistor M_1 can thus be written as $i_{\text{in}} = v_{\text{in}}/R_s$. In addition, M_1 functions as a current amplifier at high frequency with a current gain of $\beta = g_m/(j\omega C_T)$.

The overall output load of the amplifier is a parallel combination of the shunt-peaking load of R_L, L_L, and the capacitor C_{out}, and can thus be expressed as

$$Z_{\text{LOAD}} = \frac{R_L + j\omega L_L}{1 + j\omega C_{\text{out}}(R_L + j\omega L_L)} \tag{11.379}$$

The voltage gain of the amplifier can now be written using (11.379) and the current gain expression $\beta = g_m/(j\omega C_T)$ as

$$\frac{v_{\text{out}}}{v_{\text{in}}} = -\beta\frac{Z_{\text{LOAD}}}{R_s} = -\frac{g_m}{j\omega C_T R_s}\cdot\frac{R_L + j\omega L_L}{1 + j\omega C_{\text{out}}(R_L + j\omega L_L)} \tag{11.380}$$

Equation (11.380) clearly shows that, at lower frequencies, R_L plays an important role in determining the voltage gain while, at higher frequencies, the current gain roll-off is compensated by the load inductor L_L. It is particularly noted that possible spurious resonance caused by C_{out} and L_L is a concern in this amplifier design. To avoid potential adverse effects to the gain and input matching, proper values for C_{out} and L_L need to be used to keep this resonance out of the amplifier's pass-band.

Noise Figure

The noise performance of the cascoded common-source inductively degenerated amplifier can be analyzed considering two major noise contributors – one is the loss associated with the input passive network (Chebyshev band-pass filter) and another is the noise generated by the transistor M_1. For the input network, as shown in Figure 11.105, in which on-chip metal–insulator–metal (MIM) capacitors and spiral inductors are used, the MIM capacitors have much higher quality factor (Q) than those of the spiral inductors. The noise contribution of the band-pass filter is, therefore, dominated by the on-chip spiral inductors. High Q inductors are thus critical in order to reduce its noise contribution. Since the Q of these passive components dictate their noise contribution, the noise analysis of these passive components is implicitly understood. The noise contribution from M_1, for specific bias currents, can be minimized by selecting proper transistor width. The analysis of the noise performance of amplifiers operating over broad frequency ranges is particularly challenging. Both minimum noise figure and average in-band noise figure should be investigated in order to achieve an optimum noise performance across the entire bandwidth.

The analysis of the noise contributed by transistor M_1 follows [2, 22]. It is based on two-port topology with input-referred noise current and voltage sources. In the analysis, the low frequency $1/f$ noise is ignored due to the amplifier's high operating frequency and the assumption that the up-converted noise to RF frequency through mixing of the $1/f$ noise and the RF signal has negligible effect to the amplifier's noise figure. Figure 11.107(a) shows the metal oxide semiconductor (MOS) transistor noise sources with the loading effect of the feedback inductor L_s. i_{ng} represents the induced gate noise and i_{nd} is the drain noise current, whose PSD are given in (11.331) and (11.333), respectively.

Employing the conventional input-referred noise scheme, the current noise sources of M_1 are replaced with the two correlated noise generators of i_n and v_n as shown in Figure 11.107(b). As can be seen in Figure 11.107(b), when the input is short-ended, only noise generator i_n exists. As the transistor can be considered a current amplifier, the noise source i_{nd} and the input-referred noise current $i_{nd,\text{input}}$ are related by

$$\frac{i_{nd}}{\sqrt{\Delta f}} = \frac{g_m}{j\omega C_T}\cdot\frac{i_{nd,\text{input}}}{\sqrt{\Delta f}} \tag{11.381}$$

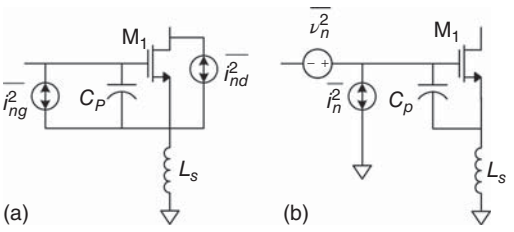


Figure 11.107. Transistor M_1 with current noise sources (a) and input-referred equivalent noise generators (b) used for the noise analysis.

where $1/j\omega C_T$ is the impedance of the output parasitic capacitor, from which we obtain

$$\frac{i_{nd,\text{input}}}{\sqrt{\Delta f}} = \frac{j\omega C_T}{g_m} \cdot \frac{i_{nd}}{\sqrt{\Delta f}} \tag{11.382}$$

The total noise current generator i_n can then be derived from (11.381) and (11.382) as

$$\frac{i_n}{\sqrt{\Delta f}} = \frac{i_{ng}}{\sqrt{\Delta f}} + \frac{i_{nd,\text{input}}}{\sqrt{\Delta f}} = \frac{i_{ng}}{\sqrt{\Delta f}} + \frac{j\omega C_T}{g_m}\frac{i_{nd}}{\sqrt{\Delta f}} \tag{11.383}$$

When the input is open-ended, the noise generator v_n can be expressed, from Figure 11.107(b), as:

$$\frac{v_n}{\sqrt{\Delta f}} = \frac{v_{n,1}}{\sqrt{\Delta f}} + \frac{v_{n,2}}{\sqrt{\Delta f}} \tag{11.384}$$

where $v_{n,1}$ is the input-referred noise from transistor M_1 and $v_{n,2}$ is that from L_s caused by i_n, which are expressed respectively as

$$\frac{v_{n,1}}{\sqrt{\Delta f}} = \frac{i_{nd}}{g_m\sqrt{\Delta f}} \tag{11.385}$$

and

$$\frac{v_{n,2}}{\sqrt{\Delta f}} = j\omega L_s \frac{i_n}{\sqrt{\Delta f}} \tag{11.386}$$

The equivalent noise voltage generator can be obtained from (11.384) to (11.386) as

$$\frac{v_n}{\sqrt{\Delta f}} = \frac{i_{nd}}{g_m\sqrt{\Delta f}} + j\omega L_s \frac{i_n}{\sqrt{\Delta f}} \tag{11.387}$$

Normally, the input-referred noise voltage source v_n is partially correlated with the input-referred noise current source i_n. Hence the PSD of v_n can be expressed as the sum of two PSD components: one fully correlated, $\overline{v_{n,c}^2}$, and the other, $\overline{v_{n,u}^2}$, uncorrelated to the noise current source i_n as

$$\frac{\overline{v_n^2}}{\Delta f} = \frac{\overline{v_{n,c}^2}}{\Delta f} + \frac{\overline{v_{n,u}^2}}{\Delta f} \tag{11.388}$$

and the corresponding correlation impedance Z_c can be expressed as

$$Z_c = \sqrt{\frac{\overline{v_{n,c}^2}}{\overline{i_n^2}}} = j\omega L_s + \frac{1}{j\omega C_T} \cdot \frac{1 + |c|p\alpha\chi}{1 + 2|c|p\alpha\chi + (p\alpha\chi)^2} \tag{11.389}$$

where the correlation coefficient c between the gate noise and drain noise is defined in (11.334) and

$$p = \frac{C_{gs}}{C_T} \tag{11.390}$$

$$\chi = \sqrt{\frac{\delta}{5\gamma}} \tag{11.391}$$

$$\alpha = \frac{g_m}{g_{d0}} \tag{11.392}$$

with g_{do} being the channel conductance at $V_{ds} = 0$. α represents the short-channel effect and is used to estimate the reduction in transconductance due to velocity saturation and mobility decrease for vertical fields.

The other two important parameters in noise figure analysis are the uncorrelated noise sources i_n and $v_{n,u}$. The corresponding equivalent noise conductance and resistance can be expressed by

$$G_n = \frac{\frac{\overline{i_n^2}}{\Delta f}}{4KT} = \frac{\gamma}{\alpha^2 g_{d0}}(\omega C_T)^2[1 + 2|c|p\alpha\chi + (p\alpha\chi)^2] \tag{11.393}$$

$$R_u = \frac{\frac{\overline{v_{n,u}^2}}{\Delta f}}{4KT} = \frac{\gamma}{\alpha^2 g_{d0}} \cdot \frac{(p\alpha\chi)^2(1 - |c|^2)}{1 + 2|c|p\alpha\chi + (p\alpha\chi)^2} \tag{11.394}$$

respectively.

Following the conventional two-port system noise analysis, the noise figure of the amplifier can be derived as

$$F = 1 + \frac{R_u + |Z_c + Z_S|^2 G_n}{R_s} \tag{11.395}$$

where $Z_S = R_s + jX_S$ is the source impedance of the amplifier. A minimum noise figure is achieved when the amplifier is noise-matched at its input and the corresponding source impedance Z_S is equal to $Z_{\text{opt}} = R_{\text{opt}} + jX_{\text{opt}}$ where

$$R_{\text{opt}} = \sqrt{\frac{R_u}{G_n} + R_c^2} = \frac{p\alpha\chi\sqrt{1 - |c|^2}}{\omega C_T[1 + 2|c|p\alpha\chi + (p\alpha\chi)^2]} \tag{11.396}$$

and

$$X_{\text{opt}} = -X_c = -\omega L_s + \frac{1}{\omega C_T} \cdot \frac{1 + |c|p\alpha\chi}{1 + 2|c|p\alpha\chi + (p\alpha\chi)^2} \tag{11.397}$$

From (11.334) and (11.390)–(11.392), $|c| = 0.395$, $p < 1$, $\alpha \le 1$, and $\chi < 1$, making the factor $(1 + |c|p\alpha\chi)/(1 + 2|c|p\alpha\chi + (p\alpha\chi)^2)$ of $1/\omega C_T$ in (11.397) close to one. The optimum source impedance can therefore approximately be achieved if the series combination of C_T and L_s is resonated across the interested frequency band. With the help of the three-section Chebyshev band-pass filter input network that provides a wideband input matching for the amplifier, the overall input reactance looking into the filter is resonated over a wide bandwidth, resulting in $X_{\text{opt}} = 0$. A quasi-minimum noise figure can hence be realized over the entire amplifier bandwidth. Correspondingly, the noise figure can be simplified from (11.395) as

$$F \simeq 1 + \frac{R_u}{R_s} + G_n R_s \tag{11.398}$$

Substituting (11.393)–(11.397) into (11.398) yields the following expression for the noise figure:

$$F \simeq 1 + \frac{\gamma}{\alpha g_m R_s}\left\{\frac{(p\alpha\chi)^2(1 - |c|^2)}{1 + 2|c|p\alpha\chi + (p\alpha\chi)^2} + (\omega C_T R_s)^2[1 + 2|c|p\alpha\chi + (p\alpha\chi)^2]\right\} \tag{11.399}$$

Equation (4.18) shows that larger transconductance g_m produces better noise figure. For given g_m, transistor with smaller W/L results in larger α and hence better noise figure as seen in (11.399). However, to maintain a fixed value for g_m with smaller transistor, larger I_D is needed according to Eq. (9.10). Therefore, large bias current is preferred for noise performance.

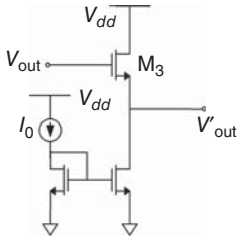


Figure 11.108. Source-follower buffer for the amplifier.

TABLE 11.3. Final Optimized Lumped-Element Values of the Amplifier

L_1 (nH)	C_1 (pF)	L_2 (nH)	C_2 (pF)	L_G (nH)	C_P (pF)	L_s (nH)	L_L (nH)	R_L (Ω)
1.08	0.65	1.63	0.45	1.32	0.06	0.61	2.65	85

As mentioned earlier, the average noise figure over the entire operating frequency band is an important parameter to evaluate because of the wide bandwidth of the amplifier. For a specific bias current I_{bias}, there is a range for the width of the transistor M_1 that can be chosen to achieve minimum average noise figure. It is recognized that for given on-chip inductors with certain Q, the noise performance of the considered cascoded amplifier is mainly limited by the noise contribution of M_1. Therefore, for broadband amplifier design, better noise performance can be achieved if larger bias current is applied.

Design and Performance

The cascoded common-source inductively degenerated amplifier was designed as a LNA operating from 3.1 to 10.6 GHz using Jazz 0.18-μm RFCMOS process [18]. A bias current $I_{\text{bias}} = 5$ mA was assumed and a minimum length of 0.18 μm was selected for both transistors M_1 and M_2. Considering the balance between the thermal drain noise and induced gate noise, the size of M_1 was selected as 260 μm based on (11.399). For M_2, a smaller size is preferred in order to reduce the parasitic capacitances. A smaller transistor size, however, produces higher noise, leading to the need of setting a minimum width for M_2 according to its maximum tolerable noise contribution. In the designed amplifier, a width of 60 μm was selected for M_2. On-chip inductors and capacitors are used for the amplifier.

To facilitate on-wafer measurements, a typical source-follower buffer is also included in the amplifier to drive an external 50-Ω load. The buffer is shown in Figure 11.108 and the 50-Ω load is connected to the source of M_3. The buffer consists of transistor M_3 and a current mirror to provide an independent biased current source I_{bias} of 5 mA. The length and width of the two transistors in the current mirror were optimized to produce high output impedance. The size of M_3 was selected as 60 μm to achieve a transconductance of $g_{m3} \simeq 1/50\,\text{S}$. As the output voltage of the buffer is only half of that produced by the amplifier without buffer, the gain of the complete amplifier with buffer is 6 dB lower than that of the amplifier core itself.

In the initial design, the gate–drain capacitance C_{gd} of M_1 was omitted to simplify the design. This capacitance, however, was included in the circuit simulations through which the values of on-chip inductors and capacitors were determined through optimization. Table 11.3 shows the values of the final optimized components. The capacitors are implemented using MIM capacitors available in the design kit. The inductors are octagonal spiral inductors with PGS, particularly designed and optimized for high Q and desired inductance values across the desired operating frequency range.

As indicated earlier, the Q of the inductors used in the input network (i.e., the band-pass filter) plays an important role in reducing the noise figure of the entire amplifier. The top-most metal layer M6 with the thickest metallization was selected for the octagonal spiral inductors due to its smallest resistance among the metal layers. The bottom-most metal layer M1 was used for the PGS. IE3D was used to optimize the parameters of all the spiral inductors, including inductors L_1, L_2, L_G, L_s, and L_L, used in the amplifier as well as

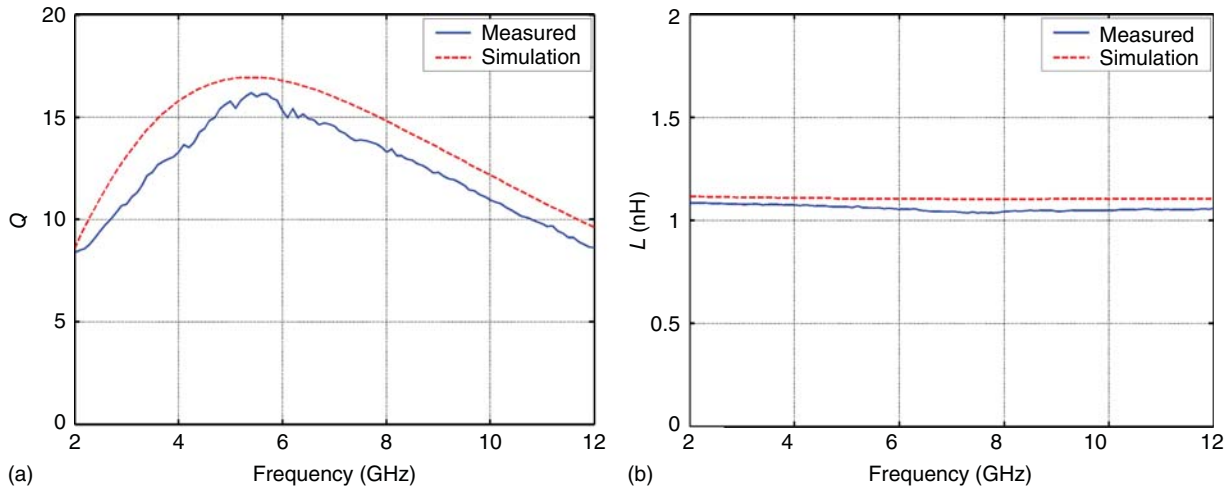


Figure 11.109. Performance of the patterned ground shield inductor L_1: (a) quality factor Q and (b) inductance L.

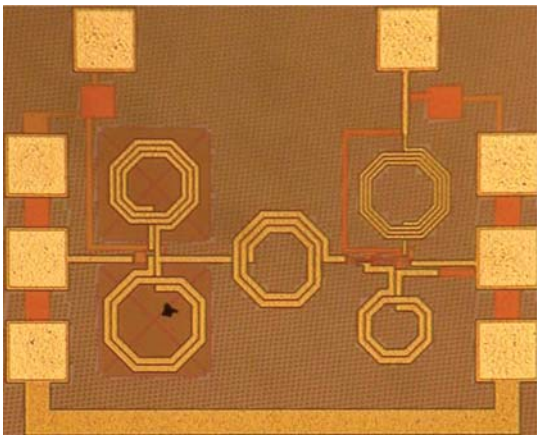


Figure 11.110. Photograph of the designed amplifier chip.

the patterned ground plane to achieve Q as high as possible and desired inductance values over the operating frequency band. These inductors were then fabricated separately on the same amplifier die for design verification via S-parameters measurement. Calibration components including open, short and through patterns along with RF metal pads were also fabricated and measured on the same wafer to de-embed the RF pad's effects from the measured results. Two-port S-parameters were measured on the fabricated inductors using a vector network analyzer and RF probe station over the frequency range of 2–12 GHz. The parameter extraction was performed by IE3D with the de-embedded S parameters to generate the measured Q and inductance values over the entire frequency band. All five inductors used in the amplifier were measured. Figure 11.109 shows the measured and simulated results for inductor L_1. As can be seen, the Q reaches a maximum value of 16 at around 5.5 GHz and is larger than 10 over the frequency range from 2.7 to 10.7 GHz. From Figure 11.109(b), it is also clear that the inductance is almost independent to frequency variation within 2–12 GHz. As the Q of these inductors dominates the noise performance of the amplifier, these relatively good Q values promise possibly good noise figure for the amplifier.

Figure 11.110 shows a photograph of the designed amplifier with an overall size of 0.88 mm $\times$ 0.7 mm, including on-wafer RF and DC-bias pads. On-wafer measurement was performed for the amplifier. Figure 11.111 shows the measured and calculated return losses at the input port, showing reasonably well agreement between them. From 2.9 to 12 GHz, the measured return loss is below −10 dB, which validates the IMN design. Figure 11.112 shows the measured and simulated return losses at the output port which

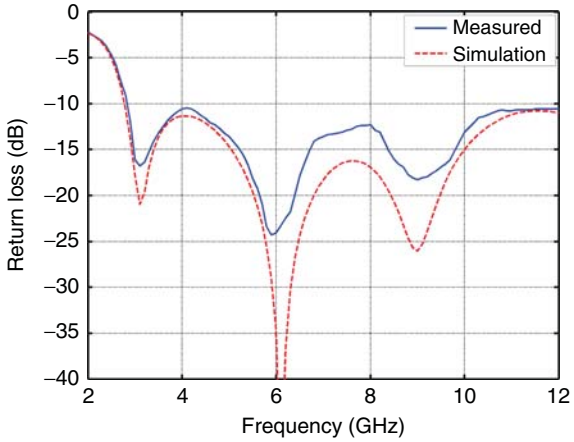


Figure 11.111. Input return loss of the designed amplifier.

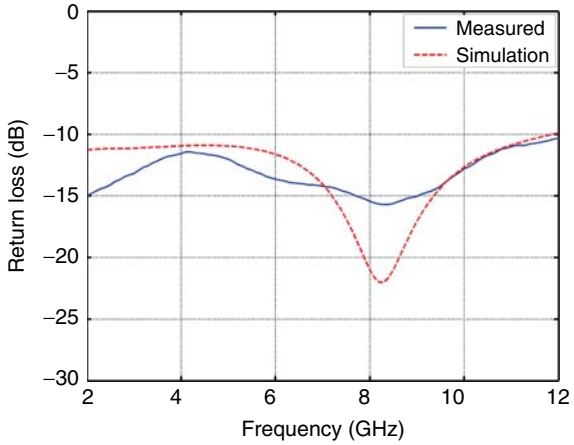


Figure 11.112. Output return loss of the designed amplifier.

also closely match each other. The output return loss is below −10 dB across 2–12 GHz which is reasonably good that validates the output buffer design. Figure 11.113 shows the measured and simulated results for the (reverse) isolation of the designed amplifier. The measured isolation is better than 40 dB across 2–12 GHz, which proves the effectiveness of the amplifier, particularly for system integration. Figure 11.114 presents the gain performance of the amplifier including buffer stage. A maximum gain of 12.4 dB is achieved. Across the 3-dB bandwidth of 2.6–9.8 GHz, the amplifier exhibits a minimum gain of 9.4 dB. Furthermore, the gain ripple is relatively small over the whole 3-dB bandwidth. The difference between the measured and simulated results at the high frequency end is caused by the extra parasitic capacitance from the output buffer, which was not fully considered during the simulation, leading to only partial compensation of the total parasitic capacitance by the inductor load. Figure 11.115 shows the noise performance of the designed amplifier with the measured noise figure matching reasonably well to that calculated. The amplifier exhibits a minimum noise figure of 4 dB at 5.2 GHz.

11.7 CURRENT MIRRORS

Current mirrors are circuits that replicate a DC current through their output, keeping the current constant regardless of the voltage applied (DC or transient) at the output node. They are thus essentially current sources. Current mirrors are widely used in many RFICs. This section presents some basics of current mirrors.

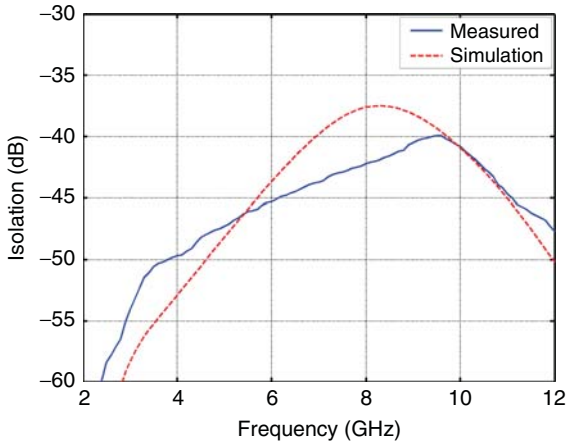


Figure 11.113. Isolation of the designed amplifier.

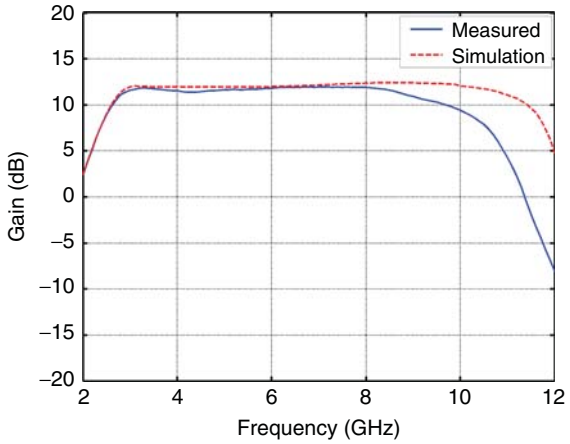


Figure 11.114. Power gain of the designed amplifier including buffer.

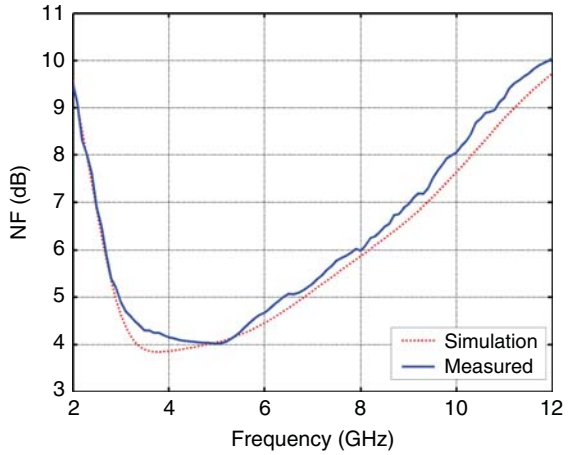


Figure 11.115. Noise figure of the designed amplifier.

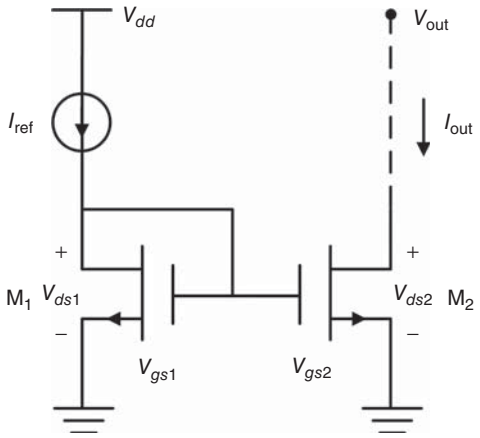


Figure 11.116. A simple current mirror.

11.7.1 Basic Current Mirror

Figure 11.116 shows a simple CMOS current mirror consisting of two transistors (M_1 and M_2). Assuming both transistors are in the active region, we can write:

$$I_{\text{ref}} = \frac{1}{2}\mu_n C_{\text{ox}} \left(\frac{W_1}{L_1}\right)(V_{gs1} - V_{T1})^2(1 + \lambda_1 V_{ds1}) \tag{11.400}$$

$$I_{\text{out}} = \frac{1}{2}\mu_n C_{\text{ox}} \left(\frac{W_2}{L_2}\right)(V_{gs2} - V_{T2})^2(1 + \lambda_2 V_{ds2}) \tag{11.401}$$

where μ_n is the electron mobility, W_i, L_i, λ_i, and V_{Ti} ($i = 1, 2$) are the gate width, gate length, channel-length modulation, and threshold voltage of transistor i, respectively, and C_{ox} is the (oxide) capacitance per unit area between the gate electrode and substrate. The basic principle of the circuit is to draw a current I_{ref} through the reference transistor M_1 that sets up a corresponding voltage V_{gs1}. Following I_{ref}, an output current I_{out} would flow through transistor M_2, which can be controlled by the ratio of the device dimensions as seen in (11.401).

Assuming $V_{gs1} = V_{gs2}$ and $V_{T1} = V_{T2}$, we can obtain from (11.400) and (11.401):

$$\frac{I_{\text{out}}}{I_{\text{ref}}} = \frac{W_2/L_2}{W_1/L_1}\frac{1 + \lambda_1 V_{ds2}}{1 + \lambda_2 V_{ds1}} \tag{11.402}$$

which shows that, even when the transistors are identical ($W_1 = W_2$ and $L_1 = L_2$), current mismatch would occur due to the differences between V_{ds1}, V_{ds2}, and λ_1, λ_2. To reduce the effect of the drain–source-voltage difference, a cascode current mirror as described in the following section, can be used.

11.7.2 Cascode Current Mirror

Figure 11.117(a) shows a cascode current mirror that can reduce the current mismatch caused by the difference between V_{ds1} and V_{ds2}. Transistor M_3 shields the voltage V_Y at node Y from variations in V_{out}. The bias voltage V_b for M_3 is set so that the voltage V_X at node X is equal to V_Y and the output current I_{out} closely follows the reference current I_{ref}. Figure 11.117(b) shows a cascode current mirror with transistor M_4 placed on top of M_1 to provide bias V_b to M_3. The minimum output voltage $V_{\text{out,min}}$ needed for keeping M_2 and M_3 in the saturation region is given as

$$V_{\text{out,min}} = V_N - V_T = V_{gs1} + V_{gs4} - V_T = V_{gs2} + V_{gs3} - V_T \tag{11.403}$$

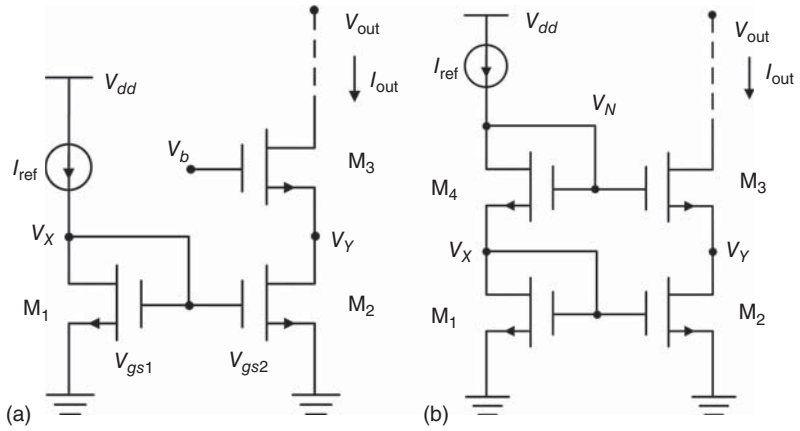


Figure 11.117. Cascode current mirror without (a) and with (b) biasing transistor.

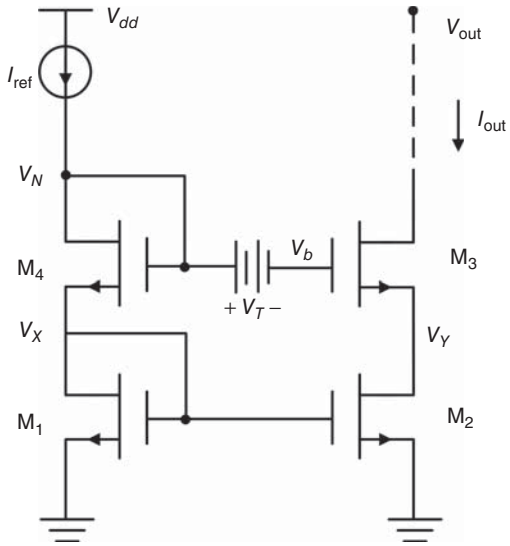


Figure 11.118. Improved cascode current mirror for high output swing.

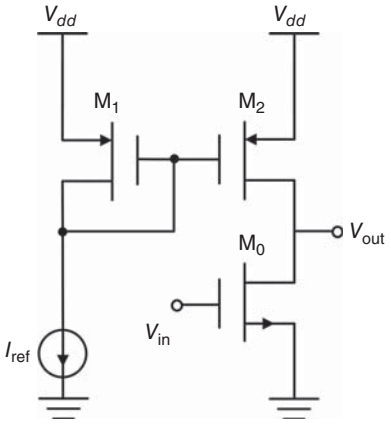


Figure 11.119. A common-source amplifier using a current mirror as an active load.

where V_N is the voltage at node N, V_{gsi}, $i = 1, 2, 3, 4$, is the gate–source voltage of transistor M_i, and all the transistors are assumed to have the same threshold voltage V_T. Equation (11.403) shows that this cascode current mirror reduces the output voltage headroom. Consider the right-side branch of the circuit in Figure 11.117(b), for M_2 and M_3 to be in the saturation region, their respective drain-source voltages V_{ds2} and V_{ds3} should be $V_{ds2} \geq V_{gs2} - V_T$ and $V_{ds3} \geq V_{gs3} - V_T$, which leads to

$$V_{\text{out,min}} = V_{ds2} + V_{ds3} \geq V_{gs2} + V_{gs3} - 2V_T \tag{11.404}$$

In order to increase the output voltage swing range, a current-mirror circuit as shown in Figure 11.118 can be used [3]. In this circuit, a voltage shift of V_T is implemented between the gates of M_3 and M_4. Under this condition, V_Y is not equal to V_X, but having $V_Y = V_{gs2} - V_T$ such that M_3 is biased at the boundary between the active and triode regions. The minimum output voltage can be obtained as

$$V_{\text{out,min}} \simeq V_{gs2} + V_{gs3} - 2V_T \tag{11.405}$$

Current mirrors are also widely used as active loads. Figure 11.119 shows an example of a current mirror used as an active load in common-source amplifiers. High impedance output can be obtained without consuming too much headroom, so the amplifier can achieve a large voltage gain.

REFERENCES

1. G. Gonzalez, *Microwave Transistor Amplifiers: Analysis and Design*, Second Edition, Upper Saddle River, New Jersey, Prentice Hall, 1966.
2. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, Cambridge, UK, 1998.
3. P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, Fourth Edition, John Wiley & Sons, Inc., New York, New York, 2001.
4. W. S. Percival, "Thermionic valve circuits," British Patent 460,562, Jan. 29, 1937.
5. E. L. Ginzton, W. R. Hewlett, J. H. Jasberg, and J. D. Noe, "Distributed Amplification," *Proc. IRE*, Vol. 36, No. 8, pp. 956–969, Aug. 1948.
6. C. S. Aitchison, "The Intrinsic Noise Figure of the MESFET Distributed Amplifier," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-33, No. 6, pp. 460–466, Jun. 1985.
7. Y. Ayasli, R. L. Mozzi, J. L. Vorhaus, L. D. Reynolds, and R. A. Pucel, "A Monolithic GaAs 1–13-GHz Traveling-Wave Amplifier," *IEEE Trans. Microwave. Theory Tech.*, Vol. MTT-30, No. 7, pp. 976–981, Jul. 1982.
8. A. S. Podgorski and L. Y. Wei, "Theory of Traveling-Wave Transistors," *IEEE Trans. Electron Devices*, Vol. 29, No. 12, pp. 1845–1854, Dec. 1982.
9. C. S. Aitchison, and M. Mbabele, "Third Order Intermodulation improvement in Distributed Amplifiers," in *Proc. Of 31st European Solid St. Conf. Dig.*, pp. 1–4, Oct. 2001.
10. X. Guan, Y. Jin, and C. Nguyen, "Design of High-Performance Compact CMOS Distributed Amplifiers with On-Chip Patterned-Ground-Shield Inductors," *IEE Electron. Lett.*, Vol. 45, No. 15, pp. 791–792, Jul. 2009.
11. X. Guan and C. Nguyen, "A 0.25-μm CMOS Ultra-Wideband Amplifier for Time-Domain UWB Applications," *2005 IEEE Radio Frequency Integrated Circuits Symposium*, Long Beach, California, Jun. 2005.
12. X. Guan and C. Nguyen, "Low-Power-Consumption and High-Gain CMOS Distributed Amplifiers Using Cascade of Inductively Coupled Common-Source Gain Cells for UWB Systems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-54, No. 8, pp. 3278–3283, Aug. 2006.
13. M. Chirala, X. Guan, and C. Nguyen, "A DC-20 GHz CMOS LNA Using Novel Multi-layered Transmission Lines and Inductors," *IEE Electron. Lett.*, Vol. 42, No. 22, pp. 1273–1275, Oct. 2006.
14. TSMC 0.18-μm CMOS Process, MOSIS Foundry, Marina del Rey, CA, USA.
15. M. K. Chirala, X. Guan, and C. Nguyen, "Integrated Multi-Layered On-Chip Inductors for Compact CMOS RFICs and Their Use in Miniature Distributed Low-Noise-Amplifier Design for Ultra-Wideband Applications," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-56, pp. 1783–1789, Aug. 2008.

16. H. Shigematsu, M. Sato, I. Hirose, F. Brewser, and M. Rodwell, "40Gb/s CMOS distributed amplifier for fiber-optic communication systems," in *Proc. Of IEEE Int. Solid-State circuits Conf.*, pp. 476–477, 2004.
17. H. M. Greenhouse, "Design of Planar Rectangular Microelectronic Inductors," *IEEE Trans. Parts, Hybrids, and Packag.*, Vol. PHP-10, pp. 101–109, Jun. 1974.
18. Jazz Semiconductor, 4321 Jamboree Road, Newport Beach, CA 92660.
19. H. S. Black, "Inventing the Negative Feedback Amplifier," *IEEE Spectr.*, Vol. 14, pp. 55–60, Dec. 1977.
20. B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, Boston, MA, 2001.
21. Cadence, Cadence Design Systems, Inc., San Jose, CA, USA.
22. A. Bevilacqua and A. M. Niknejad, "An Ultrawideband CMOS Low-Noise Amplifier for 3.1 — 10.6-GHz Wireless Receivers," *IEEE J. Solid-State Circuits*, Vol. 39, pp. 2259–2268, Dec. 2004.

PROBLEMS

11.1 Derive Eqs. (11.30)–(11.36).

11.2 Derive (11.46) and (11.47).

11.3 Derive (11.57).

11.4 Derive (11.58)–(11.61).

11.5 Derive (11.62).

11.6 Derive (11.63)–(11.65).

11.7 Derive (11.67).

11.8 Derive (11.72).

11.9 Derive (11.79), (11.80).

11.10 Consider a four-stage RF amplifier. Assume the noise figure/gain of the second stage are 6/15 dB and those of the third and fourth stages are 10/25 dB and 15/35 dB, respectively. Plot the overall noise figure of the amplifier versus noise figure of the first stage from 2 to 10 dB in 0.5-dB steps for 10, 15, 20, and 25-dB gain of the first stage. Comment on the results.

11.11 Consider a 0.25-μm MOSFET with the following measured *S*-parameters:

Frequency (GHz)	S_{11}	S_{12}	S_{21}	S_{22}
8	$0.9244\angle -45.1°$	$0.1103\angle 53.4°$	$1.9542\angle 137.2°$	$0.7176\angle -43.9°$
9	$0.9073\angle -50.8°$	$0.1211\angle 49°$	$1.8677\angle 132.1°$	$0.7015\angle -48.8°$
10	$0.8960\angle -54.4°$	$0.1275\angle 46.3°$	$1.8096\angle 129°$	$0.6914\angle -51.9°$
11	$0.8797\angle -59.4°$	$0.1359\angle 42.4°$	$1.7244\angle 124.2°$	$0.6771\angle -56.4°$
12	$0.8693\angle -62.7°$	$0.1409\angle 40°$	$1.6685\angle 121.3°$	$0.6682\angle -59.2°$

In these *S*-parameters, ports 1 and 2 are at the gate and drain, respectively, and the source and bulk are connected to ground.

a) Calculate and plot the unilateral FOM versus frequency.
b) Determine the error in using the unilateral gain equation and assess possible use of the equation for the MOSFET.

11.12 Draw the constant gain circles for $G_S = -5$, 0, and 5 dB at 10 GHz using the MOSFET given in Problem 11.11.

11.13 Describe in details a graphical gain-design procedure using constant-gain circles for the available gain G_A, similar to that typically used for unilateral amplifier gain design based on the transducer power

gain, for the design of amplifiers employing unconditionally stable bilateral MOSFETs. Derive all necessary equations used in the procedure.

11.14 Repeat Problem 11.13 for potentially unstable bilateral devices.

11.15 A CMOS amplifier is to be designed for maximum (unilateral) gain at 10 GHz using the 0.25-μm MOSFET given in Problem 11.11.

a) Design the IMNs and OMNs using lumped elements.

b) Design the IMNs and OMNs using microstrip lines in a CMOS structure. The signal line (top conductor) is located on the top metal layer and the (assumed infinitely large) ground plane is on another metal layer spaced 5 μm below the top metal layer. The metal is copper with thickness of 0.5 μm and conductivity of 5.8×107 mhos/m. The dielectric between the metal layers is SiO_2 having relative dielectric constant of 3.9 and loss tangent of 0.0002.

11.16 A 0.18-μm NMOSFET has the following S-parameters at 20 GHz: $S_{11} = 0.818\angle -119.6°$, $S_{12} = 0.065\angle 20.3°$, $S_{21} = 5\angle 87.5°$, and $S_{22} = 0.632\angle -67.4°$. Ports 1 and 2 are at the gate and drain, respectively, and the source and bulk are connected to the ground.

a) Draw the gain circle for $G_S = 3$ dB and design the IMN with the greatest degree of stability using lumped elements.

b) Design the IMN using microstrip lines having parameters given in Problem 11.15(b).

11.17 Consider the 0.25-μm MOSFET provided in Problem 11.11. Assume that this device has the following noise parameters at 10 GHz: $F_{\min} = 1.031$ dB and $r_n = 1.497$.

a) Assume the optimum input reflection coefficient for minimum noise figure is $\Gamma_{So} = 0.902\angle 56.162°$ at 10 GHz. Can a stable amplifier be designed at 10 GHz? Provide your rationale along with proper design information leading to stability if a stable amplifier can be designed.

b) Assume $\Gamma_{So} = 0.601\angle 57.53°$ at 10 GHz. Design a lumped-element LNA using this MOSFET for minimum noise figure and maximum possible gain at 10 GHz. Use two lumped elements for each of the IMNs and OMNs. Include the amplifier's schematic and the Smith charts used in the design and calculate the corresponding amplifier's gain. Include stability circles in the Smith chart. The source and load impedances are assumed to be 50 Ω.

c) Use a circuit simulator such as Agilent ADS to calculate and plot the S-parameters of the amplifier designed in Part b from 8 to 12 GHz.

11.18 Consider again the 0.25-μm MOSFET provided in Problem 11.11. Assume that this device has the following noise parameters at 10 GHz: $F_{\min} = 1.031$ dB, $r_n = 1.497$, and $\Gamma_{So} = 0.902\angle 56.162°$.

a) Can an amplifier be designed with this MOSFET to achieve $G_S = 5$ dB and 3-dB noise figure at 10 GHz?

b) Design an amplifier using this MOSFET to achieve $G_S = 5$ dB with a noise figure of 2 dB at 10 GHz. Use two lumped elements for each of the IMNs and OMNs. Include the amplifier's schematic and the Smith charts used in the design and calculate the resultant amplifier's gain. Include stability circles in the Smith chart. The source and load impedances are assumed to be 50 Ω.

c) Use a circuit simulator to calculate and plot the S-parameters of the designed amplifier from 8 to 12 GHz.

11.19 Consider a simple differential amplifier as shown in Figure 11.35 with $R_D = 50$ kΩ, $g_m = 50$ mS, and $R_{CS} = 800$ kΩ. Calculate and plot the differential-mode (G_d), common-mode (G_c), common-to-differential-mode (G_{cd}) and differential-to-common-mode (G_{dc}) voltage gains, and the common -mode (CMRR), common-differential-conversion (DCRR) and differential-common-conversion (CDRR) rejection ratios versus Δg_m from 0 to 5 mS with ΔR_D varied from 0 to 5 kΩ in 1-kΩ steps. Comment on the results.

11.20 Consider a differential amplifier as shown in Figure 11.35 with $R_D = 100\,\text{k}\Omega$. Investigate and compare the effects of bias currents on the common-mode level of the input signals to the amplifier and hence its gain and output voltage swing for $R_{CS} = 1\,\text{M}\Omega$ and $0\,\Omega$ at 2.45 GHz. Use any available submicron MOSFET in the simulation.

11.21 Consider the differential amplifier in Problem 11.19. Calculate and comment on the results for V_{ic}, V_{id}, V_{o1}, V_{o2}, V_{oc}, V_{od}, G_c, G_d, G_{cd}, G_{dc}, CMRR, DCRR, and CDRR for the following two cases with ($\Delta g_m = 0$ mS and $\Delta R_D = 0\ \Omega$) and ($\Delta g_m = 0.25, 5$ mS and $\Delta R_D = 2.5, 5\ \text{k}\Omega$):

a) $V_{i1} = 5\angle 0°$ V and $V_{i2} = 5\angle\theta$ V, where θ varies from 160° to 200° in steps of 5°.

b) $V_{i1} = 5\angle 0°$ V and $V_{i2} = |V_{i2}|\angle 180°$ V, where $|V_{i2}|$ changes from 4 to 6 V in steps of 0.25 V.

c) $V_{i1} = 5\angle 0°$ V and $V_{i2} = 6\angle 190°$ V.

11.22 DC offset occurs in a differential amplifier when there is imbalance between the two halves of the differential amplifier such as mismatches between the transistors and between the drain resistors. The DC offset can degrade the amplifier's performance and limit its potential use or integration with other active RFICs in systems. Describe in details possible adverse effects caused by nonzero DC offset.

11.23 The magnitude of DC offset to a differential amplifier can be assessed by calculating the differential output voltage $V_o = V_{o1} - V_{o2}$ as a function of the differential input voltage $V_i = V_{i1} - V_{i2}$ The higher V_i needed to produce zero V_o the more severe of the DC offset caused by the amplifier. Consider a 5.8-GHz differential amplifier as shown in Figure 11.35. Assume $R_D = 80\,\text{k}\Omega$, $g_m = 50$ mS, $\Delta g_m = 0$ mS, and $R_{CS} = 1\,\text{M}\Omega$. Calculate and plot V_i, corresponding to zero V_o, versus ΔR_D from 0 to 10 kΩ in 500-Ω steps. Use any available submicron MOSFET in the simulation. Comment on the results.

11.24 Consider the amplifier described in Problem 11.23 with $\Delta R_D = 0$. Calculate and plot V_i, corresponding to zero V_o, versus Δg_m from 0 to 1 mS. Comment on the results.

11.25 Design a 2.3–2.7 GHz differential amplifier using any available submicron CMOS process and on-chip spiral inductors. There are no required specifications for this amplifier. However, you are required to try your best to achieve best possible gain, noise figure, isolation, match, linearity, and power consumption. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption.

11.26 Design an 8–12 GHz balanced amplifier using 90° 3-dB microstrip Lange couplers and on-chip inductors on any available submicron CMOS process. There are no required specifications for this amplifier. However, you are required to try your best to achieve best possible gain, noise figure, isolation, match, linearity, and power consumption. The microstrip line is realized using the top-most metal layer (metal 1) for the signal line and the third metal layer (metal 3) for the ground plane. The properties of these metals including metallization thickness and loss tangent should be available from the process design kit (PDK) of the employed CMOS process. The dielectrics below the top-most metal can also be found in the PDK. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption. Compare the performance with that of the single amplifier used in the balanced amplifier. Also, plot and discuss results of the S-parameters of the designed Lange coupler.

11.27 Derive Eq. (11.203).

11.28 Derive (11.213).

11.29 Derive (11.215).

11.30 Derive (11.233)–(11.235).

11.31 Derive (11.236).

11.32 Derive (11.243) and (11.246)–(11.248) based on References [7, 8].

11.33 Assume the input signal to the 90° balanced amplifier is $V_i = V_m \cos \omega t$. Derive the output signal of the amplifier. Discuss about the cancellation and attenuation of harmonics of the 90° balanced amplifier.

11.34 Both 90° balanced amplifier and push–pull amplifier are often used for high power applications. Compare and list the advantages and disadvantages of these amplifiers for both narrow and broadband PAs.

11.35 Is the 90° balanced amplifier more stable than the push–pull amplifier? Provide your rationale.

11.36 Design an 8–12 GHz push–pull amplifier using a broadband balun and on-chip inductors on any available submicron CMOS process. There are no required specifications for this amplifier. However, you are required to try your best to achieve best possible gain, noise figure, isolation, match, linearity, and power consumption. If transmission lines are employed then use the top-most metal layer (metal 1) for the signal line and the third metal layer (metal 3) for the ground plane if the ground plane and signal line are at two different planes. The properties of these metals including metallization thickness and loss tangent should be available from the PDK of the employed CMOS process. The dielectrics below the top-most metal can also be found in the PDK. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption. Compare the performance with that of the single amplifier used in the push–pull amplifier. Also, plot and discuss results of the S-parameters of the designed balun. Compare the performance with that of the balanced amplifier in Problem 11.26 if this amplifier is also designed.

11.37 One problem of the push–pull amplifier is the inability of its input balun to cancel out the signals reflected back to the amplifier's input port, degrading the matching. Consider a push–pull amplifier designed at 10 GHz with 50-Ω source and load impedances. The S-parameters of its constituent components are listed below.
Balun's S-parameters: $S_{11} = 0.2\angle 50°$, $S_{21} = 0.707\angle 20°$, and $S_{31} = 0.707\angle 200°$. Port 1 is the input port and ports 2 and 3 are at the output ports. Amplifiers A and B's S-parameters: $S_{11} = 0.3\angle -110°$, $S_{12} = 0.065\angle 30°$, $S_{21} = 5\angle 140°$, and $S_{22} = 0.28\angle -70°$. Ports 1 and 2 are the input and output ports, respectively. Assume the input signal to the amplifier is $V_i = 1\angle 0°$ V.

a) Calculate the signal reflected back to the input port of the push–pull amplifier caused by the mismatch between the input balun and the individual amplifiers.
b) Calculate the signal reflected back to the input port of the push–pull amplifier caused by the mismatch between the output balun and the individual amplifiers.
c) Calculate the input and output return losses of the push–pull amplifier.
d) Discuss the contributions to the push–pull amplifier's matching from the matching of its constituent components and possible ways to alleviate this problem.

11.38 A distributed amplifier is to be designed using identical MOSFETs having the following parameters: $R_g = 5\,\Omega$, $R_D = 250\,\Omega$, $C_{gs} = 165$ fF, and $g_m = 50$ mS. Assume the characteristic impedances of the gate and drain lines are 50 Ω.

a) Determine the optimum number of stages that produce maximum gain at 18 GHz.
b) Calculate and plot the amplifier gain (dB) versus frequency from 1 to 30 GHz for $N = 5$, 10, 15, 20, and 25. Provide comments on the results.

11.39 Consider a 10-stage distributed amplifier using identical MOSFETs having $R_D = 250\,\Omega$, $C_{gs} = 165$ fF and $g_m = 50$ mS. Calculate the amplifier's gain (dB) versus frequency from 1 to 30 GHz for r_g from 0 to 20 Ω in 5-Ω steps. Comment on the results.

11.40 Consider a 10-stage distributed amplifier using identical MOSFETs having $R_g = 5\,\Omega$, $C_{gs} = 165$ fF and $g_m = 50$ mS. Calculate the amplifier's gain (dB) versus frequency from 1 to 30 GHz for R_D from 100 to 1000 Ω in steps of 100 Ω. Comment on the results.

11.41 Calculate the difference in dB for the carrier-to-third-order-intermodulation-power ratio (C/IM3) for different number of stages (N) from 1 to 5 of a distributed amplifier for both lossless and lossy input and output transmission lines with total attenuation of $\alpha_i = 0.2$ Np and $\alpha_o = 0.10$ Np per section, respectively.

11.42 Design a five-stage distributed amplifier operating from 1 to 10 GHz using identical common-source MOSFETs from any available submicron CMOS process and ideal inductors with equal and different inductor values for the gate and drain lines. The input and output ports of the amplifier are 50 Ω. The amplifier does not have specific performance requirements such as gain, noise figure, linearity, matching, and power consumption.

a) Draw the amplifier's schematic including all initially designed parameters. Calculate and plot the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K) from 1 to 20 GHz and power consumption using available CAD program for two cases: equal and different inductance values for the gate and drain inductors. Discuss the results of all the parameters.
b) Use available CAD program to optimize the initially designed parameters to produce best possible performance in gain, noise figure, isolation and match for the amplifier across 1–10 GHz. Draw the amplifier's schematic including all final parameters. Plot and discuss the amplifier's performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K) from 1 to 20 GHz, and power consumption. Comment on the performance obtained here and that obtained in Part a.
c) Investigate the effects of the inductors of the gate and drain lines on the gain, input and output return loss, and noise figure from 1 to 10 GHz when varying their inductance from the optimized value (obtained in Part b) to ±10% and 20%. Assume equal identical gate and drain inductors ($L_g = L_d$). Comment on the results.
d) Instead of using identical MOSFETs for all stages of the distributed amplifier, what do you think the performance of the distributed amplifier might be if different MOSFETs are properly chosen? Provide your rationale and support your conclusion with simulations.

11.43 Redesign the distributed amplifier designed in Part b of Problem 11.40 using on-chip spiral inductors. Optimize the amplifier to produce best possible performance in gain, noise figure, isolation, and match from 1 to 10 GHz. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance of the designed amplifier including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K) from 1 to 20 GHz, and power consumption. Comment on its performance with respect to that of the amplifier with ideal inductors designed in Part b of Problem 11.40.

11.44 A distributed amplifier is to be designed using identical common-source MOSFETs from any available submicron CMOS process. The characteristic impedances of the synthetic gate and drain transmission lines are assumed to be 50 Ω. Investigate the trade-off between gain and bandwidth of the amplifier by performing the following:

a) Calculate and plot the cut-off frequency (f_c) of the gate line from (11.254) and the transconductance (g_m) as function of the MOSFET's gate width from 50 to 100 μm using available MOSFET's model from the design kit and CAD program. Discuss the results.

b) Repeat Part a for the drain line.

11.45 A distributed amplifier is designed to work across 1–10 GHz using three identical common-source MOSFETs and microstrip lines. The microstrip line is realized using the top-most metal layer for the signal line and another metal layer for the ground plane. These metals are copper having a thickness of 0.5 μm and conductivity of 5.8×107 mhos/m. The dielectric between the metals is 5-μm thick SiO_2 having relative dielectric constant of 3.9 and loss tangent of 0.0002. Assume the MOSFETs have the gate–source capacitance (C_{gs}) and drain–source capacitance (C_{ds}) of 180 and 130 fF, respectively, and each microstrip line section (ℓ_g and ℓ_d) of the gate and drain lines is less than a quarter-wavelength ($\lambda/4$). It is also assumed that the characteristic impedances (Z_g and Z_d) of the gate and drain lines and the amplifier's input and output impedances are 50 Ω.

a) Determine the characteristic impedances and lengths of the microstrip lines used in the gate and drain lines including those at the input and output of the transmission lines. Calculate the insertion loss and return loss (*S*-parameters) of the gate and drain lines using any available CAD program.

b) Optimize the length and characteristic impedance of the microstrip lines in the gate and drain lines for lowest possible insertion loss and highest possible return loss using an available CAD program. Compare and comment on the results of the length, characteristic impedance, insertion loss and return loss with those obtained in Part a.

c) Now assume the MOSFETs have transconductance (g_m) of 50 mS and output drain–source resistance (R_{ds}) of 200 Ω. Simulate the gain and return loss of the amplifier using nonoptimized and optimized values for the microstrip lines obtained in Parts a and b, respectively. Discuss the results.

11.46 Design a five-stage distributed amplifier operating from 1 to 10 GHz using identical common-source MOSFETs from any available submicron CMOS process and (finite-ground) CPW. Consider both equal and different characteristic impedances for the CPW used in the gate and drain lines. The CPW is realized using the top-most metal layer for the signal line and ground strips. The properties of this metal including metallization thickness and loss tangent should be available from the PDK of the employed CMOS process. The dielectrics below the top-most metal can also be found in the PDK. The input and output ports of the amplifier are 50 Ω. The amplifier does not have specific performance requirements such as gain, noise figure, linearity, matching, and power consumption.

a) Draw the amplifier's schematic including all initially designed parameters. Calculate and plot the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K) from 1 to 20 GHz and power consumption using available CAD program for two cases: equal and different characteristic impedances for the gate and drain CPW. Discuss the results of all the parameters.

b) Use available CAD program to optimize the initially designed parameters to produce best possible performance in gain, isolation and match for the amplifier across 1–10 GHz. Draw the amplifier's schematic including all final parameters. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption. Comment on the performance obtained here and that obtained in Part a.

c) Investigate the effects of the characteristic impedance of the CPW on the gain, input and output return loss, and noise figure from 1 to 10 GHz when varying the CPW characteristic impedance from the optimized value (obtained in Part b) to ±10% and 20%. Assume equal characteristic impedances ($Z_{og} = Z_{od}$). Comment on the results.

d) Compute and plot the ratios between the RF power dissipated in the gate (R_{gt}) and drain (R_{dt}) termination resistors and the RF input power to the amplifier from 1 to 10 GHz.

e) In practice, C_{gs} and C_{ds} change from their nominal values listed in the PDK which are used in the design of the distributed amplifier. Assume C_{gs} and C_{gd} vary within ±20% from the nominal values while all other design parameters of the distributed amplifier remain the same. Compute and plot the ratios between the RF power dissipated in the gate (R_{gt}) and drain (R_{dt}) termination resistors and the RF input power to the amplifier from 1 to 10 GHz for five cases: C_{gs} and C_{ds} as in the PDK, C_{gs} and C_{ds} varied ±10%, and C_{gs} and C_{ds} varied ±20%. Comment on the results.

11.47 Consider the cascode gain cell shown in Figure 11.57(b).

a) Derive expressions for the transconductance and its cut-off frequency.

b) Calculate and plot the transconductance of the cascode gain cell, cascaded common-source gain cell in Figure 11.57(d), and single common-source MOSFET versus frequency from 1 to 20 GHz using any available submicron CMOS process. Discuss the results.

11.48 Consider the current-reuse gain cell shown in Figure 11.57(c).

a) Derive expressions for the transconductance and its cut-off frequency.

b) Calculate and plot the transconductance of the current reuse gain cell, cascaded common-source gain cell in Figure 11.57(d), and single common-source MOSFET versus frequency from 1 to 30 GHz using any available submicron CMOS process. Discuss the results.

11.49 Consider the cascaded common-source gain cell shown in Figure 11.57(d) with MOSFETs from any available submicron CMOS process. Calculate and plot the transconductance of the gain cell versus frequency from 1 to 30 GHz without peaking inductor and with peaking inductor values from 0.5 to 3 nH in 0.5-nH steps. Discuss the results.

11.50 Consider the cascaded common-source gain cell shown in Figure 11.57(d) with MOSFETs from any available submicron CMOS process. The peaking inductor value is 2 nH. Calculate and plot the transconductance of the gain cell versus frequency from 1 to 30 GHz with resistor R varying from 50 to 150 Ω in steps of 20 Ω. Discuss the results.

11.51 Consider the cascaded common-source gain cell shown in Figure 11.57(d) with MOSFETs from any available submicron CMOS process. The peaking inductor value is 2 nH. Calculate and plot the stability factor K versus frequency from 1 to 30 GHz for different values of resistor R from 50 to 150 Ω in steps of 20 Ω. Determine the values of R for achieving an unconditionally stable gain cell.

11.52 Design a five-stage distributed amplifier operating from 1 to 10 GHz using the cascaded common-source gain cells shown in Figure 11.57(d) with MOSFETs from any available submicron CMOS process and on-chip inductors. The input and output ports of the amplifier are 50 Ω. The amplifier does not have specific performance requirements such as gain, noise figure, linearity, matching, and power consumption.

a) Draw the amplifier's schematic including all initially designed parameters. Calculate and plot the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K) from 1 to 20 GHz and power consumption using available CAD program. Discuss the results of all the parameters.

b) Use an available CAD program to optimize the initially designed parameters to produce best possible performance in gain, isolation and match for the amplifier across 1–10 GHz. Draw the amplifier's schematic including all final parameters. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance

including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption. Comment on the performance obtained here and those obtained in Part a and Problem 11.43.

11.53 Derive (11.375) for a CMOS source follower.

11.54 Design a 3–10 GHz shunt feedback amplifier using any available submicron CMOS process and on-chip spiral inductors. There are no required specifications for this amplifier. However, you are required to try your best to achieve best possible gain, noise figure, isolation, match, linearity, and power consumption. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption.

11.55 Derive (11.378).

11.56 Design a 3–10 GHz cascoded common-source inductively degenerate amplifier using any available submicron CMOS process and on-chip spiral inductors. There are no required specifications for this amplifier. However, you are required to try your best to achieve best possible gain, noise figure, isolation, match, linearity, and power consumption. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including gain, noise figure, input and output return loss, isolation, input IP3 (IIP3), output IP3 (OIP3), stability factor (K), and power consumption.

11.57 As stated in Section 11.4, the amplifier's DC-RF conversion efficiency depends on the waveform of the RF excitation. Prove that, for ideal Class A PAs, the DC-RF conversion efficiency can reach 100% under a square-wave driving signal.

11.58 Design a Class A PA operating at 10 GHz using any available submicron CMOS process. There are no required specifications for this amplifier. However, you are required to try your best to achieve the highest possible output power, gain, and power-added efficiency. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out. Perform post-layout simulations. Plot and discuss the performance including output power, power-added efficiency, DC-RF conversion efficiency, gain, input and output return loss, and stability factor (K).

11.59 A Class-A PA is designed to operate under sinusoidal excitation using a single transistor in a CMOS process with $V_{dd} = 5$ V. The required RF output power is 10 W. Calculate the following fundamental parameters: load resistance R_L, RF voltage amplitude V_{dm} and current amplitude I_{dm} at the drain, DC current $I_{\rm dc}$, DC input power $P_{\rm dc}$, power dissipation P_d, and DC-RF conversion efficiency $\eta_{\rm dc}$.

11.60 Prove that, for ideal Class-B PAs, the DC-RF conversion efficiency can reach 100% under a square-wave driving signal.

11.61 Design a single-transistor Class-B PA operating at 10 GHz using any available submicron CMOS process. There are no required specifications for this amplifier. However, you are required to try your best to achieve the highest possible output power, gain and power-added efficiency. Draw the schematic with all element values. Prepare the layout of the amplifier ready for tape-out. Perform post-layout simulations. Plot and discuss the performance including output power, power-added efficiency, DC-RF conversion efficiency, gain, input and output return loss, and stability factor (K).

11.62 Consider a single-transistor PA operated under RF sinusoidal signals. Assume the transistor's maximum allowable power dissipation is 100 W.

a) Assume the PA is operated in Class A under the maximum output power operation, what is the highest RF output power that the amplifier can possibly deliver to the load?

b) Assume the PA is operated in Class B under the maximum output power operation, what is the highest RF output power that the amplifier can possibly deliver to the load?

11.63 A Class-B PA is designed to operate under sinusoidal excitation using a single transistor in a CMOS process with $V_{dd} = 5$ V. The required RF output power is 10 W. Calculate the following fundamental parameters: load resistance R_L, RF voltage amplitude V_{dm} and current amplitude I_{dm} at the drain, average (DC) current I_{dc}, DC input power P_{dc}, power dissipation P_d, and DC-RF conversion efficiency η_{dc}.

11.64 Derive an expression for the maximum drain-current amplitude, $I_{d,\max}$ in Class-C amplifiers in terms of the supplied voltage V_{dd}, conduction angle θ, and load resistance R_L. What is the value of $I_{d,\max}$ as θ is reduced to zero?

11.65 Sketch the waveforms of the voltage across and current through the load in Class-C amplifiers assuming the OMN completely suppresses all the harmonics generated at the drain.

11.66 Consider a MOSFET having $S_{11} = 0.75$ $S_{11} = 0.75\angle 85°$ at 30 GHz. Draw the constant gain circles for $G_s = -7$, 1, and 2 dB.

11.67 Consider a 0.18-μm NMOS whose *S*-parameters across 1–20 GHz are given in Problem 9.1, but assume $|S_{12}| \simeq 0$ at 15 GHz to make it unilateral at 15 GHz. You can also use the same device that you use in Problem 9.1 but assume $|S_{12}| \simeq 0$ at 15 GHz. Design an amplifier using this MOSFET for maximum (unilateral) gain at 15 GHz.

a) Design the IMNs and OMNs using ideal lumped elements.

b) Design the IMNs and OMNs using (finite-ground) CPW in a CMOS structure. The CPW is realized using the top-most metal layer for the signal line and ground strips. The metal is copper having a thickness of 0.5 μm and conductivity of 5.8×10^7 mhos/m. There are five SiO_2 dielectric layers, each is 5-μm thick with relative dielectric constant ε_r of 3.9 and loss tangent of 0.0002, between the metal and the 250-μm silicon substrate having $\varepsilon_r = 11.7$ and resistivity $\rho = 0.5\,\Omega$-cm.

In these designs, draw the amplifier's schematic including values for all elements, check the stability at 15 GHz, and calculate and plot the designed amplifier's gain (S_{21} in dB) and return losses (S_{11} and S_{22} in dB) versus frequency from 1 to 20 GHz using a commercially available computer program. Compare and comment between the calculated and desired gain.

11.68 Consider a 0.18-μm NMOS FET whose *S*-parameters are given in Problem 9.1.

a) Design a lumped-element amplifier in a 50-Ω system using this MOSFET for maximum (unilateral) gain at 17.1 GHz using the unilateral design procedure even though the device is bilateral as seen in its S_{12} at 17.1 GHz. Calculate the amplifier's gain at 17.1 GHz and compare with the maximum unilateral gain $G_{u,\max}$. Check the stability of the device and amplifier at 17.1 GHz. Plot the source and load stability circles and indicate the stable and unstable regions. Can the amplifier be designed to have maximum gain at 17.1 GHz while maintaining its stability? If not, then choose Z_S and Z_L that will lead to stability while achieving gain as high as possible, design the IMNs and OMNs using lumped elements, and calculate the corresponding gain at 17.1 GHz. What is the unilateral FOM of the device? Is the design assumption of unilateral device valid in this case?

b) If you are convinced that the unilateral assumption is not justified, then redesign the amplifier (at 17.1 GHz) using the bilateral procedure based on the operating power gain outlined in Section Design Based on Operating Power Gain. Can this amplifier be designed to achieve the maximum stable gain? Indicate the range of gain over which the amplifier should not be designed for to avoid unstability or potential unstability. Choose an appropriate gain to guarantee stability with some margin and design the amplifier using ideal lumped elements to achieve that gain. Calculate and plot the designed amplifier's gain (S_{21} in dB) and return losses (S_{11} and S_{22} in dB) versus frequency from 1 to 20 GHz using a commercially available computer program. Comment on the gain result with respect to your desired gain. Provide calculations confirming that the designed amplifier is stable at 17.1 GHz. Is this amplifier stable or unstable across 1–20 GHz or parts of this frequency ranges? Provide a rationale for your answer.

11.69 Derive Eqs. (11.70) and (11.71) for the available power gain G_A and operating power gain G_O using signal flow graphs, respectively.

11.70 Derive Eqs. (11.293) and (11.294) for the gate line's characteristic impedance and propagation constant, respectively, using binomial series $(1+a)^{\pm 1/2} \simeq 1 \pm \frac{a}{2}$ when $a \ll 1$.

11.71 Derive Eqs. (11.296) and (11.297) for the drain line's characteristic impedance and propagation constant, respectively, using binomial series $(1+a)^{\pm 1/2} \simeq 1 \pm \frac{a}{2}$ when $a \ll 1$.

APPENDIX 11: SIGNAL FLOW GRAPH

Signal flow graph [A1, A2] is a relatively simple technique for deriving parameters for circuits involving reflection and transmission of signals. This technique, as noted in this chapter, Chapters 10 and 12, has often been used to derive expressions for the reflection coefficients and gains for amplifiers. Its details are given in this appendix.

A11.1 FUNDAMENTALS OF SIGNAL FLOW GRAPH

For illustration purpose via a simple analysis, without lost of generation, we consider a two-port network consisting of a MOSFET, represented by its scattering (S) parameters, and IMNs and OMNs as shown in Figure A11.1(a). This network indeed represents a very basic amplifier. Figure A11.1(b) represents an equivalent circuit of the two-port network and Figure A11.1(c) shows another representation of the two-port network by decomposing it into three constituent blocks.

The drawing and analysis of signal flow graphs depend on several rules, hereafter referred to as the "signal flow graph rules," as described in the following. We consider each (normalized) voltage wave (a_i, b_i, where $i=1$, 2, or b_S) as a circuit "node," which the corresponding wave arrives to or leaves from, and each of the reflection and transmission coefficients, including S-parameters S_{ij} (i, $j=1$, 2), as representing a circuit "branch" connecting every two nodes according to the relation defined by the corresponding transmission and reflection coefficients, and S-parameters. We also refer to a node as an "incident node" or "reflected node" if that node represents an incident wave or reflected wave, respectively. For instance, a_i and b_i are called "incident node" and "reflected node" for the device's S-parameter block in Figure A11.1(c), respectively. It is noted that an incident wave represents an independent variable and a reflected wave represents a dependent variable. Furthermore, a branch always enters a reflected node (dependent node) and leaves an incident node (independent node) in a network. For example, in the device's S-parameter block, S_{11} is represented by a branch starting at note a_1 and ending at node b_1, while S_{21} is the branch going from a_1 to b_2. Since in a network consisting of several subnetworks, an incident wave in one subnetwork can act as a reflected wave in another subnetwork or vice versa, a node can be considered an incident node in one subnetwork (or network) while representing a reflected node in another subnetwork (or network). Moreover, a node is equal to the summation of all the branches entering it. The signal flow graph rules enable us to draw signal flow graphs for one-, two-, and multiport network utilizing the incident and reflected waves and associated S-parameters as well as reflection and transmission coefficients.

A11.2 SIGNAL FLOW GRAPH OF TWO-PORT NETWORKS

We consider again the two-port network in Figure A11.1(a). The signal flow graph of this two-port network would consist of the signal flow graphs for its individual blocks superimposed together. Therefore, the first step is to draw the signal flow graphs for the transistor, IMN, and OMN seen in Figure A11.1(c), and then superimpose them to obtain the overall signal flow graph of the composite two-port network. These steps are described in details as follows.

A11.2.1 Transistor's Signal Flow Graph

For the transistor represented by its S-parameters as shown in Figure A11.2(a), the (normalized) incident and reflected voltage waves are a_i and b_i, respectively. Applying the signal flow graph rules enables us to draw the corresponding signal flow graph as shown in Figure A11.2(b). Adding all the branches entering (reflected) nodes b_1 and b_2 and equating to the corresponding nodes, we get

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \tag{A11.1}$$

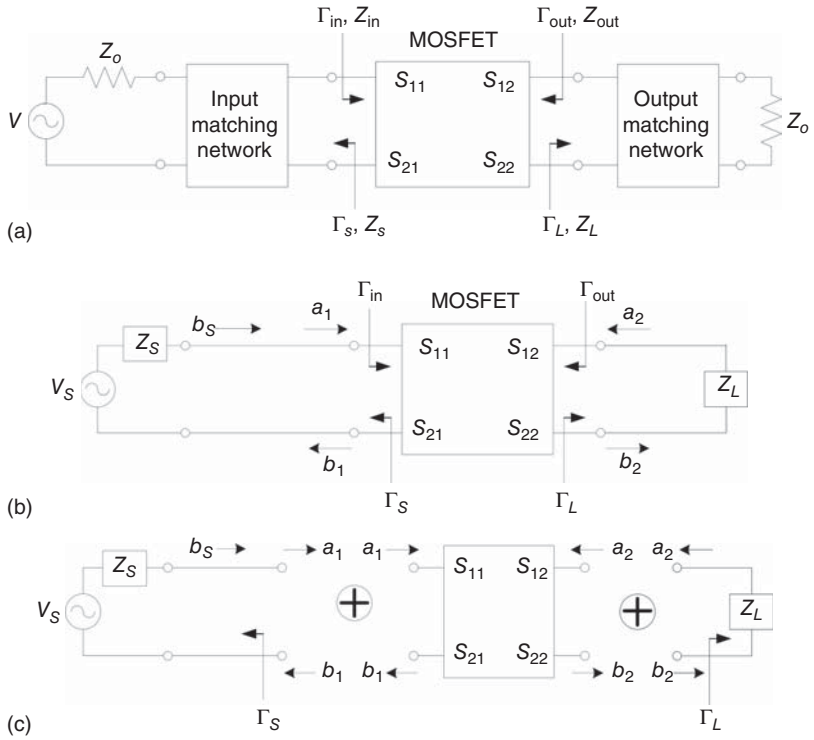


Figure A11.1. Two-port network (a), equivalent circuit (b), and its decomposition into three blocks (c). b_S represents the (normalized) voltage wave from the source V_S, and a_i and b_i ($i = 1, 2$) represent the (normalized) voltage waves at port i. S_{ij} ($i, j = 1, 2$) are the S-parameters of the transistor.

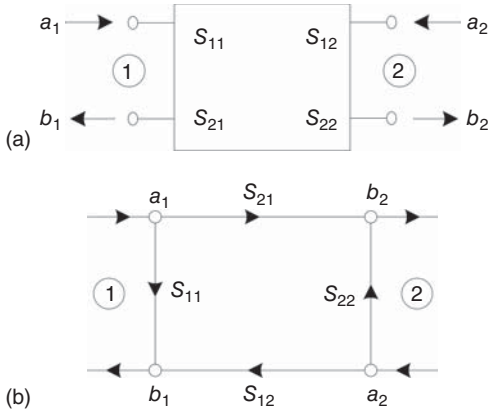


Figure A11.2. Two-port transistor represented by its S-parameters (a) and the corresponding signal flow graph (b). a_i and b_i are the incident and reflected nodes, respectively.

which are recognized as the standard equations for S-parameters that relate reflected to incident voltage waves obtained in Eq. (7.24). The signal flow graph thus allows us to derive equations relating the incident and reflected waves quickly.

A11.2.2 Input Matching Network's Signal Flow Graph

We consider the one-port IMN represented by its input impedance Z_s as shown in Figure A11.1(c) and redraw it in Figure A11.3(a) along with its signal flow graph in Figure A11.3(b). It is noted that while the branch between b_S and a_1 is represented as a unit transmission coefficient, a_1 is not equal to b_S unless there is no

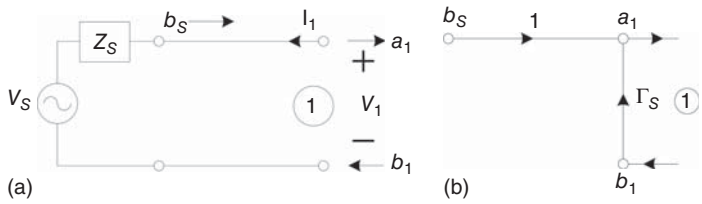


Figure A11.3. Input matching network (a) and its signal flow graph (b). b_S and b_1 are the incident nodes and a_1 is the reflected node. The branch between a_1 and b_S represents a transmission coefficient whose magnitude is 1.

reflection (i.e., $\Gamma_S = 0$). Applying the rules of the signal flow graphs, we can write from Figure A11.3(b):

$$a_1 = b_S + \Gamma_S b_1 \tag{A11.2}$$

which can be verified analytically as follows. We obtain from Figure A11.3(a):

$$V_1 = V_1^+ + V_1^- = V_S + Z_S(I_1^+ + I_1^-) = V_S + Z_S\left(\frac{V_1^+}{Z_o} - \frac{V_1^-}{Z_o}\right) \tag{A11.3}$$

utilizing $Z_o = V_1^+/I_1^+ = -V_1^-/I_1^-$. Dividing (A11.3) by $\sqrt{Z_o}$ and rearranging, we get

$$\frac{V_1^+}{\sqrt{Z_o}}\left(\frac{Z_o - Z_S}{Z_o}\right) = \frac{V_S}{\sqrt{Z_o}} - \frac{V_1^-}{\sqrt{Z_o}}\left(\frac{Z_o - Z_S}{Z_o}\right) \tag{A11.4}$$

Recognizing that $V_1^+/\sqrt{Z_o} = b_1$ and $V_1^-/\sqrt{Z_o} = a_1$ represent the normalized incident and reflected voltage waves at port 1 of the IMN, respectively, we can rewrite (A11.4) as

$$b_1\left(\frac{Z_o - Z_S}{Z_o}\right) = \frac{V_S}{\sqrt{Z_o}} - a_1\left(\frac{Z_o - Z_S}{Z_o}\right) \tag{A11.5}$$

from which, we obtain

$$a_1 = \frac{V_S\sqrt{Z_o}}{Z_o + Z_S} + \frac{Z_S - Z_o}{Z_S + Z_o}b_1 \tag{A11.6}$$

or

$$a_1 = \frac{\frac{V_S}{Z_o+Z_S}}{\sqrt{Z_o}} + \frac{Z_S - Z_o}{Z_S + Z_o}b_1 = b_S + \Gamma_S b_1 \tag{A11.7}$$

where $b_S = V_S/(Z_o + Z_S)/\sqrt{Z_o}$ and $\Gamma_S = (Z_S - Z_o)/(Z_S + Z_o)$ are the normalized voltage wave entering Z_S and the reflection coefficient looking into Z_S, respectively. Equation (A11.7) is indeed the same as (A11.2) obtained previously using the signal flow graph in Figure A11.3(b).

A11.2.3 Output Matching Network's Signal Flow Graph

Now we consider the OMN represented by its input impedance Z_L as shown in Figure A11.1(c), which is shown again in Figure A11.4(a). We can draw the signal flow graph as shown in Figure A11.4(b) and, from which, can write immediately

$$a_2 = \Gamma_L b_2 \tag{A11.8}$$

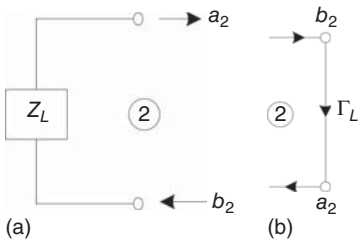


Figure A11.4. Output network (a) and its signal flow graph (b). b_2 and a_2 are the incident and reflected nodes, respectively.

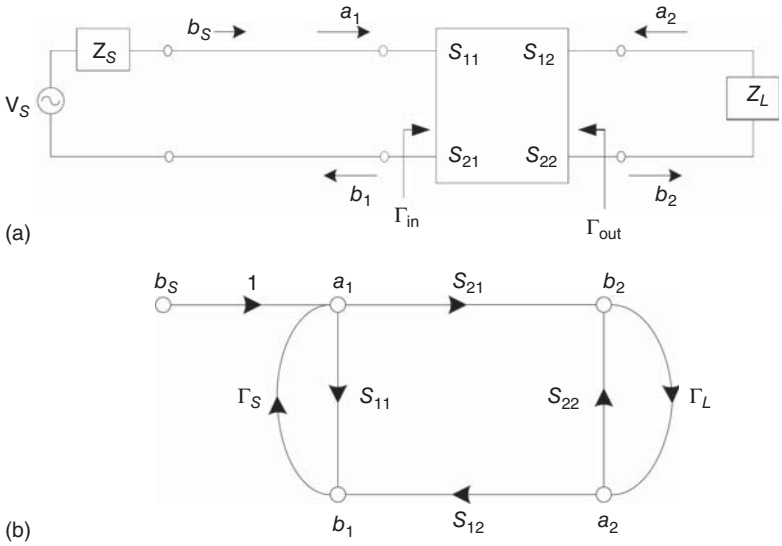


Figure A11.5. Two-port network (a) and its overall signal flow graph (b).

where $\Gamma_L = (Z_L - Z_o)/(Z_L + Z_o)$ which is the standard equation for the reflection coefficient Γ_L. Equation (A11.8) obtained using the signal flow graph indeed represents the same result derived directly from the network in Figure A11.4(a).

A11.2.4 Signal Flow Graph of the Composite Two-Port Network

The three blocks representing the transistor and the IMNs and OMNs are now superimposed to produce the desired two-port network as shown in Figure A11.1(a). The signal flow graph of the composite two-port network can be drawn directly from the same network shown in Figure A11.1(a) or by superimposing the three signal flow graphs of the constituent networks seen in Figures A11.2(b), A11.3(b), and A11.4(b). This signal flow graph is shown in Figure A11.5 along with the two-port network.

The foregoing formulation for two-port networks can be extended for multiport networks and their signal flow graphs can be drawn in a fashion similar to that for two-port networks.

A11.3 DERIVATION OF NETWORK'S PARAMETERS USING SIGNAL FLOW GRAPHS

The primary application of signal flow graphs is deriving expressions for parameters of circuits. To illustrate this application, we consider again the two-port network and its resultant signal flow graph as shown in Figure A11.5. The derivation of parameters using signal flow graphs rely on Mason's rule, also known as the nontouching loop rule. This rule determines the ratio between a dependent variable (reflected wave) b and

an independent variable (incident wave) a according to the following equation:

$$T = \frac{b}{a} = \frac{P_1\left[1 - \sum L(1)^{(1)} + \sum L(2)^{(1)} - \cdots\right] + P_2\left[1 - \sum L(1)^{(2)} + \cdots\right] + \cdots}{1 - \sum L(1) + \sum L(2) - \sum L(3) + \cdots}$$
$$= \frac{\sum\limits_{i=1,2,3,\ldots} P_i\left[1 - \sum\limits_{j=1,3,\ldots\,(\text{odd})} L(j)^{(i)} + \sum\limits_{j=2,4,\ldots\,(\text{even})} L(j)^{(i)}\right]}{1 - \sum\limits_{j=1,3,\ldots\,(\text{odd})} L(j) + \sum\limits_{j=2,4,\ldots\,(\text{even})} L(j)} \tag{A11.9}$$

whose parameters are defined as follows.

P_i ($i = 1, 2, 3, \ldots$) are the different paths connecting the independent node a to the dependent node b along the arrows from a to b. For instance, considering the independent node b_S and dependent node b_1 in Figure A11.5(b), there are two paths connecting b_S to b_1: $P_1 = (1)(S_{11}) = S_{11}$ and $P_1 = (1)(S_{21})(\Gamma_L)(S_{12}) = S_{21}\Gamma_L S_{12}$.

$L(1)$, corresponding to $j = 1$, is the sum of all first-order loops. A first-order loop is defined as the product of all the branches encountered when moving from a node following the same direction of the arrows back to that node. For instance, there are total three first-order loops in Figure A11.5(b) considering different nodes. They are $S_{11}\Gamma_S$ (considering node a_1 or b_1), $S_{21}\Gamma_L S_{12}\Gamma_S$ (considering node a_1, b_1, a_2, or b_2), and $S_{22}\Gamma_L$ (considering node a_2 or b_2).

$L(2)$, corresponding to $j = 2$, is the sum of all second-order loops. A second-order loop is defined as the product of any two nontouching first-order loops. For instance, there are two nontouching first-order loops in Figure A11.5(b): $S_{11}\Gamma_S$ and $S_{22}\Gamma_L$, and the second-order loop is hence $S_{11}\Gamma_S S_{22}\Gamma_L$.

$L(3)$, corresponding to $j = 3$, is the sum of all third-order loops. A third-order loop is defined as the product of any three nontouching first-order loops. Figure A11.5(b) does not have three nontouching first-order loops and so there is no third-order loop for this network. It is noted that more complicated network (e.g., those having feedback paths) may have third- or higher-order loops such as fourth-order loop $L(4)$, fifth-order loop $L(5)$, etc.

$L(j)^{(i)}$, $j = 1, 2, 3, \ldots$ and $i = 1, 2, 3, \ldots$, is the sum of all jth-order loops that do not touch the ith path connecting the independent node a to the dependent node b. To illustrate this, we consider Figure A11.5(b) and nodes a_1 and b_1. $L(1)^{(1)}$ is the sum of all first-order loops that do not touch the first path connecting a_1 to b_1. The first path from a_1 to b_1 is $P_1 = S_{11}$ and the only first-order loop in the network that does not touch P_1 is $S_{22}\Gamma_L$, and so $L(1)^{(1)} = S_{22}\Gamma_L$. The second path from a_1 to b_1 is $P_2 = S_{21}\Gamma_L S_{12}$ and there is no first-order loop that does not touch P_2, and so $L(1)^{(2)} = 0$. $L(2)^{(1)}$ is the sum of all second-order loops that do not touch the first path P_1 connecting a_1 to b_1. Since there is no second-order loop that does not touch P_1, $L(2)^{(1)} = 0$. $L(1)^{(2)}$ is the sum of all first-order loops that do not touch the second path P_2. Since there is no first-order loop that does not touch P_2, $L(1)^{(2)} = 0$. Similarly, $L(j)^{(i)} = 0$, $j = 3, 4, \ldots$ and $i = 3, 4, \ldots$

A11.3.1 Examples of Derivation

We consider again the two-port network represented by a signal flow graph as shown in Figure A11.5. As an example, we now determine the ratio between the voltage waves b_1 and b_S using the signal flow graph Figure A11.5(b) and Mason's rule. To do that, we first identify the following:

- Paths connecting independent node to dependent node (from b_S to b_1): There are two paths along the arrows from b_S to b_1: $P_1 = S_{11}$ and $P_2 = S_{21}\Gamma_L S_{12}$.
- Sum of first-order loops: There are three first-order loops following the arrows from a node (a_1, b_1, a_2, or b_2) back to the same node in the network: $S_{11}\Gamma_S$, $S_{21}\Gamma_L S_{12}\Gamma_S$, and $S_{22}\Gamma_L$). Hence, $L(1) = S_{11}\Gamma_S + S_{21}\Gamma_L S_{12}\Gamma_S + S_{22}\Gamma_L$.

Sum of second-order loops: Among the three first-order loops, there are only two loops that do not touch each other: $S_{11}\Gamma_S$ and $S_{22}\Gamma_L$. Therefore, $L(2) = S_{11}\Gamma_S S_{22}\Gamma_L$.

First-order loops not touching the paths: The first-order loop $S_{22}\Gamma_L$ does not touch the path $P_1 = S_{11}$, and so $S_{22}\Gamma_L$ is a nontouching loop with respect to the path P_1, and hence $L(1)^{(1)} = S_{22}\Gamma_L$. With respect to the path $P_2 = S_{21}\Gamma_L S_{12}$, all the first-order loops touch this path there and so $L(1)^{(2)} = 0$.

Second- and higher-order loops not touching the paths: There are no second- and higher-order loops that do not touch the paths $P_1 = S_{11}$ and $P_2 = S_{21}\Gamma_L S_{12}$ and so $L(j)^{(i)} = 0$, $j = 2, 3, \ldots$ and $i = 1, 2$.

Finally, applying the foregoing results to Mason's rule (A11.9), we obtain

$$\frac{b_1}{b_S} = \frac{S_{11}(1 - \Gamma_L S_{22}) + S_{21}\Gamma_L S_{12}}{1 - (S_{11}\Gamma_S + S_{21}\Gamma_L S_{12}\Gamma_S + S_{22}\Gamma_L) + S_{11}\Gamma_S S_{22}\Gamma_L} \tag{A11.10}$$

As another example, we consider the nodes b_S and b_2 in Figure A11.5(b). Following the same procedure, we find that S_{21} is the only path from b_S to b_2, there is no nontouching node with respect to this path, and the sums of the first, second and higher-order loops are $1 - (S_{11}\Gamma_S + S_{21}\Gamma_L S_{12}\Gamma_S + S_{22}\Gamma_L) + S_{11}\Gamma_S S_{22}\Gamma_L$ as obtained for b_1/b_S. Therefore, we can write using Mason's rule:

$$\frac{b_2}{b_S} = \frac{S_{21}}{1 - (S_{11}\Gamma_S + S_{21}\Gamma_L S_{12}\Gamma_S + S_{22}\Gamma_L) + S_{11}\Gamma_S S_{22}\Gamma_L} \tag{A11.11}$$

which can be simplified as

$$\frac{b_2}{b_S} = \frac{S_{21}}{(1 - S_{11}\Gamma_S)(1 - S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_S\Gamma_L} \tag{A11.12}$$

A11.3.2 Derivation of Reflection Coefficients and Power Gain

We now derive the expressions for the input and output reflection coefficients and transducer power gain, or simply power gain, that we use in Chapter 10 and this chapter using signal flow graphs.

A11.3.2.1 Reflection Coefficients. The signal flow graph for the input reflection coefficient Γ_{in} designated in Figures A11.1(a) and A11.5(a) can be drawn as shown in Figure A11.6. This reflection coefficient is defined as the ratio between the two voltage waves b_1 and a_1 $\left(\Gamma_{in} = \frac{b_1}{a_1}\right)$ and hence can be conveniently derived using its signal flow graph and Mason's rule. Following the same procedure described earlier, we can find the following:

Paths from a_1 to b_1: $P_1 = S_{11}$ and $P_2 = S_{21}\Gamma_L S_{12}$.
Sum of first-order loops: $L(1) = S_{22}\Gamma_L$.
Sum of second-order loops: $L(2) = 0$.
Nontouching first-order loops to the paths: $L(1)^{(1)} = S_{22}\Gamma_L$ and $L(1)^{(2)} = 0$.

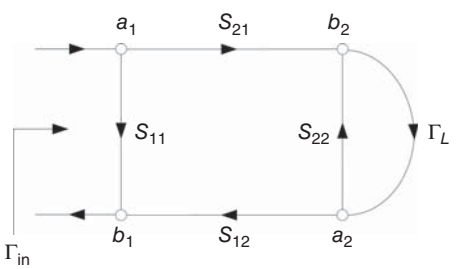


Figure A11.6. Signal flow graph for the input reflection coefficient Γ_{in} of the two-port network in Figure A11.1(b).

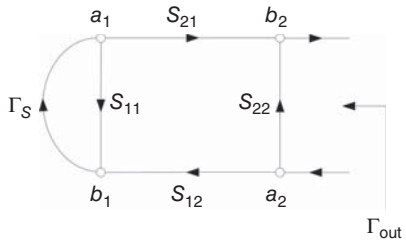


Figure A11.7. Signal flow graph for the output reflection coefficient Γ_{out} of the two-port network in Figure A11.1(b).

Nontouching second- and higher-order loops to the paths: $L(j)^{(i)} = 0, j = 2, 3, \ldots$ and $i = 1, 2$.

Applying Mason's rule, we then obtain

$$\Gamma_{in} = \frac{S_{11}(1 - S_{22}\Gamma_L) + S_{21}\Gamma_L S_{12}}{1 - S_{22}\Gamma_L} \tag{A11.13}$$

or

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} \tag{A11.14}$$

Similarly, the output reflection coefficient Γ_{out}, as seen in Figures A11.1(a) and A11.5(a), can be derived from $\Gamma_{out} = b_2/a_2$ using the signal flow graph shown in Figure A11.7 and Mason's rule as

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} \tag{A11.15}$$

or directly from (A11.14) by replacing S_{11} with S_{22} and vice versa, and Γ_L with Γ_S, recognizing the symmetry between Γ_{out} and Γ_{in} (without b_S).

A11.3.2.2 Power Gain. The transducer power gain of the two-port network as shown in Figure A11.1(b) is defined as

$$G = \frac{P_L}{P_{AVS}} \tag{A11.16}$$

where P_L and P_{AVS} are the power delivered to the load Z_L and power available from the source, respectively.

As seen from Figure A11.1(b), b_2 and a_2 are the incident and reflected voltage waves with respect to the load Z_L. The power delivered to the load is equal to the difference between the incident to and reflected power from the load, and can therefore be obtained as

$$\begin{aligned} P_L &= |b_2|^2 - |a_2|^2 \\ &= |b_2|^2(1 - |\Gamma_L|^2) \end{aligned} \tag{A11.17}$$

Substituting (A11.12) into (A11.17), we obtain

$$P_L = \frac{|b_S|^2|S_{21}|^2(1 - |\Gamma_L|^2)}{|(1 - S_{11}\Gamma_S)(1 - S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_S\Gamma_L|^2} \tag{A11.18}$$

We know from the circuit theory that the power available from a source is equal to the power delivered to a conjugately matched load, which is the maximum power delivered to a load. Considering Figure A11.1(b), this condition is met when $\Gamma_{in} = \Gamma_S^*$. Figure A11.8 shows the corresponding network and its signal flow graph. As can be recognized from Figure A11.8(a), the transmission coefficients from node b_S to node a_1', node a_1'

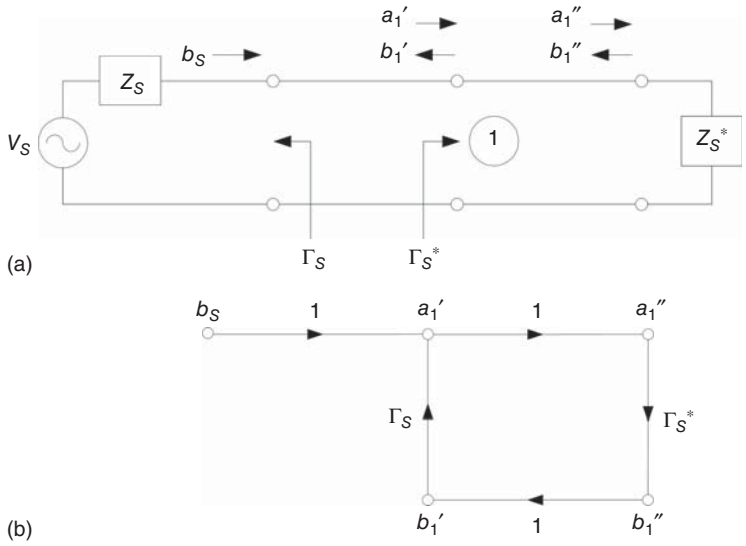


Figure A11.8. Signal flow graph for a source terminated with a conjugately matched load (a) and its signal flow graph (b).

to node a_1'', and node b_1'' to node b_1' are equal to 1, as shown in Figure A11.8(b). However, the corresponding voltage waves (e.g., b_S and a_1') are not equal due to reflections. The available power from the source is obtained as the difference between the incident power and reflected power at port 1:

$$P_{\text{AVS}} = |a_1'|^2 - |b_1'|^2 \tag{A11.19}$$

Going from node b_S to node a_1' in Figure A11.8(b), we find that $P_1 = 1$, $L(1)^{(1)} = 0$, $L(1) = \Gamma_S^*\Gamma_S$ for node a_1', and the other parameters in Mason's rule are zero, hence we can write using Mason's rule:

$$\frac{a_1'}{b_S} = \frac{1}{1 - \Gamma_S\Gamma_S^*} \tag{A11.20}$$

Now going from node b_S to node b_1', we find that $P_1 = \Gamma_S^*$, $L(1)^{(1)} = 0$, $L(1) = \Gamma_S\Gamma_S^*$ for node b_1', and the other parameters in Mason's rule are zero. We can then obtain using Mason's rule:

$$\frac{b_1'}{b_S} = \frac{\Gamma_S^*}{1 - \Gamma_S\Gamma_S^*} \tag{A11.21}$$

Substituting (A11.20) and (A11.21) into (A11.19) results in

$$P_{\text{AVS}} = \frac{|b_S|^2}{1 - |\Gamma_S|^2} \tag{A11.22}$$

The transducer power gain can now be obtained from (A11.16), (A11.18), and (A11.22) as

$$G = \frac{|S_{21}|^2(1 - |\Gamma_S|^2)(1 - |\Gamma_L|^2)}{|(1 - S_{11}\Gamma_S)(1 - S_{22}\Gamma_L) - S_{12}S_{21}\Gamma_S\Gamma_L|^2} \tag{A11.23}$$

Expanding the denominator of (A11.23) and making use of (A11.14) and (A11.15), we can rewrite the transducer power gain in the following forms:

$$G = \frac{1 - |\Gamma_S|^2}{|1 - \Gamma_{\text{in}}\Gamma_S|^2}|S_{21}|^2\frac{1 - |\Gamma_L|^2}{|1 - S_{22}\Gamma_L|^2} \tag{A11.24}$$

and

$$G = \frac{1 - |\Gamma_S|^2}{|1 - S_{11}\Gamma_S|^2} |S_{21}|^2 \frac{1 - |\Gamma_L|^2}{|1 - \Gamma_{\text{out}}\Gamma_L|^2} \tag{A11.25}$$

which are convenient for the analysis of amplifiers.

REFERENCES

1. J. K. Hunton, "Analysis of Microwave Measurement Techniques by Means of Signal Flow Graphs," *IRE Trans. Microwave Theory Tech.*, Vol. 8, No.2, pp. 206–211, Mar. 1906.
2. D. Anderson, L. Smith, and J. Gruszynski, "S-Parameter Techniques for Faster, More Accurate Network Design," Hewlett-Packard Test & Measurement Application Note 95–1, 1996, pp. 18–24.

CHAPTER 12

OSCILLATORS

Oscillators are always needed for radio frequency (RF) power generation, whether as a stand-alone component in RF systems or as an integral element of other components such as mixers. In contrast to amplifiers, which require stability, oscillators can work only in unstable circuit environments, which result in negative impedances enabling oscillation to occur. While RF power is an important parameter, in most typical applications, phase noise plays a more critical role; a low-phase noise is generally needed to produce a clean RF signal either for transmitters or receivers. Oscillators are typically designed as either single-ended or balanced with the latter more preferable due to possibly higher power and lower phase noise. In components or systems requiring RF power over a frequency range, voltage-controlled oscillators (VCO) are generally used. This chapter discusses the fundamentals of oscillators, the theory of phase noise, and the design of both single-ended and balanced oscillators for radio frequency integrated circuits (RFICs).

12.1 PRINCIPLE OF OSCILLATION

We consider a general oscillator circuit represented schematically as shown in Figure 12.1. The impedance $Z_{\text{osc}}(V,\omega) = R_{\text{osc}}(V,\omega) + jX_{\text{osc}}(V,\omega)$ represents the impedance of the indicated (active) oscillating network, consisting of active device(s) and associated circuitry, as a function of the RF voltage V produced by the active devices and frequency ω, and $\Gamma_{\text{osc}}(V,\omega)$ is the corresponding reflection coefficient. $Z_{\text{out}}(\omega) = R_{\text{out}}(\omega) + jX_{\text{out}}(\omega)$ represents the impedance looking into the indicated (passive) output network which consists of a load or output impedance, which is typically 50 Ω, and a matching circuit that transfers the load to $Z_{\text{out}}(\omega)$, and $\Gamma_{\text{out}}(\omega)$ is the corresponding reflection coefficient. There are two possibilities for the operation of the circuit shown in Figure 12.1. They are specified as given below:

Stable Operation: Stable operation occurs when

$$\text{Re}[Z_{\text{osc}}(V,\omega) + Z_{\text{out}}(\omega)] > 0 \tag{12.1}$$

or

$$R_{\text{osc}}(V,\omega) + R_{\text{out}}(\omega) > 0 \tag{12.2}$$

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

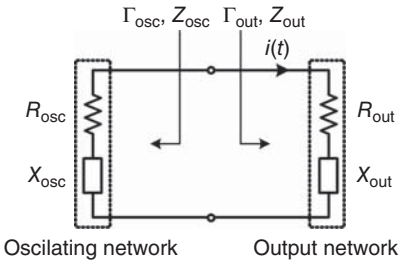


Figure 12.1. General oscillator circuit.

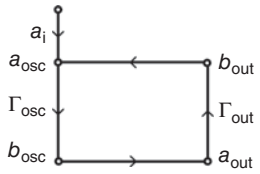


Figure 12.2. Signal flow graph for the circuit in Figure 12.1.

which is equivalent to

$$R_{\text{out}}(\omega) > |R_{\text{osc}}(V, \omega)| \tag{12.3}$$

Unstable Operation: Instability occurs when the condition stated by Eqs. (12.1), (12.2), or (12.3) is not satisfied – for instance,

$$R_{\text{osc}}(V, \omega) + R_{\text{out}}(\omega) < 0 \tag{12.4}$$

Figure 12.2 shows a signal flow graph of the circuit in Figure 12.1, following the procedure in the Appendix (Signal Flow Graphs) of Chapter 11 (Amplifiers), where a_{osc} and b_{osc} represent the (normalized) incident and reflected voltage waves with respect to the oscillating network; a_{out} and b_{out} represent the (normalized) incident and reflected voltage waves with respect to the output network; and a_i stands for (normalized) signal incident to the oscillating network. Using this signal flow graph, we can derive

$$\frac{a_{\text{out}}}{a_i} = \frac{\Gamma_{\text{osc}}(V, \omega)}{1 - \Gamma_{\text{osc}}(V, \omega)\Gamma_{\text{out}}(\omega)} \tag{12.5}$$

Under oscillation, a signal is generated and delivered to the output network even though there is no RF signal input to the circuit[1] except DC bias voltages. In order for this signal to be sustained indefinitely (i.e., $a_{\text{out}} \neq 0$), either $\Gamma_{\text{osc}}(V, \omega)$ approaches infinity, which is not feasible, or $1 - \Gamma_{\text{osc}}(V, \omega)\Gamma_{\text{out}}(\omega) = 0$. The latter condition indicates that sustained oscillation occurs when

$$\Gamma_{\text{osc}}(V, \omega)\Gamma_{\text{out}}(\omega) = 1 \tag{12.6}$$

12.1.1 Oscillation Conditions

The (RF) current flow in the circuit loop shown in Figure 12.1, in general, consists of the fundamental frequency ω and its harmonics. We assume that the circuit is operated in the sinusoidal steady state and all the harmonic signals are sufficiently suppressed by filters and, hence, are negligibly small. This current can be expressed as

$$i(t) \simeq I_o \cos \omega t \tag{12.7}$$

[1] Actually, there is always a signal entering the circuit; this signal is the inherent noise generated in the active device.

At steady state, the (maximum) current amplitude I and its corresponding voltage V become I_o and V_o, respectively, and we can write at frequency ω:

$$[Z_{\text{out}}(\omega) + Z_{\text{osc}}(V_o, \omega)]I_o = 0 \tag{12.8}$$

from which, we have

$$Z_{\text{out}}(\omega) + Z_{\text{osc}}(V_o, \omega) = 0 \tag{12.9}$$

Equation (12.9) is equivalent to

$$\begin{aligned} R_{\text{out}}(\omega) + R_{\text{osc}}(V_o, \omega) &= 0 \\ X_{\text{out}}(\omega) + X_{\text{osc}}(V_o, \omega) &= 0 \end{aligned} \tag{12.10}$$

which represents the oscillation criteria in more specific details. Since R_{out} is always positive for a passive network, (12.10a) indicates that

$$R_{\text{osc}}(V_o, \omega) < 0 \tag{12.11}$$

which implies that the oscillating network indicated in Figure 12.1 is unstable per condition (12.4). It is noted that the only condition for this instability to occur is the active device needs to be potentially unstable; it does not have to be completely unstable. The entire circuit of Figure 12.1 is unstable when

$$|R_{\text{osc}}(V_o, \omega)| > R_{\text{out}} \tag{12.12}$$

Under this condition, an oscillation can be initiated at frequency ω_o, according to (12.10b), with an initial RF voltage. Once the oscillation starts and continues to build up, the current *i(t)* will grow, leading to increasing RF voltage and current amplitudes until the steady state where they reach V_o and I_o, respectively. The steady-state oscillation occurs at ω_o when the oscillation conditions are met as

$$\begin{aligned} R_{\text{out}}(\omega_o) + R_{\text{osc}}(V_o, \omega_o) &= 0 \\ X_{\text{out}}(\omega_o) + X_{\text{osc}}(V_o, \omega_o) &= 0 \end{aligned} \tag{12.13}$$

12.1.2 Oscillation Determination

An ultimate goal in designing an oscillator is to determine its oscillation, which involves the steady-state voltage (V_o) and current (I_o) amplitude and oscillating frequency ω_o, all of which are contained in (12.13). It is recalled that the output impedance $Z_{\text{out}}(\omega_o) = R_{\text{out}}(\omega_o) + jX_{\text{out}}(\omega_o)$ is a function[2] of ω_o at a specific V_o, while the oscillating impedance $Z_{\text{osc}}(V_o, \omega_o) = R_{\text{osc}}(V_o, \omega_o) + jX_{\text{osc}}(V_o, \omega_o)$ depends on both ω_o and $V_o X_{\text{osc}}(V_o, \omega_o)$, however, usually depends weakly on V_o and, hence, the condition (12.13b) can be used to determine ω_o.

Once the oscillation condition (12.13) is established and the output network is obtained, $Z_{\text{osc}}(V_o, \omega)$ normally varies slowly with ω as compared to $Z_{\text{out}}(\omega)$, which is transformed from a load via a matching network, over a narrow frequency range due the frequency sensitivity of the matching network, and so we can neglect the frequency dependence of $Z_{\text{osc}}(V_o, \omega)$. Figure 12.3 illustrates the output and oscillating impedances. In general, condition (12.13a) is met at several frequencies while (12.13b) is not satisfied at different voltage amplitudes other than V_o since $X_{\text{osc}}(V, \omega_o)$ is amplitude-dependence; that is, $X_{\text{osc}}(V_1, \omega_o) \neq X_{\text{osc}}(V_o, \omega_o)$. Therefore,

[2]In general, Z_{out} is also indirectly dependent upon V besides ω per Eq. (12.6), since Z_{osc} is a function of both ω and V.

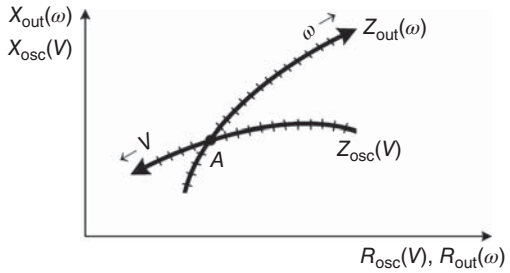


Figure 12.3. Sketches of output and oscillating impedances. Point A represents the steady-state operating point corresponding to ω_o and V_o.

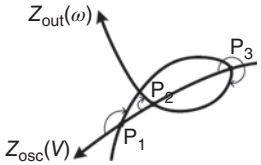


Figure 12.4. Oscillation locations.

the oscillation condition (12.13) may lead to oscillations that are not stable. The sustained steady-state oscillation is possible if and only if the following conditions are met [1]:

$$R_{\text{out}}(\omega_o) + R_{\text{osc}}(V_o, \omega_o) = 0$$
$$X_{\text{out}}(\omega_o) + X_{\text{osc}}(V_o, \omega_o) = 0$$
$$\left.\frac{\partial R_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dX_{\text{out}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} - \left.\frac{\partial X_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dR_{\text{out}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} > 0 \tag{12.14}$$

Typically, R_{out} is not a function of frequency; that is, $\frac{dR_{\text{out}}(\omega)}{\partial \omega} = 0$.

Figure 12.4 illustrates the oscillation points. Under the stable oscillation condition given in (12.14), the intersecting angle measured clockwise from the oscillating impedance line's arrow direction to the output impedance line's arrow direction must be less than 180° for stable oscillation. As examples, Figure 12.4 shows that P_1 and P_2 represent possible stable operating points with corresponding angles less than 180°, while P_3 stands for unstable operating point since the intersecting angle is greater than 180°.

12.2 FUNDAMENTALS OF OSCILLATOR DESIGN

The most crucial requirement in the oscillator design is the potential of the employed active device to have oscillation. The active device must be potentially unstable to fulfill this requirement. There are typically two approaches to ensure this to happen. One is to terminate a certain port of the device with appropriate impedance, which is essentially equivalent to providing an internal feedback in the device. Another is to provide an external feedback either to the device itself or to the circuit on which the device is embedded. This section will discuss the fundamentals of oscillator design including using proper terminations to cause oscillation and feedback oscillator basics. It is noted that typical oscillator design is based on large-signal techniques. Small-signal design techniques, while may result in reasonable prediction of the frequency of oscillation, give less accurate results for other parameters such as power and phase noise, and hence are not preferred. However, small-signal techniques can be useful for cases where large-signal information, such as *S*-parameters or nonlinear models, is not available.

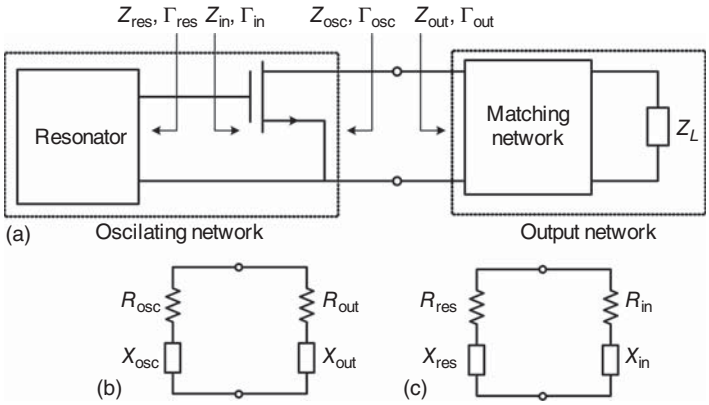


Figure 12.5. Basic oscillator topology (a) and its equivalent circuits (b) and (c).

12.2.1 Basic Oscillators

We begin by considering a very basic yet general oscillator topology as shown in Figure 12.5(a), which consists of a common-source MOSFET, a resonator (or resonant or tank circuit), a matching network, and a load (Z_L). A reactive network can also be used to terminate the source terminal to improve the oscillation. The MOSFET is assumed to be potentially unstable and hence can have possible oscillations with proper terminations. The resonator mainly controls the oscillation frequency and the matching network is used to transform the impedance Z_{out} to the load which is either a typical 50 Ω or the input impedance of a successive component such as amplifier. The circuit in Figure 12.5(a), specified with the oscillating and output networks, is electrically equivalent to the circuit shown in Figure 12.1. From Figure 12.5(a), we can identify the function of the oscillating and output networks. The resonator in the oscillating network is used to cause $|\Gamma_{osc}| > 1$ necessary for oscillation to begin. The output network sets the oscillation frequency based on the resonator or resonant circuit and delivers the generated signal to the output load according to the oscillation condition (12.13). Figure 12.5(b) and (c) shows two equivalent circuits of the basic oscillator shown in Figure 12.5(a) referenced to the interface between the output network and the MOSFET and between the resonator and the MOSFET, respectively.

Under oscillation and considering Figure 12.5(a) or (b), we obtain according to (12.6):

$$\Gamma_{out}(\omega) = \frac{1}{\Gamma_{osc}(V, \omega)} \tag{12.15}$$

where Γ_{osc} is obtained from Eq. (11.8) of Chapter 11 as

$$\Gamma_{osc} = S_{22} + \frac{S_{12}S_{21}\Gamma_{res}}{1 - S_{11}\Gamma_{res}} \tag{12.16}$$

with $S_{ij}(i,\ j = 1,\ 2)$ being the S-parameters of the MOSFET. Substituting Γ_{osc} in (12.16) into (12.15) gives

$$\Gamma_{out} = \frac{1 - S_{11}\Gamma_{res}}{S_{22} - \Delta\Gamma_{res}} \tag{12.17}$$

where

$$\Delta = S_{11}S_{22} - S_{12}S_{21} \tag{12.18}$$

as given in Eq. (10.3) of Chapter 10 (Stability). We can obtain from (12.17)

$$\Gamma_{res} = \frac{1 - S_{22}\Gamma_{out}}{S_{11} - \Delta\Gamma_{out}} \tag{12.19}$$

Γ_{in} is given in Eq. (11.7) of Chapter 11 as

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_{out}}{1 - S_{22}\Gamma_{out}} = \frac{S_{11} - \Delta\Gamma_{out}}{1 - S_{22}\Gamma_{out}} \tag{12.20}$$

Taking the product between (12.19) and (12.20) gives

$$\Gamma_{res}\Gamma_{in} = 1 \tag{12.21}$$

which implies that the equivalent circuit shown in Figure 12.5(c) also oscillates. Similarly, assuming oscillation for the circuit shown in Figure 12.5(c), it can also be proved that the circuit shown in Figure 12.5(a) or (b) is oscillated as well. In general, it can be proved that, for oscillators, oscillation occurs in the oscillator's equivalent circuit at any point of the oscillator. Another word, when a circuit is oscillated, all locations in the circuit oscillate simultaneously, which is expected as the entire circuit is unstable.

We assume that the large-signal *S*-parameters of the MOSFET are available either through measurements, calculations, or large-signal device models. Under appropriate bias conditions, $|R_{osc}(V, \omega_o)| > R_{out}(\omega)$ corresponding to $R_{osc}(V, \omega_o) + R_{out}(\omega) < 0$, the circuit oscillates and the oscillation begins to build up from a noise level. The RF voltage and current of the generating signal, once initiated, continue to grow until they reach the steady state. It is noted that, while oscillation can be possible only if the device is potentially unstable and the total loop resistance $[R_{osc}(V, \omega_o) + R_{out}(\omega)]$ is negative, the ability of a transistor to begin oscillation in a given circuit is determined by the noise level. As the oscillation grows, the signal's voltage V and power increase, and the resistance $R_{osc}(V, \omega_o)$ reduces until the device is operated under the steady state, at which $|R_{osc}(V, \omega_o)| = R_{out}(\omega)$ and $|X_{osc}(V, \omega_o)| = -X_{out}(\omega)$, where $V = V_o$, according to (12.14).

The oscillator design begins with choosing a potentially unstable MOSFET. It is recalled from Chapter 10 that a potential instability for devices occurs when $|S_{11}| > 1$ and/or $|S_{22}| > 1$, or either K<1 or $|\Delta| > 1$ when $|S_{11}| < 1$ and $|S_{22}| < 1$, where K is given as

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{12}S_{21}|} \tag{12.22}$$

A design procedure for oscillators described in Figure 12.5(a) is briefly outlined below.

1. Check the device's stability to make sure it is potentially unstable at the interested frequency.
2. Plot the stability circle in the (input) oscillating-network plane on a Smith chart with the center and radius given respectively in Eqs. (10.19) and (10.20) of Chapter 10:

$$C_S = (C_{SR}, C_{SI}) = \frac{(S_{11} - \Delta S_{22}^*)^*}{|S_{11}|^2 - |\Delta|^2} \tag{12.23}$$

and

$$R_S = \left| \frac{S_{12}S_{21}}{|S_{11}|^2 - |\Delta|^2} \right| \tag{12.24}$$

where Δ is given in (12.18).

3. Choose Γ_{res} or Z_{res} inside the Smith chart and within the unstable region of the stability circle. Make sure it is sufficiently far away from the stability circle to avoid potential shift into the stable region due to device's process variations.
4. Calculate Γ_{osc} using (12.16) with the device's large-signal *S*-parameters and chosen value for Γ_{res}. Make sure $|\Gamma_{osc}| > 1$ to ensure possible oscillation. Ideally, it is also desired that $|\Gamma_{osc}| < 1$ at frequencies other than the desired oscillation frequency to guarantee out-of-band stability to avoid spurious oscillations.

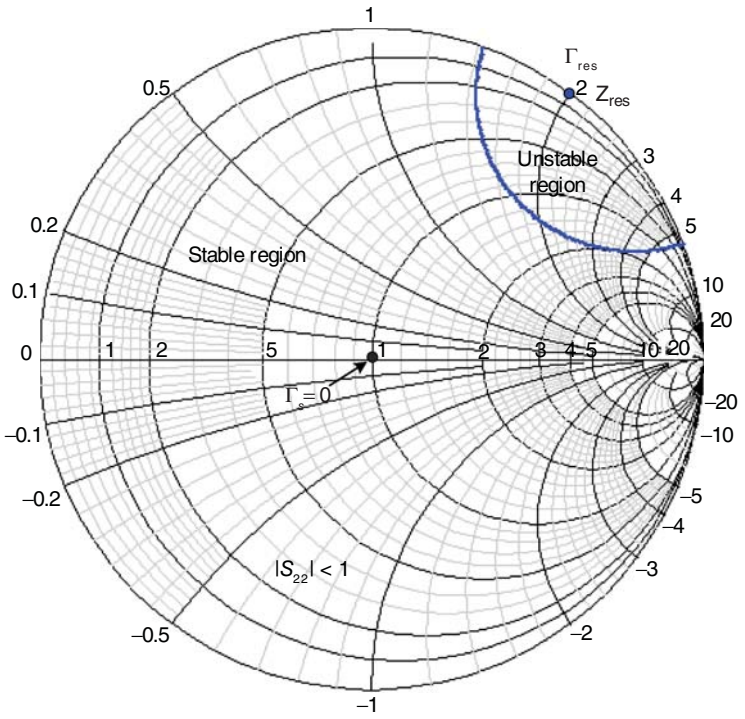


Figure 12.6. Stability circle showing the unstable region and selected Γ_{res}.

5. Choose the output impedance Z_{out} according to the steady-state oscillation condition: $R_{out} = -R_{osc}(V_o, \omega_o)$ and $X_{out} = -X_{osc}(V_o, \omega_o)$. Under small-signal operation, R_{out} is typically chosen as 1/3 to 1/2 of $|R_{osc}(V, \omega_o)|$, where $V < V_o$ is the small-signal voltage, to ensure oscillation.
6. Design a matching network to transfer Z_{out} to a load (typically 50 Ω or the input impedance of a subsequent component.)
7. Analyze and optimize the designed oscillator using a nonlinear analysis method such as that from a commercially available computer program to achieve desired performance for parameters such as output power, frequency spectrum, and phase noise.

As an example of oscillator design, we consider a MOSFET whose large-signal *S*-parameters at 5 GHz are:

$$S_{11} = 0.9702\angle -27.45^{\circ} \quad S_{12} = 0.0681\angle 69.56^{\circ} \quad S_{21} = 3.0771\angle 155.75^{\circ} \quad S_{22} = 0.7366\angle -27.22^{\circ}$$

Following the foregoing design procedure, we first check the MOSFET's stability at 5 GHz. Since $K = 0.043 < 1$, the device is potentially unstable. Figure 12.6 shows the stability circle in the (input) oscillating-network plane, which is centered at $C_S = 1.126\angle 45.085^{\circ}$ and has radius $R_S = 0.476$. Since $|S_{22}| < 1$, the unstable region is inside the Smith chart and stability circle. $\Gamma_{res} = 0.6 + j0.8$, corresponding to $Z_{res} = j100\,\Omega$ (assuming 50-Ω normalization or system), inside the unstable region is selected. An inductor whose value is 3.18 nH can be used to represent the chosen Z_{res} at 5 GHz. $\Gamma_{osc} = 1.122 - j0.4$ is then calculated, from which $Z_{osc} = -(120 + j228.78)\,\Omega$ is determined. Since $|\Gamma_{osc}| > 1$, an oscillation is possible. The output impedance Z_{out} is determined as $R_{out} = -R_{osc} = 120\,\Omega$ and $X_{out} = -X_{osc} = j228.78\,\Omega$. A matching network is finally designed using Smith chart to transform Z_{out} to a 50-Ω load. Figure 12.7 shows the schematic of the designed oscillator with respect to a 50-Ω load.

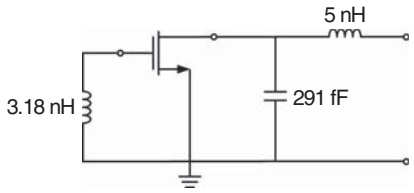


Figure 12.7. Oscillator schematic.

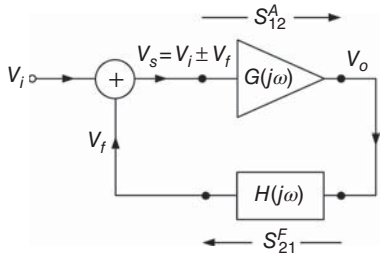


Figure 12.8. Basic feedback oscillator block diagram. V_o is the RF voltage of the oscillator, but not the actual RF voltage feeding a subsequent component; a power divider or coupler is typically needed at the output of the amplifying network to extract an RF output signal with voltage $V_{\text{out}} < V_o$.

12.2.2 Feedback Oscillators

Many forms of oscillators that can be implemented in CMOS RFIC are based on "feedback" topologies – for example, Colpitts, Hartley, Clapp, Seiler, Pierce [2] and the balanced or differential configurations of these. Even a (non-feedback) oscillator, whose physical appearance does not conform to a feedback structure, can be viewed electrically as a feedback topology due to the feedback mechanisms occurred internally in the active device itself, such as the gate-drain capacitance, and externally within the oscillator circuit such as coupling effects. Effectively, oscillators, feedback or no-feedback, can be electrically considered equivalent to some feedback topology. Feedback oscillator topology hence represents the foundation for many CMOS RFIC oscillators.

Figure 12.8 shows a basic feedback oscillator block diagram, which consists of a summation or combining (+) network, an amplifying network (which, in the simplest form, can be a MOSFET) with frequency-dependent voltage gain $G(j\omega)$, which is equal to the S_{21} parameter of the amplifying network $S_{21}^{A}(\omega)$, a feedback network with frequency-dependent voltage transfer function $H(j\omega)$, which is equal to the insertion-loss parameter S_{21} of the feedback network $S_{21}^{F}(\omega)$. $G(j\omega)$ represents the gain between the input and output when there is no feedback and, hence, is called the open-loop gain. The amplifier operates over a bandwidth containing the oscillation frequency ω_o of the oscillator. The feedback network can be lumped elements (L, C or a combination of L and C), a resonator (either lumped element, transmission line, or another distributed resonator.) The positive and negative feedback correspond to $V_s = V_i + V_f$ and $V_s = V_i - V_f$, respectively. We obtain from Figure 12.8

$$V_o = G(j\omega)[V_i + V_f] \tag{12.25}$$

assuming positive feedback, and

$$V_f = H(j\omega)V_o \tag{12.26}$$

The RF output voltage with the feedback in place can be derived from (12.25) and (12.26) as

$$V_o = \frac{V_i G(j\omega)}{1 - H(j\omega)G(j\omega)} \tag{12.27}$$

Equation (12.27) suggests that there will be no RF signal (V_o) if there is no input signal (V_i), implying that oscillation would never be established. This, however, happens only under perfect circuit environment and ideal circuit elements including solid-state devices, where noise never exists. In practice, noise always exists in circuits across the electromagnetic (EM) spectrum; the most pronounced one perhaps come from the solid-state devices themselves. So, in reality, there is always noise that enters the feedback oscillator described in Figure 12.8 which, although not real RF signal, acts as an RF input signal driving the amplifying network into possible oscillation. Therefore, under proper bias conditions, a feedback oscillator begins to oscillate from the RF noise level without a real RF input signal.[3] The noise is needed only initially to start oscillation; once the oscillation is initiated, an RF signal is generated and takes place of the noise as the input signal to sustain the oscillation and push the oscillation further into steady state. The oscillator's output voltage increases as the oscillation builds up and finally reaches an output voltage $V_o \neq 0$ at the steady state. Since the noise level is presumably very small, this RF signal generation phenomenon is possible only if either $G(j\omega)$ is very large (approaching infinity which is not feasible), or $1 - H(j\omega)G(j\omega)$ becomes extremely small, approaching zero, which is achievable. Let ω_o be the frequency at which the latter condition occurs; that is:

$$1 - H(j\omega_o)G(j\omega_o) = 0 \tag{12.28}$$

This condition is equivalent to

$$|H(j\omega_o)G(j\omega_o)| = 1 \tag{12.29}$$

and

$$\angle H(j\omega_o) + \angle G(j\omega_o) = 0 \tag{12.30}$$

or, in terms of S-parameters:

$$|S_{21}^A||S_{21}^F| = 1 \tag{12.31}$$

and

$$\angle S_{21}^A + \angle S_{21}^F = 0 \tag{12.32}$$

Condition (12.28) is referred to as the Nyquist criterion that specifies that condition for oscillation. This condition is equivalent to the Barkhausen criterion which states that, a circuit having a closed-loop voltage gain associated with positive feedback of

$$G_{L,P}(j\omega) = \frac{V_o}{V_i} = \frac{G(j\omega)}{1 - H(j\omega)G(j\omega)} \tag{12.33}$$

would oscillate if $H(j\omega)G(j\omega) \geq 1$.

Condition (12.28), or (12.29), and (12.30), states that in a positive feedback system or circuit, if the total gain of the individual open-loop gains, namely the open-loop gain, $H(j\omega)G(j\omega)$ has a unity gain and a total phase (or the phase shift around the close loop) of zero degree or a multiple of 360° at frequency ω_o then the system or circuit would oscillate at that frequency. This oscillation condition is mathematically the necessary and sufficient condition to guarantee oscillation. However, oscillation under this condition occurs only theoretically under the perfect world. In practice, non-ideal circuit environment, condition, and circuit elements prevent the oscillation from happening when the open-loop gain is exactly one and/or the total phase is exactly zero or multiple of 360°. Practical feedback oscillators need open-loop gain greater than 1 in order to initiate oscillation. Having an open-loop gain greater than one is also what RF designers want since it leads directly to higher output power as indicated in (12.27). It is recalled, as noted in Figure 12.8, that the actual RF output voltage produced by the oscillator that can be used by other circuits is smaller than V_o due to a

[3] Although, the conclusion for oscillation is drawn based on a feedback oscillator with a clear feedback loop, this phenomenon also applies to any other kinds of oscillators, such as those we considered earlier, since there is always feedback in oscillators, whether internally or externally, or intentionally or unintentionally, as mentioned earlier, and so there is indeed noise entering oscillators.

power divider or coupler used at the output of the amplifying network to extract that output signal. If the open-loop gain is less than 1 at the frequency where the total phase is 0 or a multiple of 360°, the system or circuit is stable. If the open-loop gain is greater than 1, the system or circuit is unstable, implying possible oscillation.

For negative feedback, the closed-loop gain is

$$G_{L,N}(j\omega) = \frac{V_o}{V_i} = \frac{G(j\omega)}{1 + H(j\omega)G(j\omega)} \tag{12.34}$$

which results in oscillation at ω_o when

$$1 + H(j\omega_o)G(j\omega_o) = 0 \tag{12.35}$$

or

$$|H(j\omega_o)||G(j\omega_o)| = 1 \tag{12.36}$$

and

$$\angle H(j\omega_o) + \angle G(j\omega_o) = 180^{\text{o}} \tag{12.37}$$

or, in terms of *S*-parameters:

$$|S_{21}^A||S_{21}^F| = 1 \tag{12.38}$$

and

$$\angle S_{21}^A + \angle S_{21}^F = 180^{\text{o}} \tag{12.39}$$

Equations (12.36)–(12.38), or (12.39) indicate that the circuit would oscillate at ω_o when the magnitude of the open-loop gain is equal to 1 and the total phase shift around the loop is equal to 180°.

Since the feedback network is passive, $|S_{21}^F| < 1$, hence leading to $|S_{21}^A| > 1$ which signifies, from (12.31) and (12.39), that the amplifier must produce gain (>0 dB) in order for the oscillation to be established. In terms of signal propagation, at the oscillation frequency, the propagating time of the generated RF signal around the loop for the positive feedback must be an integer of the oscillation period or the equivalent electrical length of the loop must be 0 or a multiple of 360°. For the negative feedback, the propagating time must be an odd-integer of half of the oscillation period or the equivalent electrical length of the loop must be an odd-integer of 180°.

12.2.2.1 Feedback Oscillator Analysis. For the convenience of analysis and illustration of different feedback oscillator configurations, we consider a general feedback oscillator as shown in Figure 12.9. The feedback network consists of three elements (Y_1, Y_2, and Y_3), each of them being capacitor or inductor, or a combination of these, and produces a frequency-dependent transfer function $H(j\omega)$ as noted in Figure 12.8. In lieu of lumped elements, distributed elements may also be used in the feedback network. The feedback network may also be configured to work as a lumped-element RLC or distributed resonator. Under ideal conditions, the feedback network contains all lossless elements and accordingly has an infinite quality factor (Q). Practical feedback network, however, has losses due to real inductors, capacitors, and distributed elements, resulting in finite Q which affects the phase noise of the oscillator.

Different feedback configurations, and hence feedback oscillators, can be formed by setting V_1, V_2, V_3, and/or V_4 to obtain different grounded terminals for the MOSFET as well as different MOSFET terminals between which the feedback network is connected. The common-source with the feedback between the gate and drain terminals and the load (Z_L) at the drain, as shown in Figure 12.10, is specified by

$$\begin{aligned} V_2 &= 0 \\ V_3 &= V_4 \end{aligned} \tag{12.40}$$

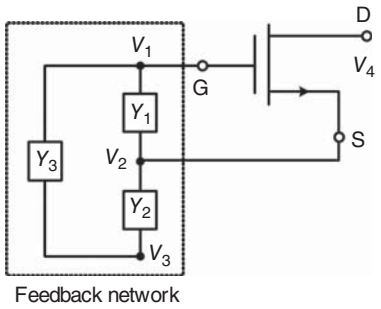


Figure 12.9. General feedback oscillator.

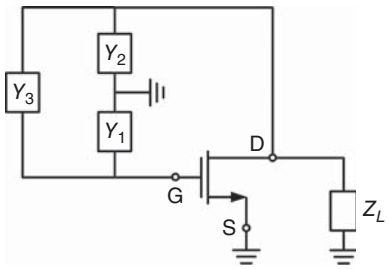


Figure 12.10. Common-source feedback oscillator.

The common-gate having the feedback between the source and drain terminals with the load at the drain is obtained using

$$V_1 = 0$$
$$V_3 = V_4 \tag{12.41}$$

The common-drain with the feedback between the gate and source terminals and the load at the source is specified with

$$V_4 = 0 \tag{12.42}$$

Moreover, different feedback oscillator topologies – for instance the Colpitts, Clapp, Hartley, Seiler and Pierce oscillator topologies – can be formed by using different types of the feedback network.

Assuming the MOSFET can be represented by a simple unilateral device model as shown in Figure 12.11, we can draw an equivalent circuit for the general feedback oscillator (Figure 12.9) in Figure 12.12. Applying Kirchoff's current law to nodes V_1, V_2, V_3, and V_4 of the circuit in Figure 12.12, we get

$$(Y_1 + Y_3 + G_{gs})V_1 - (Y_1 + G_{gs})V_2 - Y_3V_3 = 0 \tag{12.43}$$

$$(Y_1 + G_{gs} + g_m)V_1 - (Y_1 + Y_2 + G_{gs} + G_{ds} + g_m)V_2 + Y_2V_3 + G_{ds}V_4 = 0 \tag{12.44}$$

$$Y_3V_1 + Y_2V_2 - (Y_2 + Y_3)V_3 = 0 \tag{12.45}$$

and

$$g_mV_1 - (g_m + G_{ds})V_2 + G_{ds}V_4 = 0 \tag{12.46}$$

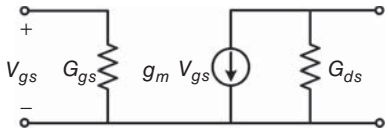


Figure 12.11. A simple unilateral MOSFET model. g_m: transconductance, G_{gs}: gate-source conductance, and G_{ds}: drain-source conductance.

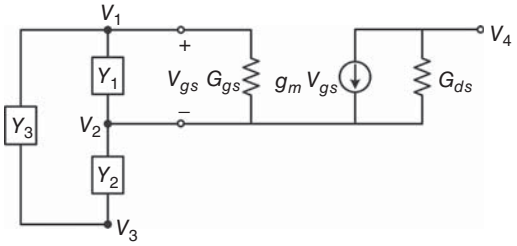


Figure 12.12. Equivalent circuit of the general feedback oscillator shown in Figure 12.9.

respectively, which can be written in matrix form as

$$\begin{bmatrix} Y_1+Y_3+G_{gs} & -(Y_1+G_{gs}) & -Y_3 & 0 \\ -(Y_1+G_{gs}+g_m) & Y_1+Y_2+G_{gs}+G_{ds}+g_m & -Y_2 & -G_{ds} \\ -Y_3 & -Y_2 & Y_2+Y_3 & 0 \\ g_m & -(G_{ds}+g_m) & 0 & G_{ds} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = 0 \tag{12.47}$$

Equation (12.47) can be used for different voltage settings for different feedback and oscillator configurations according to (12.40)–(12.42).

The common-source feedback oscillator as shown in Figure 12.10 can be represented by the equivalent circuit shown in Figure 12.13. Eq. (12.47) can be rewritten for this circuit by not considering node V_2 and combining nodes V_3 and V_4, corresponding to $V_2 = 0$ and $V_3 = V_4 \equiv V$, respectively, as

$$\begin{bmatrix} Y_1+Y_3+G_{gs} & -Y_3 \\ g_m & -Y_3 \end{bmatrix} \begin{bmatrix} V_1 \\ V \end{bmatrix} = 0 \tag{12.48}$$

where G_{gs} is neglected due to its typically small value. This equation, containing the solutions for V_1 and V, can also be derived directly from Figure 12.13 considering nodes V_1 and V. Under oscillation, RF signals are generated and hence V_1 and V must be different from zero. Accordingly, the determinant in (12.48) must be equal to zero:

$$\begin{vmatrix} Y_1+Y_3+G_{gs} & -Y_3 \\ g_m & -Y_3 \end{vmatrix} = 0 \tag{12.49}$$

which leads to

$$Y_2(Y_1+Y_3+G_{gs}) + Y_3(Y_1+G_{gs}+g_m) = 0 \tag{12.50}$$

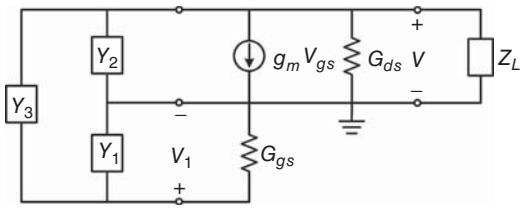


Figure 12.13. Equivalent circuit of the common-source feedback oscillator.

TABLE 12.1. Possible Elements for Y_1, Y_2, and Y_3

Y_1	Y_2	Y_3	Feedback oscillator
Inductor	Inductor	Capacitor	Hartley
Capacitor	Capacitor	Inductor	Colpitts

signifying the condition that the constituent parameters must satisfy. The transconductance g_m and gate-source conductance G_{gs} are both real and positive, and so (12.50) is not valid if all of the remaining (passive) elements Y_1, Y_2, and Y_3 are also real and positive. Since these elements are passive, they must represent either inductors and/or capacitors. Assume they are ideal elements and hence lossless, we can then let $Y_i = jB_i$, where $i = 1, 2, 3$, and, upon substituting into (12.50) and setting the real and imaginary parts to zero, we obtain

$$\frac{1}{B_1} + \frac{1}{B_2} + \frac{1}{B_3} = 0 \tag{12.51}$$

$$\frac{1}{B_3} + \left(1 + \frac{g_m}{G_{gs}}\right)\frac{1}{B_2} = 0 \tag{12.52}$$

Solving (12.51) and (12.52) gives

$$\frac{B_2}{B_1} = \frac{g_m}{G_{gs}} \tag{12.53}$$

and

$$\frac{B_2}{B_3} = -\left(1 + \frac{g_m}{G_{gs}}\right) \tag{12.54}$$

Since g_m and G_{gs} are real and positive, (12.53) and (12.54) indicate respectively that the elements represented by Y_1 and Y_2 must be of the same type, either all capacitors or all inductors, and the element associated with Y_3 must be of the opposite type, either inductor or capacitor, respectively. Table 12.1 shows the possibilities for Y_1, Y_2, and Y_3 and the resultant feedback oscillator topologies.

Colpitts Oscillator

The Colpitts oscillator corresponds to Y_1, Y_2, and Y_3 being inductive and capacitive, respectively. Figure 12.14 shows the schematic of the common-source Colpitts oscillator. Applying (12.51) to this oscillator, we obtain at the oscillation frequency ω_o:

$$\omega_o L - \frac{1}{\omega_o}\left(\frac{C_1 + C_2}{C_1 C_2}\right) \tag{12.55}$$

from which, we can derive the oscillation frequency as

$$\omega_o = \frac{1}{\sqrt{LC}} \tag{12.56}$$

where C is the series combination of C_1 and C_2:

$$C = \frac{C_1 C_2}{C_1 + C_2} \tag{12.57}$$

The relation between the feedback capacitors C_1, C_2, and the MOSFET's parameters is obtained from (12.53) as

$$\frac{C_2}{C_1} = \frac{g_m}{G_{gs}} \tag{12.58}$$

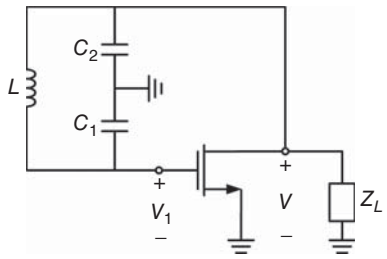


Figure 12.14. Schematic of a common-source Colpitts oscillator.

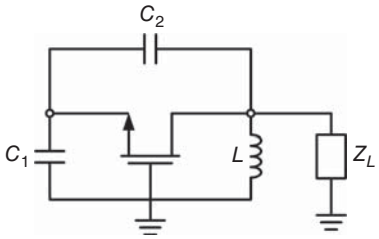


Figure 12.15. Schematic of a common-gate Colpitts oscillator.

Equation (12.57) can be rewritten upon using (12.58) as

$$C = C_1\left(1 + \frac{g_m}{G_{gs}}\right) \tag{12.59}$$

or

$$C = C_2\left(1 + \frac{G_{gs}}{g_m}\right) \tag{12.60}$$

Equations (12.57), (12.59), or (12.60), together with (12.56), shows that the oscillation frequency at a given MOSFET's bias condition can be tuned by using variable capacitance such as that provided by a varactor for either C_1, C_2, or both. For fixed C_1 and C_2, although the frequency tuning cannot be theoretically achieved by changing the MOSFET's bias voltages to vary g_m and G_{gs}, as the oscillation frequency is fixed for C_1 and C_2 (with given L) per (12.56), a narrow tuning range can be achieved in practice by varying the bias voltages. The voltage transfer function $H(j\omega)$ of the feedback network can be derived from Figure 12.14 as

$$H(j\omega) = \frac{V_1}{V} = \frac{B_3}{B_1 + B_3} = -\frac{B_{2\prime}}{B_1} = -\frac{C_{2\prime}}{C_1} \tag{12.61}$$

making use of (12.51). This transfer function also relates to the device's parameters upon using (12.58) as

$$H(j\omega) = -\frac{g_m}{G_{gs}} \tag{12.62}$$

The Colpitts oscillator can also be formed by grounding the gate or drain terminal. Figure 12.15 shows the schematic of the common-gate Colpitts oscillator with feedback between the source and drain.

Hartley Oscillator

Figure 12.16 shows a common-source Hartley oscillator with Y_1, Y_2, and Y_3 being inductors and capacitor, respectively. Utilizing (12.51) and (12.53) as for the Colpitts oscillator, we can derive the oscillation frequency

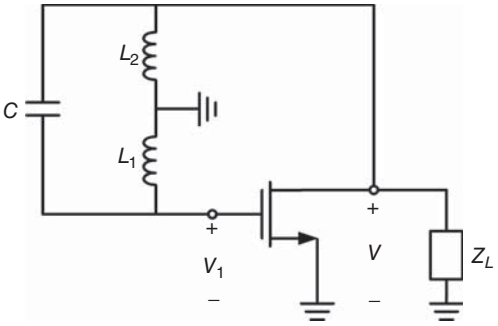


Figure 12.16. Schematic of a common-source Hartley oscillator.

and the relation between the inductors L_1 and L_2 as

$$\omega_o = \frac{1}{\sqrt{LC}} \tag{12.63}$$

where

$$L = L_1 + L_2 \tag{12.64}$$

and

$$\frac{L_1}{L_2} = \frac{g_m}{G_{gs}} \tag{12.65}$$

Equation (12.64) can be rewritten upon using (12.65) as

$$L = L_1 \left(1 + \frac{G_{gs}}{g_m}\right) \tag{12.66}$$

or

$$L = L_2 \left(1 + \frac{g_m}{G_{gs}}\right) \tag{12.67}$$

Equations (12.64), (12.66), or (12.67), together with (12.63), shows that the oscillation frequency at a given MOSFET's bias condition can be tuned by using variable capacitance, such as that provided by a varactor, for C or variable inductance, such as that provided by active inductors, for L_1 and/or L_2. Using active inductors, however, is not very desirable due to possible increase in noise. Furthermore, in practice, a narrow tuning range can also be achieved by varying the bias voltages applied to the MOSFET for fixed C_1, C_2, and L. The voltage transfer function $H(j\omega)$ of the feedback network can be derived from Figure 12.16, making use of (12.51), as

$$H(j\omega) = \frac{V_1}{V} = -\frac{L_{1'}}{L_2} \tag{12.68}$$

which becomes upon using (12.65):

$$H(j\omega) = -\frac{g_m}{G_{gs}} \tag{12.69}$$

which is identical the voltage transfer function of the Colpitts' feedback network given in (12.62) as expected. Figure 12.17 shows another configuration of the Hartley oscillator with common-gate MOSFET and feedback between the source and drain.

The foregoing analyses employ a very simple unilateral model for MOSFETs for a convenient illustration of the essence of the feedback oscillators. The derived equations, although not very accurate, can serve as

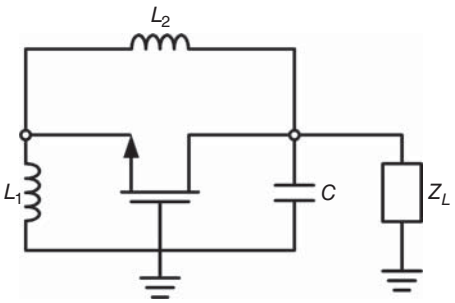


Figure 12.17. Schematic of a common-gate Hartley oscillator.

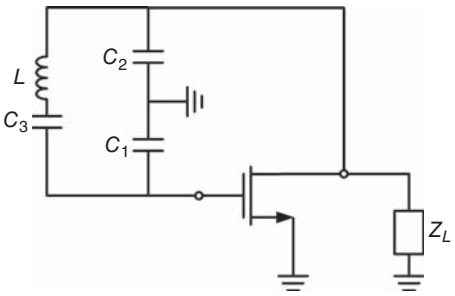


Figure 12.18. Schematic of a common-source Clapp oscillator.

the starting equations to determine initial values of the feedback elements. More complete device models as well as accurate models for the feedback elements such as inductors and capacitors are essential for accurate oscillator design.

Other Feedback Oscillators

The feedback oscillators used in the foregoing analyses employ three separate branches, each having a single element, for the feedback network. Additional branches and/or more elements in each branch, however, can also be used to form other feedback oscillators that may provide more flexibility and/or performance improvement. The analysis for more than three branches can be done similar to the previous analysis used for the three-branch feedback network. When adding additional elements into the three branches, it is reminded that the total admittance in each branch (Y_1, Y_2, and Y_3) must be so that Y_1, Y_2 are of the same type and Y_3 is of the opposite type, as stated earlier. For instance, an additional capacitor and inductor can be added into the L and C_1, C_2 branches of the Colpitts oscillator shown in Figures 12.14 and 12.15, respectively, provided that the resultant Y_1, Y_2 are of the same kind and Y_3 is of the opposite kind. In fact, when an additional capacitor is added in series with L in the L-branch of the Colpitts oscillator, the resultant oscillator is known as the Clapp oscillator. The Clapp oscillators are shown in Figures 12.18 and 12.19 for the common source and gate, respectively.

Figures 12.20 and 12.21 show two additional feedback oscillators, namely Seiler and Pierce oscillator, respectively. These oscillators can be analyzed using the approach discussed earlier.

12.3 PHASE NOISE

While power is an important parameter for oscillators, perhaps an even more critical parameter is phase noise. The phase noise degrades the performance of communication, sensing and radar systems. In fact, it is considered one of the most important factors which limit the quality of systems. In most applications, it is important that the phase noise of oscillators, or any signal-generation sources, be kept at sufficiently low level even at the sacrifice of their output power.

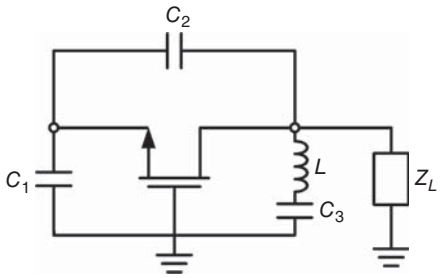


Figure 12.19. Schematic of a common-gate Clapp oscillator.

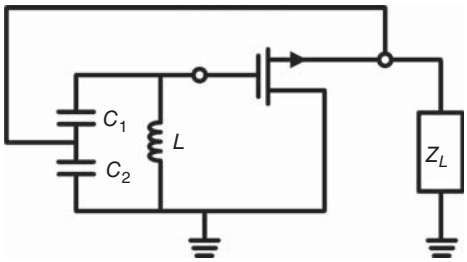


Figure 12.20. Seiler oscillator.

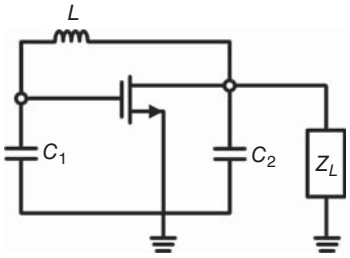


Figure 12.21. Pierce oscillator.

12.3.1 Fundamentals of Phase Noise

To effectively describe what phase noise is in oscillators or any signal-generation components, we begin by considering an ideal (non-practical) oscillator. Under a sinusoidal steady-state condition, this oscillator would produce a perfect sinusoidal signal whose voltage can be expressed as

$$V(t) = A\sin(2\pi f_o t + \phi_o) \tag{12.70}$$

where $f_o = \frac{\omega_o}{2\pi}$ and A are the frequency and maximum amplitude of the signal, respectively, and ϕ_o is an arbitrary phase constant. $(2\pi f_o t + \phi_o)$ represents the total phase of the signal that grows linearly with frequency at any instant; this phase is "pure" without any fluctuation. The signal's frequency spectrum, obtained by taking a Fourier transform of (12.70), gives an impulse or Dirac delta function $\delta(f_o)$, which exists only at f_o, signifying that the signal in the frequency domain, just like that in the time domain, is "pure" having no distortion or noise in both amplitude and phase or, another word, a perfect spectral purity.

Practical oscillators, just like any other RF components, are non-ideal, always having noise that modulates the signal and causes fluctuations in its amplitude and phase, leading to imperfect spectral purity. There are two kinds of noise in real signals: amplitude noise and phase noise. The output voltage of practical oscillators, neglecting the inter-modulation of the amplitude and phase noise due to the nonlinearity of the oscillators or subsequent components, can be expressed, taking into account the time-dependent noise-induced variations

in amplitude and phase, as

$$V(t) = [1 + a(t)]A \sin[2\pi f_o t + \phi_o + \phi_n(t)] \tag{12.71}$$

where $a(t)$ represents the amplitude-modulation (AM) noise (or simply amplitude noise or fluctuation), $[1 + a(t)]A$ is the total (noise-modulated) maximum amplitude, and $\phi_n(t)$ represents the phase-modulation (PM) noise (or simply phase noise or fluctuation). $\phi_n(t)$ describes a deviation of the phase from the original phase $(2\pi f_o t + \phi_o)$ of the pure signal. The effect of the AM noise on an RF signal source is generally much smaller than that of the PM noise, particularly at frequencies close to the carrier. It is noted that at frequencies sufficiently far from the carrier, the spectrum of an RF signal can be dominated by the amplitude noise. Additionally, the AM noise of a signal source used in circuits can be well suppressed using properly designed balanced structures, for instance, in balanced mixers. Furthermore, well-designed, high-quality oscillators usually have very stable amplitude. Therefore, it is reasonable to consider the AM noise $a(t)$, and hence the total signal amplitude, as constant over time, thereby neglecting the contribution and/or effect of the AM noise to a system's operation and/or its measured results. The noise-induced phase fluctuation is much more random in nature as compared to that of amplitude and is very uncertain. As indicated in (12.71), the phase noise $\phi_n(t)$ dictates the accuracy of the signal's total phase $[2\pi f_o t + \phi_o + \phi_n(t)]$ at any given time, and is therefore an important system parameter in many applications where both amplitude and phase information is needed. The double side-band power spectrum of (12.71) can be obtained from its Fourier transform and is a superposition of the carrier power and the power spectrum of the amplitude and phase noise. That is,

$$S(f) = \frac{A^2}{2}\delta(f - f_o) + S_A(f - f_o) + S_\phi(f - f_o) \tag{12.72}$$

where δ is again the Dirac delta function, and $S_A(f - f_o)$ and $S_\phi(f - f_o)$ represent the amplitude and phase noise power spectra, respectively. At frequencies, where external white noise is predominant, as will be seen for large offset frequencies, the power spectrum has equal contribution from the amplitude and phase noise.

The signal, as can be seen from (12.71), is no longer a pure sinusoid and its frequency spectrum, as can be seen from (12.72), is thus no longer a Dirac impulse, but exhibits sidebands around the oscillator's frequency as shown in Figure 12.22. The oscillator's frequency spectrum in Figure 12.22 also describes the phase noise spectrum of the oscillator in the frequency domain. As can be seen, the phase noise effectively causes the frequency of the oscillator to fluctuate around its nominal value f_o, hence causing an uncertainty in frequency at any time. This fluctuation in frequency is viewed as "frequency noise" and is directly related to phase noise. Hence, phase noise and frequency noise can be used to indicate the same thing.

The phase noise, as can be recognized by now, indeed directly dictates the (frequency) stability of oscillators. Considering the stability of oscillators over time, the phase noise primarily affects the short-term stability[4] of oscillators (typically considered within less than a few seconds or minutes).

The frequency spectrum as given in (12.72) and seen in Figure 12.22 suggests that the single noise-modulated signal produced by an oscillator can be considered equivalent to multiple signals

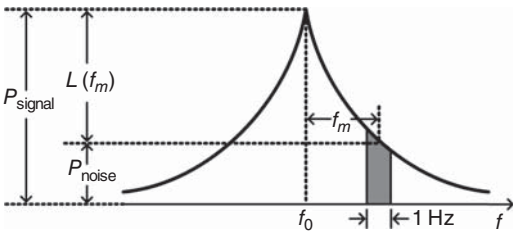


Figure 12.22. General oscillator signal spectrum showing the effect of phase noise on frequency as sidebands on the spectrum. These double-sideband noises around f_o are a part of the signal.

[4]Long-term instability refers to (slow) changes in frequency over an extended time (hours, days, months, years). It describes the stability of a source as it ages.

occurring concurrently: a (pure) principal signal having frequency f_o, also known as the carrier, and double side-band noise signals around the carrier having frequencies $f_o \pm f_m$, where f_m is the offset frequency from the carrier frequency. Therefore, when an oscillator is used in systems, whether as a local oscillator (LO) source for a mixer, a source for a transmitter, or anything else, the oscillator's signal contains not only the intended real RF signal but also other unwanted RF noise signals. It is this problem that causes degradation in the performance of corresponding components and systems employing oscillators, which, if sufficiently severe (i.e., high phase noise), can potentially disrupt the component or system function, and/or cause measurement inaccuracy. For instance, an oscillator with high phase noise produces large unwanted side-band noise signals (approaching the carrier in the limit), possibly causing radiation of unwanted signals for transmitters, unwanted mixing products in receivers, degradation of the signal-to-noise ratio (*SNR*) of receivers, etc. These side-band signals, therefore, should be as small as possible, which is possible only if the oscillator has a very low phase noise. Ideally, we want a spectrum as sharp as possible, approaching a single-frequency spectrum of an ideal (noise-less) oscillator.

We consider a general signal spectrum of oscillators as shown in Figure 12.22. The (single-sideband) phase noise at frequency f_m offset from the carrier's frequency f_o is defined as the ratio of a single-sideband noise power (P_{noise}) in a 1-Hz bandwidth at f_m to the carrier's total power (P_{signal}) as

$$\mathscr{L}(f_m) = 10\log\frac{P_{\text{noise}}}{P_{\text{signal}}} = 10\log\left(\frac{\text{Noise Power in 1 Hz at } f_o + f_m}{\text{Carrier Power}}\right) (\text{dBc/Hz}) \tag{12.73}$$

The phase noise as defined in (12.73) represents the normalized single-sideband noise power in a 1-Hz bandwidth with respect to the carrier's power. It is actually the single sideband spectral noise density. As an example, a phase noise of -85 dBc/Hz at an offset frequency of 100 KHz for a 35-GHz oscillator indicates that the single-sideband power of the signal (in a 1-Hz bandwidth) at 35 GHz $\pm$ 100 KHz is 85 dB below the signal power at 35 GHz.

The analysis of phase noise depends on the phase fluctuation $\phi_n(t)$ as can be seen in (12.71). This analysis, in general, is difficult to perform due to the fact that the phase fluctuation is random in nature. For simple illustration of the analysis without loss of generality, we assume that the phase fluctuation in (12.71) behaves as a simple sinusoidal signal of

$$\phi_n(t) = \Phi_n \sin(2\pi f_m t) \tag{12.74}$$

where Φ_n is maximum amplitude and f_m is offset-frequency from the carrier frequency f_o with $f_m t << 1$, which corresponds to offset-frequencies very close to the carrier frequency as typically considered in oscillators. Neglecting the effect of the AM noise and letting $\phi_o = 0$, we can rewrite the signal in (12.71) as

$$V(t) = A\sin[2\pi f_o t + \Phi_n \sin(2\pi f_m t)] \tag{12.75}$$

Expanding (12.75) gives

$$V(t) = A\sin(2\pi f_o t)\cos[\Phi_n \sin(2\pi f_m t)] + A\cos(2\pi f_o t)\sin[\Phi_n \sin(2\pi f_m t)] \tag{12.76}$$

which becomes, considering $f_m t << 1$,

$$V(t) \simeq A\sin(2\pi f_o t) + A\frac{\Phi_n}{2}\{\sin[(2\pi f_o + 2\pi f_m)t] + \sin[(2\pi f_o - 2\pi f_m)t]\} \tag{12.77}$$

Equation (12.77) shows that the output signal of the oscillator consists of a (pure) sinusoidal signal at f_o (principal signal) and two smaller-amplitude frequency-modulated (FM) signals at frequencies $f_o \pm f_m$ (noise signals), with modulation index Φ_n, representing the phase noise. This corresponds to a spectrum comprising a large component at f_o and two small side lobes at $f_o \pm f_m$. We can then see, as expected, that the power spectral density (PSD) of the oscillator's overall output signal is directly related to the PSD of the phase

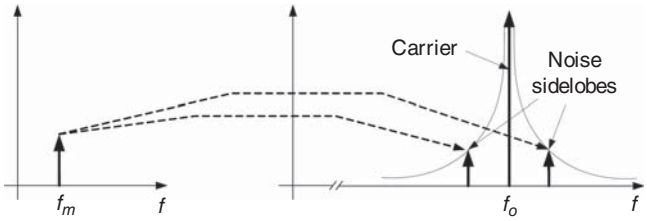


Figure 12.23. Illustration of the effect of the phase noise to the oscillator's output spectrum.

noise or phase fluctuation represented by the two FM signals besides the PSD of the carrier. The PSD of the phase noise $\phi_n(t) = \Phi_n \sin(2\pi f_m t)$ can be obtained from the real part of its Fourier transform as

$$S_\phi(f) = \frac{\Phi_n^2}{2}\delta(f - f_m) \tag{12.78}$$

where $\delta(f - f_m)$ is the Delta function. The PSD of the oscillator's output signal is obtained by taking a Fourier transform of (12.77) and making use of (12.78) as

$$S_V(f) = \frac{A^2}{2}\left[\delta(f - f_o) + \frac{1}{2}S_\phi(f - f_0) + \frac{1}{2}S_\phi(f_0 - f)\right] \tag{12.79}$$

which is directly related to the phase noise's PSD in (12.78) as noted earlier. It is noted that the phase noise's PSD given in (12.78) and used in (12.79) is for the noise at a single offset frequency f_m. As the frequency spectrum of a real oscillator's signal consists of continuous frequencies offset from the carrier frequency, the phase noise spectrum can be seen as an integration or summation of the individual phase noises from the continuously sinusoidal noise signals. These noises are viewed as the noise side lobes at both sides of the carrier frequency which, together, form the "phase-noise skirt" as seen in Figure 12.22. Figure 12.23 illustrates how the noise in the phase (phase noise) affects the oscillator's output spectrum by translating into the noise side lobes at both sides of the carrier frequency. The phase noise at an offset frequency f_m (i.e., at $f = f_o + f_m$) as defined in (12.73) can be obtained from $S_V(f)$ in (12.79), upon using (12.78), as

$$\mathscr{L}(f_m) = \frac{S_V(f_o + f_m)}{A^2/2} = \frac{S_\phi(f_m)}{2} \tag{12.80}$$

which is half of the (total) PSD $S_\phi(f_m)$ of the (phase) noise in both sidebands around the carrier, as expected.

Since frequency is the derivative of phase as given in

$$f(t) = \frac{\omega(t)}{2\pi} = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{12.81}$$

where the frequency $f(t)$ is modeled as an instantaneous parameter varying as a function of time due to the phase-noise effect, phase is, in turn, an integral of frequency and, hence, is related to frequency through $(2\pi f)^2$. Similar for the phase $\phi(t)$ that measures changes of phase with respect to time, $f(t)$ describes a deviation of frequency from its ideal value f_o at any given time. It shows that frequency in the presence of phase noise is an instantaneous quantity whose value varies over time around the (average) signal frequency f_o. The relation between frequency and phase leads to the conclusion that the phase noise can be viewed as relating to the frequency fluctuation or noise or vice versa, as noted earlier. Therefore, the PSD of frequency noise $S_f(f_m)$ at an offset frequency f_m can be derived directly from the PSD of phase noise $S_\phi(f_m)$ as

$$S_f(f_m) = (2\pi f_m)^2 S_\phi(f_m) = 2(2\pi f_m)^2 \mathscr{L}(f_m) \tag{12.82}$$

after applying (12.80).

Although a simple phase noise function as described in (12.74) is used in the foregoing analysis, the derived equation (12.80) for $\mathscr{L}(f_m)$ and the relation (12.82) for $S_\phi(f_m)$ and $\mathscr{L}(f_m)$ are valid for any time-dependence phase variation. In order to facilitate comparisons of phase noise at different carrier frequencies, we can also use the normalized frequency $\overline{f}(t) \equiv f(t)/f_o$ and define the corresponding normalized PSD $\overline{S}_\phi(f_m)$ as

$$\overline{S}_\phi(f_m) \equiv \left(\frac{f_m}{f_o}\right)^2 S_\phi(f_m) \tag{12.83}$$

In practice, the phase noise is typically characterized by $\mathscr{L}(f_m)$ or $S_\phi(f_m)$.

The phase noise in the frequency spectrum, as seen in Figure 12.22, also affects the oscillator signal in the time domain. We consider periodic signals for the oscillator such as a simple sinusoidal or square-wave signal as shown in Figure 12.24. Due to the noise-induced deviation of phase and hence frequency from the original ones at any given time, the period of the signal, which is the reciprocal of the frequency, can change at any time. As a result, the waveform does not cross the time axis precisely at the same points periodically – for instance, at every one half of the period of an ideal sinusoidal signal. The duration of the period for the signal hence can change from period to period, as illustrated in Figure 12.24. So the waveform, while still maintaining its shape, can move back and forth at any time, causing deviation from the ideal one. This problem occurring in the time domain is known as "time jitter." The amplitude of the phase noise affects directly the signal waveform and can severely disturb it if the phase noise is sufficiently large, making the signal departed significantly from its intended shape and hence is useless for either the oscillator alone or systems implementing it. Assume the maximum amplitude Φ_n of the phase fluctuation is small, which is typical for well-designed oscillators, the resultant waveform can be considered almost periodic as described in Figure 12.24 with an average period of T_o, corresponding to f_o, and a timing error of $\Delta\tau$. Several measures for the timing error can be used – for instance, "cycle-to-cycle jitter" which characterizes the difference between two consecutive periods, or "absolute jitter," which measures the timing error between two periodic waveforms with the same frequency. Since the phase noise is considered stationary and Gaussian, the standard deviation or effective variance ($\sigma\Delta\tau$) of the timing error ($\Delta\tau$) is the root mean square (*rms*) value of the timing error. A first-order formula relating the jitter to the phase noise can be derived by assuming that the white noise dominates the phase noise, and hence the white phase noise can be derived from the more general definition for jitter [3] as

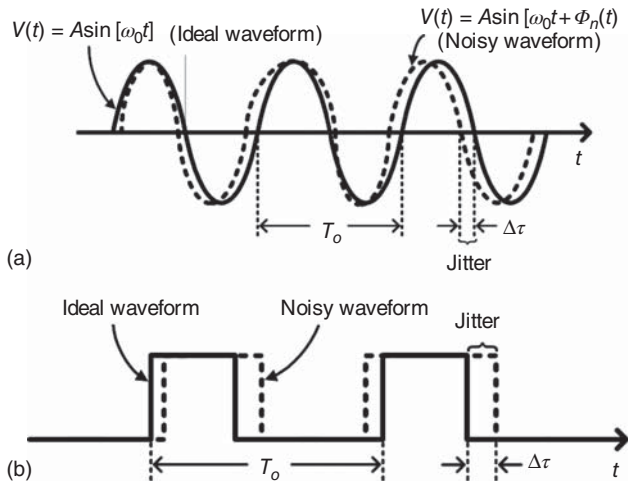


Figure 12.24. Sinusoidal (a) and square-wave (b) pure signals and their corresponding signals with phase noise showing the time-jitter effect of phase noise. T_o is the period of the carrier or ideal waveform. The time deviation $\Delta\tau$ is caused by the jittering resulting from the phase noise $\phi_n(t)$.

$$\mathcal{L}(f_m) = \frac{f_o}{f_m^2}\left(\frac{\sigma\Delta\tau}{T_o}\right)^2 \tag{12.84}$$

where T_o indicates the period of the oscillator's signal corresponding to f_o.

12.3.2 Phase Noise Modeling

For the analysis and design of oscillators, as well as of receivers, transmitters, and systems employing oscillators, it is fundamentally important to model the phase noise and identify not only the sources of noise in oscillators, including those in the constituent active elements such as transistors and amplifiers, that give rise to phase noise, but also the passive circuit elements in oscillators that affect the phase noise performance.

Again, we consider a feedback oscillator such as that shown in Figure 12.8. From that topology and the noise figure of amplifiers discussed in Chapter 11, we would expect that the thermal noise provided by $FkTB$, where $k = 1.374 \times 10^{-23}$J/°K is the Boltzmann's constant, T is the operating temperature in Kelvin degree (°K), and B is the absolute RF bandwidth of the amplifier, would enter the oscillator via the amplifier and contaminate the signal going through the amplifier, hence contributing to the oscillator's phase noise. This white noise power produces phase uncertainty across wide frequency ranges and is that of the main phase-noise contributions which depends on the noise figure of the transistor itself and the (amplifier) circuit employing that transistor. As the parameter $FkTB$ shows, this noise is constant across the oscillator's bandwidth and typically very small, hence producing a noise floor for oscillators at frequencies sufficiently far away from the carrier frequency until further suppressed by filtering of circuitry following the oscillator.

Another principal source of phase noise is "flicker noise." Transistors (or any active devices) always have flicker noise, also known as "$1/f$ noise," with spectrum varying as $1/f$ and occurring at relatively low frequencies, typically specified up to the $1/f$ corner frequency of transistors. The level of the $1/f$ noise and its corner frequency vary depending on the type of transistors and the characteristics of transistors. Bipolar junction transistors typically have lower $1/f$ corner frequency and noise than FETs. Within each type, some transistors may have higher or lower $1/f$ noise as compared to others. Since the flicker noise exists only at very low frequencies compared to RF, it does not cause any noise effect to RFICs operating as linear components or passive components such as switches. The flicker noise is a sort of correlated noise and can only provide effects in nonlinear time-variant systems. In an oscillator, the flicker noise is converted into phase noise through an up-conversion mixing process within the oscillator itself, in which the flicker noise at low frequencies is mixed with the oscillator's RF signal to produce a noise signal located next to the carrier by the same amount of the flicker-noise frequency. The $1/f$ noise region in the oscillator's spectrum, within which the $1/f$ noise dominates, typically spans from the carrier frequency f_o to the $1/f$ noise corner. The $1/f$ noise region is closest to the carrier and, within this region which covers small offset frequencies from the carrier frequency, the FM $1/f$ noise determines the oscillator's phase noise characteristic due to the fact that the $1/f$ noise is typically much higher than the white-noise floor produced by the amplifier's noise figure. Compared to bipolar transistors, CMOS transistors generate more flicker noise and, consequently, the $1/f$ phase noise is higher in CMOS oscillators as compared to other processes. As can be expected and mentioned earlier, the $1/f$ phase noise plays a dominant role near the carrier frequency, while the white noise affects the phase noise at higher frequencies far away from the carrier frequency. A less important phase noise effect is due to the shot noise, which is a function of DC bias currents to the transistors.

Since the oscillator signal including noise passes through the amplifier and feedback loop, which is assumed to contain a resonator functioning as a band-pass filter centered at the oscillator's frequency, these components affect the phase noise performance, and so design leading to low-phase noise needs to be considered for them as well. For instance, the noise figure of the amplifier needs to be low and the (feedback) resonator should be designed to have a quality factor as high as possible.

We analyze the oscillator's phase noise based on the phase noise theory developed by Leeson [4]. To facilitate the analysis, we consider a basic feedback oscillator as shown in Figure 12.25(a), in which we assume the input signal to the amplifier, and effectively to the oscillator, consists of two parts: an input noise signal represented by voltage $V_i = FkTB$, or PSD $S_{\phi i}(f)$, and an input RF signal, whose power is P_{avs}, representing

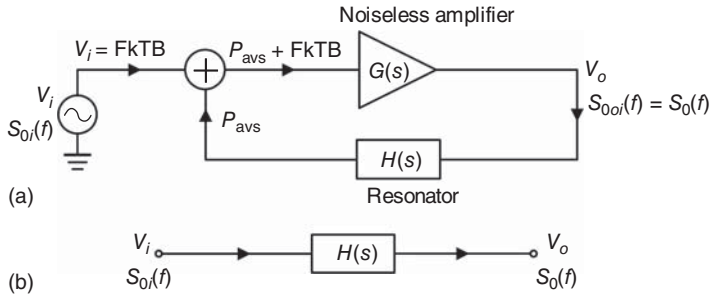


Figure 12.25. Feedback oscillator model (a) and its equivalence (b) used for phase-noise modeling. $s = j\omega$.

the power available at the input of the amplifier which is indeed the power going through the resonator and appearing at the amplifier's input. The output noise signal is shown as voltage V_o or PSD $S_{\phi o}(f) \equiv S_{\phi}(f)$. We also assume that the amplifier is noiseless, as far as the noise figure discussed in Section 11.2.1 of Chapter 11 is concerned, with gain $G(s) = 1$, where $s = j\omega$. This assumption is for the convenience of the analysis and is valid since the noise figure and gain are included externally in the noise-figure contribution via $S_{\phi i}(f)$.

The noise voltage transfer function of the oscillator can be obtained from (12.27) with $G(s) = 1$ as

$$\frac{V_o}{V_i} = \frac{1}{1 - H(s)} \tag{12.85}$$

As the gain of the amplifier is assumed to be 1, the overall noise transfer function is provided by the voltage transfer function $H(s)$ of the feedback network and, hence, we can obtain an equivalence for the feedback oscillator model as shown in Figure 12.25(b). Assume that the feedback behaves as an RLC network resonating at the oscillation frequency f_o, then its transfer function can be derived from the RLC resonators considered in Chapter 5 (Resonators) as

$$H(s) = \frac{1}{1 + j2Q_L\frac{f_m}{f_o}} \tag{12.86}$$

within a narrow bandwidth around f_o specified by $f_o \pm f_m$, where f_m is the frequency offset from f_o and Q_L is the loaded quality factor of the feed-back resonator given from (5.19) of Chapter 5:

$$Q_L = \frac{f_o}{\Delta f} \tag{12.87}$$

with Δf representing the 3-dB bandwidth of the feedback resonator.[5] It is noted, as discussed in Chapter 5, that the loaded quality factor Q_L depends on the load presented to the resonator, which basically characterizes the coupling between the resonator and the input and output of the amplifier. For very light coupling, Q_L approaches the resonator's unloaded quality factor Q_U. In practical circuits, it is reasonable to have Q_U around 2–5 times of Q_L. The noise power transfer function of the feedback oscillator can be obtained from (12.85) as

$$\left|\frac{V_o}{V_i}\right|^2 = \left|\frac{1}{1 - H(s)}\right|^2 \tag{12.88}$$

As discussed earlier, the PSD of a signal is equal to its Fourier of transform. Considering this and the fact that the transfer function of a network $H(s)$ is the Fourier transform of the network's impulse response $h(t)$, it can be deduced that (12.88) also represents the ratio between the PSD of the output noise or the PSD of

[5] In Chapter 5, we use $2\Delta f$ as the 3-dB bandwidth for resonators.

the phase noise $S_\phi(f)$ and the PSD of the input noise $S_{\phi i}(f)$. Hence, we can write this ratio, making use of (12.86) and (12.88), as

$$\frac{S_\phi(f)}{S_{\phi i}(f)} = \left|\frac{1 + j2Q_L f_m/f_o}{j2Q_L f_m/f_o}\right|^2 = \frac{1 + 4Q_L^2(f_m/f_o)^2}{4Q_L^2(f_m/f_o)^2} \tag{12.89}$$

or

$$S_\phi(f) = \left[1 + \frac{1}{4Q_L^2}\left(\frac{f_o}{f_m}\right)^2\right] S_{\phi i}(f) \tag{12.90}$$

The PSD of the input noise $S_{\phi i}(f)$ can be considered the phase noise of a standalone network consisting of the amplifier and resonator in series, which can be measured directly and used to calculate the phase noise from (12.90). For simplicity, if we neglect the noise contribution from the resonator, then we can measure the phase noise of the amplifier itself and approximate the oscillator phase noise as

$$S_\phi(f) \simeq \left[1 + \frac{1}{4Q_L^2}\left(\frac{f_o}{f_m}\right)^2\right] S_{\phi,\text{amp}}(f) \tag{12.91}$$

where $S_{\phi,\text{amp}}(f)$ is the PSD of the amplifier's phase noise.

The total output noise power P_{No} of an amplifier with unity gain is related to the amplifier's noise figure F as seen in Eq. (11.90) of Chapter 11 as

$$P_{\text{No}} = FkTB \tag{12.92}$$

Since we assume that the amplifier used in the feedback oscillator model is noiseless with gain of 1, this noise power also represents the noise that enters the amplifier and, together with the power P_{avs} available at the input of the amplifier, produces the SNR at the input of the amplifier. This SNR also represents the reciprocal of the PSD of the input noise, which can therefore be obtained for a 1-Hz bandwidth as

$$S_{\phi i}(f) = \frac{FkT}{P_{\text{avs}}} \tag{12.93}$$

As can be seen in (12.93) and, as expected from the SNR, the higher the signal power, the higher the SNR and hence the lower phase noise. Thus P_{avs} and hence the power that can be produced by the amplifier is an important factor for phase noise. The input noise PSD given in (12.93), as can be understood from the noiseless amplifier assumption earlier, only considers the effect of the noise figure and, therefore, describes a general PSD available at the input of the oscillator, neglecting other internal noise effects of the transistors and other components used in the oscillator. One particularly well-known noise occurring in solid-state devices at very low frequencies is the flicker or $1/f$ noise described earlier. This $1/f$ noise varies inversely as a function of the small frequency offset f_m near the carrier frequency according to f_c/f_m up to the "$1/f$ corner frequency" f_c, causing the $1/f$ phase noise dependence, and hence needs to be included in the phase-noise evaluation. It is recalled that (12.93) describes the input noise PSD under the (ideal) noiseless assumption for the amplifier and hence for the employed active devices. This assumption deviates with the inclusion of the $1/f$ noise, and hence it is justifiable to incorporate into (12.93) the whole effect as $(1 + A_{1/f} f_c/f_m)$, where $A_{1/f}$ represents the magnitude of the $1/f$ noise effect. With this consideration, we can then rewrite (12.93) as

$$S_{\phi i}(f) = \frac{FkT}{P_{\text{avs}}}\left(1 + A_{1/f}\frac{f_c}{f_m}\right) \tag{12.94}$$

$A_{1/f}$ and f_c for transistors can be determined by measuring the $1/f$ noise spectrum of the amplifier employing the transistor used in the oscillator. f_c is obtained approximately at the intersection of the $1/f$ noise curve and the white-noise curve (noise floor), where the $1/f$ noise power and white-noise power are equal, as can be seen later in Figure 12.25(b) for ultra-high Q resonator. Note that the $1/f$ noise and white-noise power,

and hence $A_{1/f}$ and f_c, change with the bias condition of the transistor. However, considering the facts that the $1/f$ noise only affects the phase noise at frequency $f = f_o \pm f_m$ near the carrier frequency (i.e., with very small f_m), but not at frequencies sufficiently far from it, and the effect is more pronounced at offset frequencies far away from the corner frequency f_c (i.e., $f_c/f_m >> 1$), the values of the constant $A_{1/f}$ should not be too large or too small. In view of this, it is reasonable to let $A_{1/f} = 1$, which would relatively do not cause a significant difference in $S_{\phi i}(f)$ to some extent. As a result, we can modify $S_{\phi i}(f)$ from (12.94) as

$$S_{\phi i}(f) = \frac{FkT}{P_{\text{avs}}}\left(1 + \frac{f_c}{f_m}\right) \tag{12.95}$$

The PSD of the phase noise $S_\phi(f)$ can now be derived from (12.90) and (12.95) as

$$S_\phi(f) = \frac{FkT}{P_{\text{avs}}}\left(1 + \frac{f_c}{f_m}\right)\left[1 + \frac{1}{4Q_L^2}\left(\frac{f_o}{f_m}\right)^2\right] \tag{12.96}$$

from which, the (single-sideband) phase noise of the oscillator can be obtained, making use of (12.80), as

$$\mathscr{L}(f_m) = \frac{FkT}{2P_{\text{avs}}}\left(1 + \frac{f_c}{f_m}\right)\left[1 + \frac{1}{4Q^2}\left(\frac{f_o}{f_m}\right)^2\right] \tag{12.97}$$

or, after rearrangement,

$$\mathscr{L}(f_m) = \frac{FkT}{2P_{\text{avs}}}\left[f_c\left(\frac{f_o}{2Q_L}\right)^2\left(\frac{1}{f_m^3}\right) + \left(\frac{f_o}{2Q_L}\right)^2\left(\frac{1}{f_m^2}\right) + f_c\left(\frac{1}{f_m}\right) + 1\right] \tag{12.98}$$

The phase noise $\mathscr{L}(f_m)$ is typically given in logarithm [i.e., $10\log(.)$] with unit of decibels relative to the carrier/Hertz.

Equation (12.98) shows that the phase noise is contributed by four different noises, which describe specific behaviors of the phase noise in different regions of frequency around the carrier frequency. f_m is the (offset) frequency variable at which the phase noise is evaluated for oscillators operating at f_o, while the $1/f$ corner frequency f_c and the half 3-dB bandwidth $\Delta f/2 = f_o/2Q_L$ dictated by the transistor and resonator used in the oscillator circuit, respectively, specify the frequency transitions at which the frequency-dependence behavior of the phase noise changes. Therefore, the phase noise as a function of f_m should be examined considering f_c and $f_o/2Q_L$. It is recognized from (12.98) that high- and low-quality factor correspond to small and large bandwidth Δf relative to f_o, respectively. Since the $1/f$ corner frequency f_c is relatively small for typically used transistors, at most around a few tens of MHz, for low-frequency feedback oscillators, we may specify that oscillators using low-Q and high-Q resonators correspond to[6] $\Delta f/2 > f_c$ and $\Delta f/2 < f_c$, respectively. The assumption of $\Delta f/2 < f_c$ for high Q, however, may not be valid for RF oscillators, especially those operating in high RF regions. As an example, we consider a 20-GHz feedback oscillator employing a MOSFET having $f_c = 20$ MHz, which is considered high for MOSFETs. In order for $\Delta f/2 < f_c$, Q_L of the resonator must be greater than 500 which is impossible to obtain in CMOS RFIC with current CMOS technology. For RF oscillators, even $\Delta f/2$ is relatively small with respect to f_o for high-Q resonators, it may still be much larger than f_c of transistors - for instance, a loaded quality factor of[7] 50 for a 50-GHz oscillator corresponds to $\Delta f/2 = 0.5$ GHz which should be much larger than f_c of any good transistor. Since extremely high-Q resonators are very difficult, if not possible, to obtain in CMOS RFIC, particularly at very high frequencies, most RF oscillators, whether employing low- or high-Q resonators, should be classified as corresponding to $\Delta f/2 > f_c$. Nevertheless, for the sake of understanding the insight of the frequency dependence of the phase

[6] $\Delta f/2$ is used instead of Δf since only a half of the 3-dB bandwidth, corresponding to a half of the spectrum around f_o (i.e., $f > f_o$), is considered.
[7] $Q_L = 50$ is considered very high for CMOS lumped-element resonators at 50 GHz.

noise behavior, we will examine the phase noise for two cases of resonators: low/high Q and ultra-high Q, but keeping in mind that majority of RF oscillators should conform to the low/high Q case.

Low/High Q: This condition corresponds to $\Delta f/2 > f_c$ or $f_c < f_o/2Q_L$ and can be achieved when the loaded quality factor of the feedback resonator is either low or high. This condition is met by most RF oscillators. Under this condition, when $f_m \leq f_c$ (i.e., the offset frequency f_m is very small or the frequency $(f_o + f_m)$ is very close to the oscillator frequency f_o, the phase noise, as can be seen in (12.98), is dominated by the first term and its value primarily varies as $1/f^3$, where $f \equiv f_m$, up to $f_m = f_c$, corresponding to −9 dB/octave. For $f_c \leq f_m \leq f_o/2Q_L$, the phase noise is dominated by the second term and hence behaves according to $1/f^2$ corresponding to −6 dB/octave. The $1/f^2$ noise region occurs at higher offset frequencies than the $1/f^3$ noise region. Both of these noises are FM around the carrier of the oscillator circuit. It is noted that, for all offset frequencies $f_m \leq f_o/2Q_L$ or f_m within the resonator's half 3-dB bandwidth (pass-band), the feedback-loop gain is maximum, causing the noise to be amplified to much larger than the noise floor produced by $FkT/2\ \ P_{\text{avs}}$, hence dominating the overall phase noise. Beyond $f_o/2Q_L$, where the offset frequency is sufficiently far away from the oscillator frequency f_o (stop-band), the gain of the feedback loop becomes much smaller than 1, forcing the noise signal to be largely suppressed; the fourth term becomes dominant and accordingly the phase noise shows approximately a $1/f^o$ dependence or constant behavior. In this relatively higher frequency range, the phase noise is the thermal noise, which indeed sets the noise floor for the oscillator.[8] If there are other components following the oscillator, the constant thermal noise level would continue up to frequency $f_1 = \Delta f_1/2$ set by the 3-dB filtering bandwidth Δf_1 of the subsequent components, after which, the noise level drops according to the filtering function. Furthermore, any subsequent amplifier such as driver or power amplifier also provides additional noise and change the overall phase noise. 12.26(a) shows a sketch of the phase noise with low/high Q resonators, showing the noise spectrum in three different frequency regions depending on the offset frequency from the oscillator's frequency f_o: $1/f^3$ noise region, $1/f^2$ noise region and $1/f^0$ noise region. It is especially reminded that the loaded quality of the resonator, and hence its corresponding unloaded quality, dictates the phase noise at frequencies close to the carrier frequency, making it crucial to increase the loaded quality factor in order to lower the phase noise close to the carrier frequency.

Ultra-High Q: This condition corresponds to $\Delta f < f_c$ or $f_c > f_o/2Q_L$ and can be achieved when the loaded quality factor of the feedback resonator is extremely high. This condition is rarely met by most RF oscillators due to the fact that it is extremely difficult, if not impossible, to design ultra-high Q resonators in CMOS RFIC meeting the condition $Q_L > f_o/2f_c$. For very small offset frequency $f_m \leq f_o/2Q_L$ near the carrier frequency, the phase noise, as for the low/high Q case, is also controlled by the first term and, as such, shows its frequency dependence approximately as $1/f^3$ up to $f_m = f_o/2Q_L$, corresponding to −9 dB/octave. For $f_o/2Q_L \leq f_m \leq f_c$, the phase noise is dominated by the third term and behaves according to $1/f$ corresponding to −3 dB/octave. Beyond f_c, the fourth term dominates and accordingly the phase noise approximately stays constant with frequency, whose level establishes the noise floor for the oscillator. If there are other components following the oscillator, the constant thermal noise level would continue up to frequency $f_1 = \Delta f_1/2$ set by the 3-dB filtering bandwidth Δf_1 of the subsequent components, after which, the noise level drops according to the filtering characteristics. Figure 12.26(b) shows the behavior of the phase noise versus frequency with ultra-high Q resonators in which, the noise spectrum is classified into three regions: $1/f^3$ noise region, $1/f$ noise region and $1/f^0$ noise region.

For $f_m \leq f_c$ and $f_c \leq f_m \leq f_o/2Q_L$ for the low/high Q case and $f_m \leq f_o/2Q_L$ for the ultra-high Q case, the phase noise at an offset frequency in each frequency region is always proportional to $(f_o/2Q_L)^2$. Accordingly, we can derive the relation between the phase noise $\mathscr{L}_1(f_m)$ and $\mathscr{L}_2(f_m)$ corresponding to different loaded quality factors Q_{L1} and Q_{L2}, respectively, as

$$\mathscr{L}_1(f_m) = \mathscr{L}_2(f_m)\left(\frac{Q_{L2}}{Q_{L1}}\right)^2 \tag{12.99}$$

[8]For instance, considering an oscillator employing an amplifier having $F = 5$ dB and $P_{\text{avs}} = 1$ dBm operating at room temperature, the phase-noise floor is -174 dBm + 5 dB − 1 dBm − 3 dB = − 173 dBm.

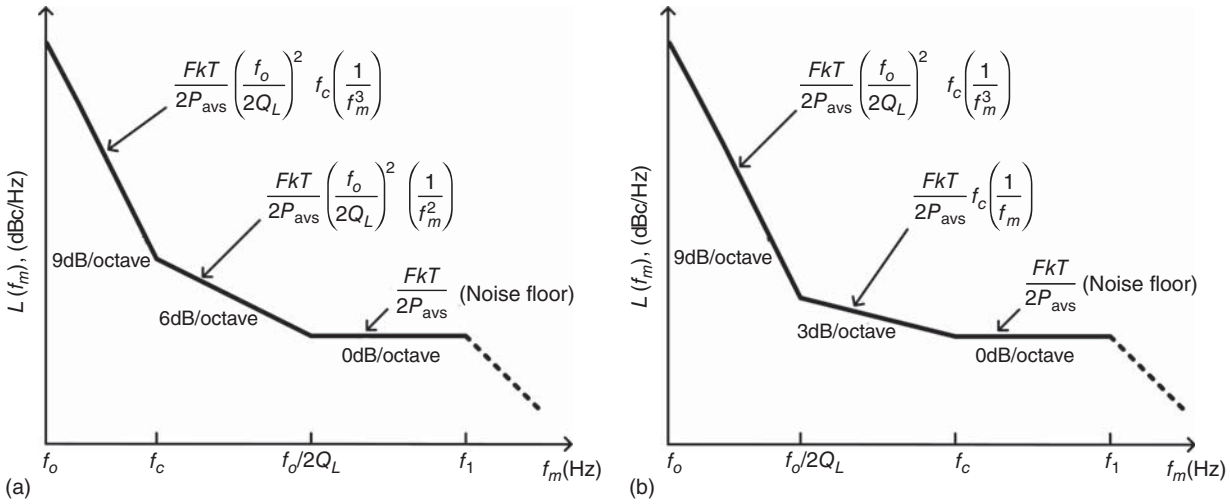


Figure 12.26. Phase noise characteristics of oscillators with low/high Q (a) and ultra-high Q (b) resonators. The phase-noise level after frequency f_1 shows the effect of filtering caused by components, if any, following oscillators. For stand-alone oscillators, the thermal noise floor is constant pass f_1.

In low phase-noise oscillator design, it is useful to determine the required quality factor for resonators. From (12.98), we can derive the required loaded quality factor Q_L for a given phase noise $\mathscr{L}(f_m)$ as

For low/high Q:

$$Q_L \geq \begin{cases} \dfrac{1}{2}\dfrac{f_o}{f_m}\sqrt{\dfrac{f_c}{f_m}}\sqrt{\dfrac{FkT}{2P_{\text{avs}}}}\dfrac{1}{\sqrt{\mathscr{L}(f_m)}}, & f_m \leq f_c \\ \dfrac{1}{2}\dfrac{f_o}{f_m}\sqrt{\dfrac{FkT}{2P_{\text{avs}}}}\dfrac{1}{\sqrt{\mathscr{L}(f_m)}}, & f_m \geq f_c \end{cases} \tag{12.100}$$

For ultra-high Q:

$$Q_L \geq \frac{1}{2}\frac{f_o}{f_m}\sqrt{\frac{f_c}{f_m}}\sqrt{\frac{FkT}{2P_{\text{avs}}}}\frac{1}{\sqrt{\mathscr{L}(f_m)}} \tag{12.101}$$

For $f_m \geq f_o/2Q_L$ or $f_m \geq f_c$ in the low/high Q or ultra-high Q, respectively, the phase noise is set by the phase-noise floor, and so $\mathscr{L}(f_m) \geq \frac{FkT}{P_{\text{avs}}}$, from which we can obtain the minimum level for the available power P_{avs}, and hence the corresponding oscillator's output power and output coupling mechanism to extract this power, needed for a given phase noise as

$$P_{\text{avs}} \geq \frac{FkT}{\mathscr{L}(f_m)} \tag{12.102}$$

and the minimum noise figure required for the amplifier as

$$F \geq \frac{P_{\text{avs}}}{kT}\mathscr{L}(f_m) \tag{12.103}$$

As an example showing the phase-noise dependence on one of the oscillator's circuit parameters, we show in Figure 12.27 the calculated phase noise of a CMOS feedback oscillator as a function of the offset frequency for different values of the resonator's loaded quality factor Q_L.

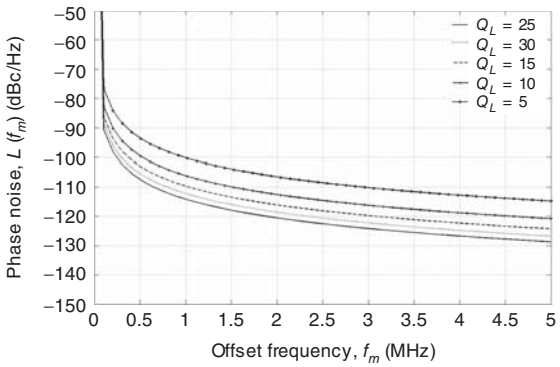


Figure 12.27. Calculated phase noise as a function of the offset frequency for different Q_L. $f_o = 35$ GHz, $F = 5$ dB, $P_{\text{avs}} = 0$ dBm, and $f_c = 200$ KHz.

12.3.3 Low Phase-Noise Design Consideration

In order to find ways to reduce the phase noise in oscillators, we examine the phase noise expression (12.98) whose spectrum is sketched in Figure 12.26. It is apparent that the phase noise can be reduced by employing proper transistors and circuit design. These include using transistors with small $1/f$ corner frequency, small $1/f$ noise and low minimum noise figure, and proper circuit elements to increase the power entering the amplifier through the feedback loop, lowering the amplifier's noise figure, and increasing the resonator's unloaded and loaded Q. Increasing the power entering the amplifier, however, also means reducing the power taken out as the oscillator output power – and so care needs to be exercised not to reduce this power to an undesired level, particularly not to be below the white-noise floor of the oscillator. It is recalled that bipolar junction transistors have much lower $1/f$ noise than FETs and so are more suitable for low phase noise. It is also recalled that the loaded Q approaches the unloaded Q as the loading effects to the resonator are reduced. Therefore, not only the resonator's unloaded Q needs to be maximized, its loaded Q in the oscillator also needs to be increased. This can be achieved by minimizing the coupling between the feedback resonator and the amplifier at both the amplifier's input and output ports. This, nevertheless, inadvertently reduces the amplifier's output power going into the feedback path and hence to the amplifier's input, thereby affecting the phase noise.

Particularly, since the flicker-induced phase noise is generated by an up-converted mixing process, which happens only in nonlinear devices, the efficiency of this conversion can be reduced by proper oscillator design. For instance, in single-ended oscillators, where only the fundamental signal is desired and all harmonics are filtered out, a noise-reduction circuit may be employed to reduce the $1/f$ -noise upconversion to the fundamental frequency. In balanced oscillators, such as push-push or differential oscillators, not only the noise-upconversion to the fundamental frequency needs to be considered, but also the conversion to harmonics needs to be taken account – for instance, proper circuit configurations need to be used to trap or suppress the undesired phase noise at both fundamental and harmonics, and the circuit should be as balanced as possible to minimize the generation of undesired harmonics, hence preventing or lessening the noise-up-conversion process. Moreover, to minimize the flicker-induced phase noise, the $1/f$ corner frequency of the transistor used in the oscillator should be considered in circuit design. For instance, the oscillator used in a phase-locked loop (PLL) synthesizer should have its $1/f$ corner frequency lower than the lowest frequency of the PLL's bandwidth, and hence the noise can be low-pass filtered and does not appear at the output of the synthesizer. If the oscillator is not carefully designed or the corner frequency is not properly considered (e.g., using high corner frequency), the flicker noise can deteriorate the phase noise at high offset frequencies (below the corner frequency.)

12.3.4 Effects of Phase Noise on Systems

As mentioned earlier, the phase noise degrades the ability and/or quality of systems such as limiting the dynamic range of a system (i.e., the minimum to maximum signal levels that a system can take). The upper

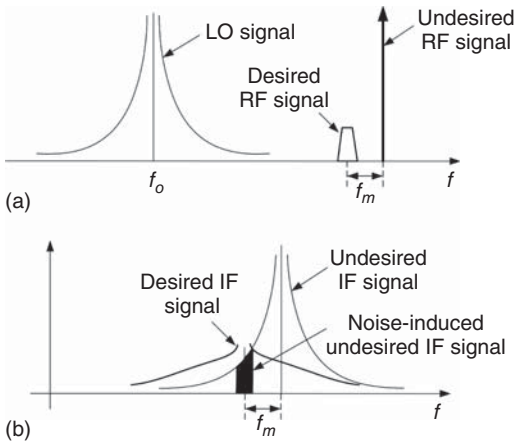


Figure 12.28. Effects of phase noise on signal-reception quality in communication and sensing systems: desired, undesired RF signal spectra and LO's spectrum (a) and corresponding down-converted IF signal spectra (b).

end of the dynamic range of a system is determined by the maximum allowable input signal to the system, which is determined by the system's linearity, while the lower end is characterized by noise, which may be the wideband thermal noise (signal's noise floor) set by the noise figure of the amplifier used in a LO or the narrow-band phase noise of a LO around its carrier frequency. The basic problem that causes performance degradation in systems lies in possibilities of phase-noise-induced noisy signals masking desired signals or causing uncertainty in measurements.

Typically, in systems such as communication or sensing, a relatively small wanted RF signal must be detected in the presence of relatively large unwanted signals such as noise from surrounding environments (Earth, buildings, trees, etc.) and other interfering RF signals. The detection occurs in the receive path, where the desired RF signal is normally down-converted to a lower intermediate frequency (IF) signal by mixing with an LO signal in a mixer. To illustration the noise effects of the LO on the RF signal reception, we consider desired and undesired RF signals whose frequencies are separated by f_m, which is same as the offset frequency f_m from the LO's frequency f_o, as shown in Figure 12.28(a). Both the desired and undesired RF signals are down-converted to corresponding IF signals in a mixer. If the LO is ideal (i.e., having no phase noise), the down-converted desired and undesired IF signals would be separated by the same frequency amount of f_m and are completely distinguished from each other. The removal of the undesired IF signal from the desired IF signal, although could be very difficult or even impossible depending on the value of f_m, can be theoretically achieved using a filter of (unrealistic) infinite quality factor (Q). A practical LO, however, is noisy, and so the desired and undesired RF signals also mix with the phase noises (which are also signals) at all frequencies in the LO's spectrum around f_o, including the main LO frequency f_o, to produce the desired and undesired noisy IF signals whose spectra are shown in Figure 12.28(b). We now specifically consider the down-conversion at the offset frequency f_m. The phase noise at this offset frequency from f_o mixes with the undesired RF signal to produce an undesired IF signal falling exactly at the same frequency of the desired IF signal, as illustrated in Figure 12.28(b). The main desired IF signal and the noise-induced undesired IF signal (noise) are all located at f_m away from the main undesired IF signal. The undesired and desired down-converted signals are overlapped and thus cannot be separated even when an (ideal) infinite-Q filter is used. The desired IF signal is thus contaminated with noise. Ultimately, the phase noise of the oscillator can partially or completely mask desired signals depending on the phase noise and signal levels, thereby effectively affecting the minimum detectable signal level of systems. If the phase noises at offset frequencies are not sufficiently low, the resultant undesired noises appearing in the IF band would seriously degrade the *SNR* of the receiver and hinder its performance.

In the transmit path of systems, the transmitting power is typically provided directly from an oscillator via power amplifiers or through a mixer (up-converter), which up-converts a low-frequency signal to an RF signal by mixing it with an LO's signal and power amplifiers. The quality of the direct-generation or

up-converted transmitting RF signal also depends upon the employed oscillator and contains noise contributed by the oscillator's phase noise. In many applications, particularly communications or non-military sensing, the high-power transmitting RF signal must satisfy a certain power mask, which limits the transmitting power. Therefore, the output spectrum of the oscillator, and hence the direct or up-converted transmitting RF signal, must also comply with the same power mask. Usually, this requirement is much less stringent than the one imposed by the reception quality for the receive path of systems. The power mask indirectly dictates the limitation of the amount of noise that can be transmitted. Under such conditions, the transmitting noise may be so small that cannot mask other real RF transmitting signals.

With the increase of spectrum traffic, complexity of systems and wireless applications, the phase noise of oscillators, whether used in receivers or transmitters, becomes even more critical to system operation. For instance, multiband or ultra-wideband systems with multiple adjacent bands or channels demand more frequency selectivity and hence better phase noise. The instability of frequency sources in such systems may cause "spill" effects on signals, degrading a system's ability in operating precisely at separate adjacent bands or channels.

12.3.5 Analysis Example of Effects of Phase Noise

As an illustration of analyses of the effects of the phase noise on system performance, we analyze in this section the effect of the phase noise of RF signal sources on the phase measurement using RF interferometry. In RF interferometers, the transmit signal from the RF source is mixed with the receive signal (obtained either by reflection from or transmission through the object) in a quadrature mixer to produce a base-band signal. This base-band signal corresponds to the phase difference (due to the time delay) between the transmit and receive signals and can be processed to produce the object information. Interferometry is basically a phase detection process and the measured phase error primarily dictates the accuracy of the target signature. The total induced phase error attributes to the combination of the quadrature mixer's imbalance and RF source's instability caused by its phase noise during the phase detection process. The contribution from the quadrature mixer imbalance can be compensated by a correction algorithm in the signal processing, leaving the phase noise of the RF source as the principal source of phase error. This phase-noise-induced phase error is analyzed as follows.

We begin by describing the transmit and receive signals in an RF interferometric system as

$$V_T(t) = [1 + a(t)]A_T \cos[2\pi f_o t + \phi_o + \phi_n(t)] \tag{12.104}$$

$$V_R(t) = [1 + a(t - t_d)]A_R \cos[2\pi f_o(t - t_d) + \phi_o + \phi_n(t - t_d)] \tag{12.105}$$

respectively, where A_T and A_R are the amplitudes of the transmit and receive signals, respectively; $a(t)$ and $\phi_n(t)$ designate the AM and PM noise effects, respectively; ϕ_o is an arbitrary phase constant; and t_d is the time delay between the transmit and receive signals. The AM noise contribution in the estimation or measurement of the phase error can be neglected as discussed in Section 12.3.1.

The in-phase and quadrature output signals of the quadrature mixer, after low-pass filtering, are given by

$$V_I(t) = A_I \cos[\Delta\phi_n(t) + \phi_S] + n(t) \tag{12.106}$$

and

$$V_Q(t) = A_Q \cos[\Delta\phi_n(t) + \phi_S] + n(t) \tag{12.107}$$

respectively, where $\Delta\varphi_n(t) = \varphi_n(t) - \varphi_n(t - t_d)$ represents the phase error due to the phase noise of the RF source; $\phi s = 2\pi f_o t_d$ is the (constant) phase difference between the transmit and receive signals corresponding to the time delay t_d; and $n(t)$ stands for any additional noise from the receiver, which can be reduced significantly by averaging the digitized data in the signal processing, making its impact on the system performance negligibly small. The most important contribution to the phase error, as can be seen, is $\Delta\phi_n(t)$, which

is dictated by the RF source's phase noise and the time delay between the transmit and receive signals. The accuracy of the phase measurement is degraded by this noise term, and hence the phase noise of the RF source, which has uncertainty due to the random nature of noise.

The PSD of the phase error $\Delta\phi_n(t)$ can be derived as

$$S_{\Delta\phi}(f) = 2S_{\phi}(f)[1 - \cos(2\pi f t_d)] \; = t_d{}^2(2\pi f)^2 S_{\phi}(f)\left[\frac{\sin(2\pi f t_d/2)}{2\pi f t_d/2}\right]^2 \tag{12.108}$$

where $S_{\phi}(f)$ is the PSD of the phase noise $\phi_n(t)$. The corresponding PSD of frequency noise is obtained from (12.82) as

$$S_{\Delta f}(f) = (2\pi f)^2 S_{\Delta\phi}(f) \tag{12.109}$$

We can rewrite (12.108), upon using (12.109), as

$$S_{\Delta\phi}(f) = t_d{}^2 S_{\Delta f}(f)\left[\frac{\sin(2\pi f t_d/2)}{2\pi f t_d/2}\right]^2 \tag{12.110}$$

The variance of the phase error $\Delta\phi_n(t)$ can be derived as

$$\sigma^2[\Delta\phi_n(t)] = \frac{t_d{}^2}{\pi}\int_0^{\infty} S_{\Delta f}(f)\left[\frac{\sin(2\pi f t_d/2)}{2\pi f t_d/2}\right]^2 df \tag{12.111}$$

As can be seen from (12.111), the mean square deviation of the phase noise, $\sigma^2[\Delta\varphi_n(t)]$, relates to the frequency noise PSD of the signal source, $S_{\Delta f}(f)$, and hence its phase noise PSD, $S_{\Delta\phi}(f)$, or phase noise, $\mathscr{L}(f)$, which can be measured using a spectrum analyzer. The standard deviation or rms value of $\Delta\phi_n(t)$ is considered the phase error, that is, the phase deviation from the expected phase corresponding to the delay time t_d due to the phase noise of the signal source. Therefore, the phase noise of the signal source directly affects the phase measurement accuracy and can be used to determine the phase error based on (12.111).

12.4 OSCILLATOR CIRCUITS

The most critical task in oscillator design, just like for any other circuits, is selecting or configuring a proper topology to potentially meet certain requirements such as output power and/or phase noise. There are various oscillator circuits and, in the previous section, we have studied the basic feedback oscillators, which can serve either as a stand-alone circuit or as a component in another oscillator topology. We recall that many oscillators can be modeled as feedback oscillators. In this section, we will examine three additionally important topologies: cross-coupled oscillators, distributed oscillators, and push-push oscillators.

12.4.1 Cross-Coupled Oscillators

Cross-coupled oscillators have relatively good phase noise, ease of implementation and differential balanced operation. They are widely used in various RF systems. The differential topology is particularly attractive as compared to a single-ended counterpart. It has better phase noise and less harmonic distortion, resulting in a more stable frequency source with higher spectrum purity. The differential output facilitates direct integration with other balanced circuits such as doubly balanced mixers, but requires a balun or transformer for connection with other single-ended components.

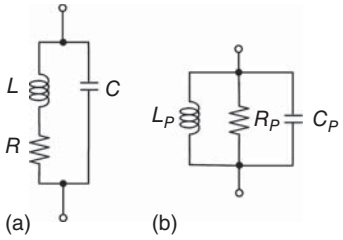


Figure 12.29. A resonator (a) and its equivalent parallel resonator (b).

12.4.1.1 Resonator or Tank Circuit. As for the basic oscillator described in Section 12.2.1, cross-coupled oscillators employ RLC resonators or "tanks" as the oscillation-frequency determined element in conjunction with active devices.

We consider a parallel resonator consisting of inductor and capacitor for cross-coupled oscillators. Typical MIM capacitors have very high Q and so we can assume that the capacitor in the resonator has negligible loss while the inductor is lossy as shown in Figure 5.2(a) of Chapter 5, which is shown here again in Figure 12.29(a). This resonator behaves as a parallel resonator shown in Figure 12.29(b), which consists of R_p, L_p, and C_p in parallel, within a narrow frequency range around an interested frequency ω as noted in Section 5.1.1 and in Problem 5.2. The equivalent elements L_p, R_p, and C_p can be derived as follows. L_p is obtained as

$$L_p = L\left(1 + \frac{R^2}{\omega^2 L^2}\right) = L\left(1 + \frac{1}{Q^2}\right) \tag{12.112}$$

where

$$Q = \omega L/R \tag{12.113}$$

Typical on-chip inductors have Q larger than 5, hence we can approximate L_p from (12.112) as

$$Lp \simeq L \tag{12.114}$$

R_p is given as

$$Rp \simeq \frac{\omega^2 L^2}{R} = RQ^2 \tag{12.115}$$

and

$$C_p = C \tag{12.116}$$

At the resonant frequency $\omega = 1/\sqrt{L_p C_p}$, the resonator reduces to a simple resistor R_p and the phase difference between the voltage and current of the resonator vanishes.

12.4.1.2 Oscillation. We now consider a possible oscillation resulting from connecting an RLC parallel resonator or tank to an active device. If a "negative resistor," represented by an active device or a circuit such as that shown in the oscillating network of Figure 12.5(a), whose value is $-R_p$ is placed in parallel with the resistor R_p of the tank as shown in Figure 12.30, the impedance of the tank becomes infinity. The quality factor of the tank then approaches infinity and the entire circuit of Figure 12.30 would oscillate indefinitely. The oscillation frequency is identical to the tank's resonant frequency given as

$$\omega_r = 1/\sqrt{LpCp} \tag{12.117}$$

As mentioned earlier, the negative resistance under a small-signal operation is typically selected as $-3R_p$ to $-2R_p$ to sustain a stable oscillation. Various ways including using proper impedance termination or positive feedback can be used to produce a negative resistance for MOSFETs. With proper terminating impedance

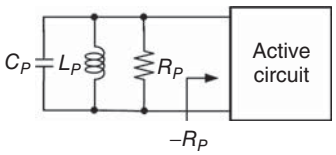


Figure 12.30. Oscillator formed by a tank and a negative-resistance active circuit.

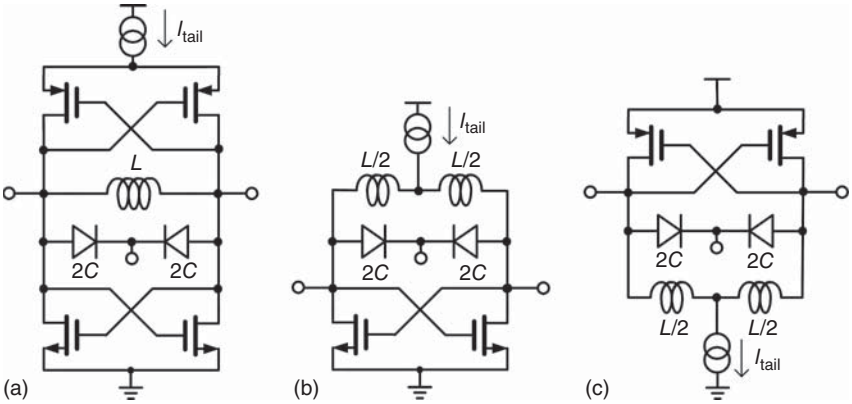


Figure 12.31. Cross-coupled VCO topologies using complementary NMOS and PMOS (a), NMOS-only (b), and PMOS-only (c) transistors. The diodes are varactors used for frequency tuning.

Z_{out} to a MOSFET, as illustrated in Figure 12.5(a), $|\Gamma_{in}| > 1$ and hence $\mathrm{Re}(Z_{in}) < 0$ can be achieved. Referring to Figure 12.8, the output impedance of the (closed-loop) feedback network can be derived as

$$Z_o = \frac{Z_{o,\mathrm{open}}}{1 + A(j\omega)} \tag{12.118}$$

where $Z_{o\ \mathrm{open}}$ is the output impedance without feedback (open-loop) and $A(j\omega) = G(j\omega)H(j\omega)$ is the total loop gain. At the resonant frequency of the tank, if the total loop gain is sufficiently negative (i.e., the feedback is sufficiently positive), a negative resistance can be obtained for the feedback network's output impedance, hence leading to a possible oscillation.

12.4.1.3 Analysis and Design. For generality, we consider a cross-coupled voltage-control oscillator (VCO) instead of a (single-frequency) cross-coupled oscillator. As discussed earlier, an oscillator must provide a negative resistance in order to potentially produce a negative loop resistance to sustain an oscillation. For single-ended design, one transistor connecting to a resonator is sufficient to produce a negative loop resistance. However, for a balanced design such as differential circuit, a minimum of two transistors is necessary. We consider common-source cross-coupled VCO topologies as shown in Figure 12.31, each employing two cross-coupled transistors to provide a positive feedback for achieving the required negative resistance. The cross-coupled transistors can be implemented with only NMOS, only PMOS, or complementary NMOS and PMOS transistors as seen in Figure 12.31. To provide current biasing, additional transistors to create a current source (I_{tail}) are also needed. These additional transistors should have no influence on the performance of a perfectly balanced oscillator.

Current- and Voltage-Limitation Operation

Two modes of operation, namely current- and voltage-limitation, can be identified for typical cross-coupled oscillators, considering the bias current as an independent variable [5]. In the current limitation, the voltage amplitude of the oscillation grows linearly with the bias current until the oscillator enters the voltage-limitation operation. In the voltage-limitation region, the oscillation amplitude is limited to a certain

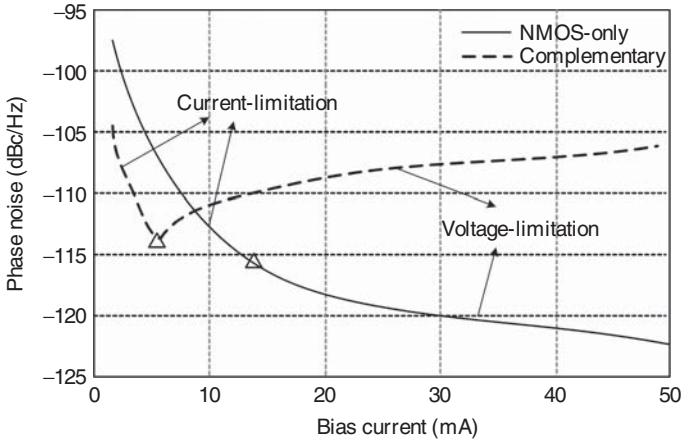


Figure 12.32. Calculated phase noise at 5-GHz oscillation and 1-MHz offset versus bias current for complementary and NMOS-only cross-coupled VCOs.

value $V_{\max}$, which is determined by the supply voltage and/or a change in the operating mode of the active devices (e.g., MOSFET entering the triode region). The voltage amplitude of the oscillation can thus be expressed as

$$V_A = \begin{cases} R_p I_{\text{tail}}, & \text{Current} - \text{Limitation} \\ V_{\max}, & \text{Voltage} - \text{Limitation} \end{cases} \tag{12.119}$$

In the current-limitation operation, the bias current determines the amplitude of the oscillation. Figure 12.32 shows the calculated phase noise of cross-coupled VCOs employing 0.25-μm NMOS-only and complementary NMOS/PMOS transistors with a typical LC tank having $Q = 10$. The calculations are done at 5-GHz oscillation and 1-MHz offset using the RF spectra simulator in Cadence [6]. The phase noise illustrates the influence of the bias current on the phase noise. The triangle sign (Δ) represents the boundary between the current-limitation and the voltage-limitation operation, which are obtained by calculations using Cadence. When the bias current is lower than the level at the triangle sign, the current-source devices always work within the saturation region. However, when the bias current is increased beyond the triangle sign, the current-source devices are pushed into the triode region due to larger oscillation swings. In the complementary topology, the amplitude is fixed by the power supply and the phase noise increases due to the noise up-conversion. For low power supply voltages, the lowest phase noise is attainable with an NMOS- or PMOS-only topology at the expense of high current. It is noted that for the complimentary topology, the bias current should be at the level such that the VCO operates at the boundary between the current-limitation and the voltage-limitation regions, which is the optimal design point for good phase noise with low power consumption represented by the triangle sign in Figure 12.32 on the dashed curve.

Comparison of Complementary and NMOS (PMOS)-only VCO

Each of the cross-coupled VCO topologies employing NMOS-only, PMOS-only, and complementary NMOS and PMOS, as shown in Figure 12.31, has its own advantages and disadvantages. To compare the performance of these different topologies, it is assumed that the bias conditions and the LC tanks used in these circuits are same. The transistors are assumed to operate as an ideal switch, fully switching on and off during each half of the period without any transient such that the drain current of the transistors is a square wave [6].

One of the requirements for cross-coupled oscillators is the circuit must be balanced. To that end, the transistors in the transistor pair of the NMOS- or PMOS-only topology must be identical and the transconductances of the NMOSFET and PMOSFET in the complimentary topology must be equal. The sum of both transconductances in the complementary topology is equal to G_M while, for the NMOS (PMOS)-only topology, $G_M = g_m/2$ where g_m is the transconductance of each transistor. In the complementary case, the bias

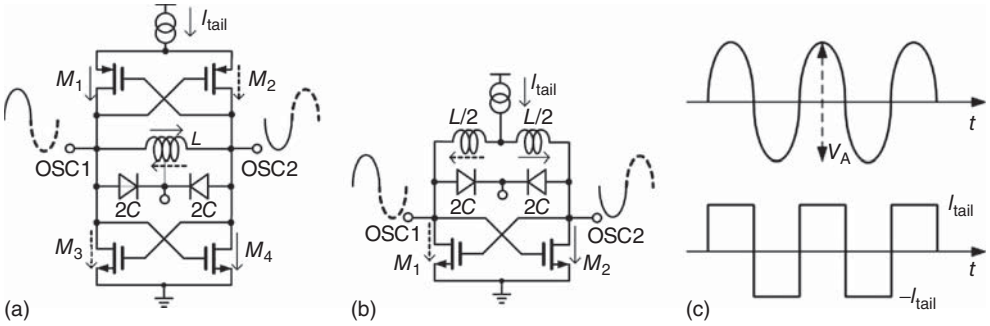


Figure 12.33. Current switching in complementary (a) and NMOS-only (b) cross-coupled VCO, and output oscillation voltage and tank current waveforms (c).

current I_{tail} is drawn through the LC tank twice every period, while in the NMOS-only case, the current is drawn through only half of the tank in each period.

Figure 12.33(a) shows the behavior of the current switching in the complimentary VCO topology with the assumption that all the transistors operate as ideal switches during the steady-state oscillation. Figure 12.33(b) shows the current switching in the NMOS-only VCO topology, which has the same LC tank as the complementary topology but with the current source feeding at the center point joining two equal inductors $L/2$ making up the tank's inductor L. In Figure 12.33(a), when the two differential outputs OSC_1 and OSC_2 undergo positive and negative oscillation swing, respectively, the transistors M_1 and M_4 are switched on while M_2 and M_3 are switched off, and the bias current, I_{tail}, flows from the bias current source to the ground through M_1, LC tank and M_4. On the other hand, when OSC_1 and OSC_2 have negative and positive oscillation swing, respectively, the bias current flows through M_2, LC tank and M_3 to the ground and the transistors M_1 and M_4 are switched off in this half period. Denoting the differential oscillation voltage as the transient voltage potential at OSC_1 with respect to OSC_2 and the positive current direction from OSC_1 to OSC_2, the oscillation voltage and current switching of Figure 12.33(a) can be described by Figure 12.33(c).

For the complimentary VCO, as shown in Figure 12.33(a), a current swing on the tank's resistor, R_p, generates a rectangular waveform for the tank current with peak current of $\pm I_{\text{tail}}$ due to the ideal switching of the transistors, as exhibited in Figure 12.33(c). The output peak-to-peak oscillation voltage is $V_A = 2R_p I_{\text{tail}}$, where R_p is the equivalent loss resistance of the tank and I_{tail} is the bias current drawn from the current source. This voltage waveform, however, is not rectangular, but approximates a sinusoid. This is due to the fact that the LC tank, through its filtering function, significantly suppresses the harmonic voltages, leaving only the fundamental voltage to appear at the two terminals of the tank. Hence, the output differential peak-to-peak voltage amplitude of the oscillation can be derived taking only the fundamental component as

$$V_A = \frac{4}{\pi} R_p I_{\text{tail}} \tag{12.120}$$

where the $2/\pi$ term comes from the Fourier coefficient of the signal at the fundamental frequency.

The oscillation voltage and current switching transient of the NMOS-only VCO in Figure 12.33(b) can also be described by Figure 12.33(c). The tank's current swing is the same as that in the complementary VCO. However, it is noted that the voltage amplitude is only a half of that of the complementary counterpart due to the fact that the bias current I_{tail} flows through only a half of the tank per cycle, as

$$V_A = \frac{2}{\pi} R_p I_{\text{tail}} \tag{12.121}$$

The available rms power produced by a resistor R_p with an AC voltage amplitude V_A across it can be expressed as

$$P_{\text{avs}} = \frac{V_A^2}{2R_p} \tag{12.122}$$

Now we recall the phase-noise expressions, such as (12.98), derived earlier, which contain the available power P_{avs}. (12.122), together with (12.98), (12.119), and (12.120), signifies that the complementary VCO topology has an intrinsic 6-dB better phase noise than the NMOS-only counterpart for the same current if the amplitude is current-limited. In the voltage-limitation region, the amplitude of the complementary VCO is limited by the power supply voltage V_{dd}, while the amplitude of the NMOS-only topology is limited by V_{gs} of the NMOS transistor.

From the calculation results shown in Figure 12.32, it can be concluded that if the power supply voltage is sufficiently high, the complementary VCO presents the best power-efficient solution for a comparable good phase noise. When the power supply voltage is low – for example, lower than $V_{\text{dd}} = 1.8$ V for the 0.25-μm CMOS process used in a cross-coupled VCO design example described later, the NMOS-only topology can achieve lower phase noise at the expense of higher current.

Phase Noise

As discussed earlier, flicker or $1/f$ noise plays the most significant role in the phase noise near the oscillator or carrier frequency. This noise is up-converted to the phase noise through a mixing process with an RF signal, and the existence of this phase noise or its magnitude depends upon the balance of the cross-coupled circuitry. The mechanism of the flicker noise upconversion in cross-coupled oscillators is explained in [7]. When a cross-coupled VCO circuit is unbalanced, an oscillation would occur at the common-mode node of the current source at the second harmonic of the carrier frequency, because the current source is pulled every time one of the CMOS transistor pair switches on while the other is off. Through channel-length modulation, the noise of the tail current source is upconverted to the second harmonic. This upconverted noise then enters the LC tank and is mixed with an RF signal at the carrier frequency, resulting in phase noise sidebands around the carrier frequency and its third harmonic. Therefore, to minimize the upconversion of flicker noise from the tail current source, balance must be preserved for the cross-coupled VCO, implying that all even harmonics are suppressed. It is noted that odd harmonics are not important to flicker noise upconversion because they do not affect the symmetry of the voltage oscillation across the LC tank.

Among the transistors employed in the cross-coupled VCO, flicker noise from the tail current source is the main contributor to the phase noise in the $1/f^3$ noise region. The CMOS transistors in the cross-coupled transistor pair function as current switching during the steady-state oscillation and, therefore, its contribution to noise is small during the steady-state operation. As indicated earlier, flicker noise is a sort of correlated noise and can exist only in nonlinear time variant systems. For transistors with ideal switching, flicker noise vanishes. Practical transistors, however, do not provide ideal switching, causing a small amount of the CMOS transistors' flicker noise to be upconverted into the phase noise around the oscillator frequency.

Another $1/f^3$ phase noise contribution, which does not originate from flicker noise, is "parametric noise" [8] such as noise produced by the nonlinearity of the varactors used in VCO. The varactor capacitance, as part of the tank, affects the oscillation frequency and, although is controlled by an applied DC voltage, is also affected by the RF voltage imposed across it. While the steady-state oscillation frequency remains the same due to the varactor's principal capacitance resulting from a fixed DC bias condition, a large RF sinusoidal VCO signal can change the varactor capacitance with time within one oscillation period, hence affecting the instantaneous oscillation frequency. Due to the strong swing of the RF signal across the varactor in the tank, the capacitance changes as $C_o(t) = C_o + C(t)$, where C_o is the principal capacitance, resulted from the varactor's bias voltage, corresponding to the steady-state oscillation frequency and $C(t)$ is the capacitance change due to the RF voltage. Consequently, the frequency varies as $f_o(t) = f_o + \Delta f(t)$, where f_o is the oscillation frequency and $\Delta f(t)$ is the noise produced by $C(t)$ on the frequency. This noise usually has a flicker noise-like PSD. An increase in capacitance results in a decrease in frequency, and this resultant frequency change behaves as noise around the oscillation frequency. It is noted that this phase noise is not actually caused by noise that gives rise to the phase noise discussed previously, but by the nonlinearity of the varactor, which is invoked from the oscillator's RF voltage swing. This kind of parametric noise combines with the flicker noise of the tail current source to produce a more pronounced effect on the overall phase noise of the VCO. The parametric noise and its relative contribution to the $1/f^3$ phase noise are difficult to model by

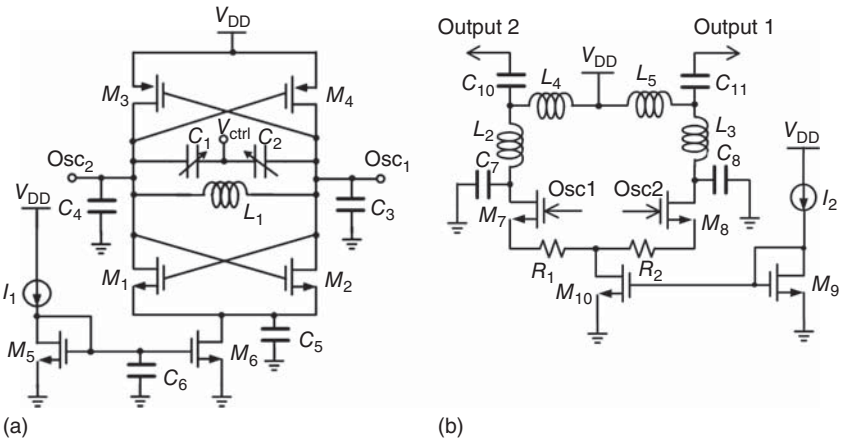


Figure 12.34. Complementary cross-coupled VCO with output differential buffer (a) and the differential buffer (b).

simple equations. They, however, can be approximately analyzed through simulations using custom-written or commercially available computer programs.

Design Example

In this section, we present a design example for a complementary cross-coupled CMOS VCO at 5.8-GHz center frequency in a 0.25-μm CMOS process. The complementary VCO can provide a large tuning frequency range for a given LC tank as the AC current goes through the entire tank inductor, resulting in a broad bandwidth.

Design. Our design goals for the complementary VCO are to achieve a wide frequency tuning range and good phase noise with minimum current consumption at 5.8-GHz. Figure 12.34 shows the VCO topology with an output buffer, which drives a pair of 50-Ω loads. The buffer facilitates measurement and connection with other subsequent 50-Ω components without any impact on the tuning range and phase noise of the VCO. This buffer, however, may not be needed when the VCO is directly integrated with other components – for instance, in a single-chip transmitter when all components are integrated or co-designed as a single complete circuitry. The VCO consists of a tank and two pairs of MOSFETs. The tank consists of an inductor, L_1, and a varactor pair, C_1 and C_2, representing two capacitors. The MOSFET pairs comprise a cross-coupled NMOS pair, M_1 and M_2, and a cross-coupled PMOS pair, M_3 and M_4, to form a positive feedback network to produce oscillation when connected with the tank. The bias current, I_{tail}, is mainly determined by the current mirror comprising transistors M_5 and M_6.

The parameters of the six transistors used in the cross-coupled pairs and current mirror and the varactors in the tank are interrelated in the design and affect the overall performance of the VCO. Particularly, it is apparent that the design requires a tank with high quality factor. The loss of the tank is the dominant factor affecting the VCO phase noise and mainly caused by the inductor, L_1, for given varactors. Moreover, the four cross-coupled transistors need to be optimized together with the bias current to reach an optimal operating point indicated by the symbol "Δ" in Figure 12.32 (on the dashed curve for the complementary VCO topology). Finally, the flicker noise upconversion from the current-bias transistor, M_6, also needs to be minimized.

The varactors of the tank, represented by C_1 and C_2 in Figure 12.34(a), are implemented by a pair of PMOS transistors with the source, drain, and bulk connected to each other, as shown in Figure 12.35. The common connection is also connected to the DC control voltage, V_{ctrl}. The capacitance of each varactor is determined by a DC voltage, provided through V_{ctrl}, across the gate and bulk, while the DC bias voltage V_{gs} of the gate is kept constant by the operating point of the VCO circuitry. Changing the DC control voltage V_{ctrl} varies the capacitance of the tank and hence the oscillation frequency, thereby resulting in frequency tuning. The biggest available size of varactors in the employed process is chosen to produce the widest range for the

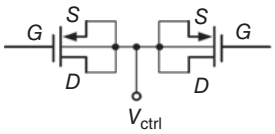


Figure 12.35. Varactor pair implemented by PMOSFETs.

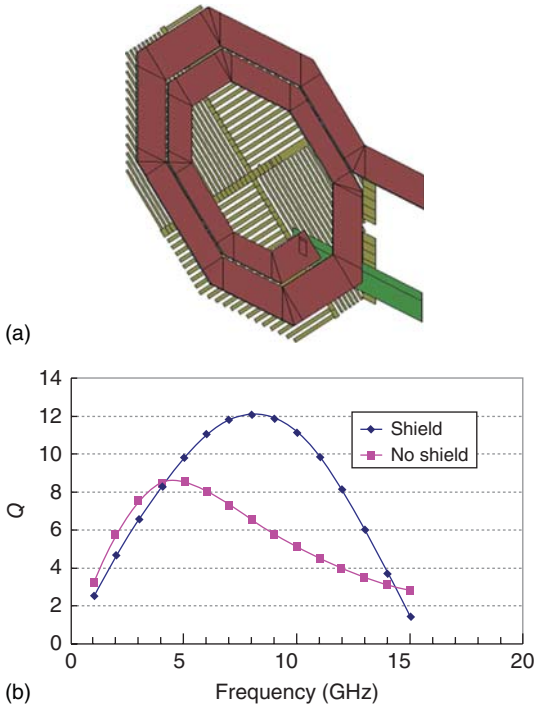


Figure 12.36. Tank's spiral inductor with PGS (a) and calculated Q (b).

capacitance in order to achieve the largest frequency tuning range. As can be seen in Figure 12.34, the gates of the PMOSFET-varactors are connected with the strong oscillation nodes (OSC_1 and OSC_2). Although the same capacitance can be reached by a reverse connection (i.e., gate connecting to V_{ctrl} instead of drain), such connection would cause much loss when a strong RF voltage oscillation occurs on the bulk, which is the VCO's output, because of its conductive substrate. Consequently, the oscillation amplitude would be lower and the phase noise would be worse due to the loss of the varactors in the tank. When the bulks of the two identical PMOSFETs are connected as shown in Figure 12.35, the feed-through signals of the balanced differential oscillation voltages are combined with 180-deg out of phase at the connection node of the bulks and hence cancel each other, eliminating signal leakage and producing higher signal amplitude with lower phase noise at RF frequencies.

An octagonal spiral inductor is used for the tank. Figure 12.36(a) shows its layout on the topmost metallization layer with a pattern ground shield (PGS) on the poly-silicon layer. Figure 12.36(b) shows the calculated Q of the inductor with and without PGS. The peak Q of the PGS inductor is more than that of the non-PGS inductor due to the loss reduction resulting from the PGS as discussed in Section Substrate Resistance R_{sub} of Chapter 3 (Lumped Elements). The PGS inductor's maximum Q is 12 from 5 to 11 GHz.

The VCO current-bias transistor, M_6 in Figure 12.34(a), is chosen to have a relatively large size with Length (L)/Width (W) = 0.75 μm/500 μm for small flicker noise. A current ratio of 10 is obtained by choosing the size for the NMOS transistor M_5 as 0.75 μm/50 μm. A large capacitor, C_5, is connected with the drain of M_6 to provide an RF short to the ground for the common-mode current at the second harmonic so that even harmonics are blocked from injecting into the LC tank, hence minimizing the flicker-noise upconversion.

The NMOS cross-coupled transistors, M_1 and M_2, are chosen to be identical for achieving good differential balance. The PMOS transistors, M_3 and M_4, are particularly selected so that $G_{\text{Mp}} = G_{\text{Mn}} = 3/2R_p$ at the design frequency of 5.8-GHz, where G_{Mp} and G_{Mn} are the transconductances of the PMOS and NMOS transistors, respectively, which result in a symmetric oscillation in which the positive and negative peak amplitudes of the voltage are equal. These symmetrically differential balanced oscillations inject less noise into the LC tank from the amplifier's transistors than any imbalance oscillations. The combination of the transconductances of the NMOS and PMOS transistor pairs are thus equal to three times of the loss resistance R_p in the LC tank, enabling a stable oscillation to occur, as mentioned earlier.

The bias current, I_{tail}, is optimized along with the size of the cross-coupled NMOS and PMOS transistor pairs to reach an optimum operating point as indicated by the "Δ" mark in Figure 12.32. It is particularly noted that increasing I_{tail} over such optimal value results in a stronger oscillation amplitude but with a poorer phase noise.

The differential oscillator's outputs, OSC_1 and OSC_2 in Figure 12.34, are DC-coupled to the input ports of the following differential buffer, which are at the gates of the transistors M_7 and M_8, respectively, as shown in Figure 12.34(b). In order to keep the quality factor of the LC tank unaffected by the low input impedance of the buffer, which is actually a differential common-source amplifier pair, two on-chip MIM capacitors, C_3 and C_4, are placed in parallel with the input ports of the buffer, enabling the loading of the VCO to be dominated by these capacitances which provide high impedances at the design frequency.

The complete complementary cross-coupled VCO including buffer is designed and fabricated using the TSMC 0.25-μm CMOS process [9]. Table 12.2 lists the parameters of the designed VCO without the buffer.

The VCO's phase noise at 5.8-GHz is simulated using Cadence. Table 12.3 summarizes the phase noise and noise contributions from the VCO elements at 1-MHz offset.

The differential common-source buffer is designed to take very little current and primarily intended for attenuating the strong oscillation voltage produced by the VCO to a small output level. To this end, its transistors M_7 and M_8, as seen in Figure 12.34(b), are chosen to have a small gate width needed for small transconductance. The 50-Ω impedance matching at the design frequency of 5.8-GHz is achieved by matching networks comprising the MIM capacitors C_7 and C_8 and spiral inductors L_2, L_3, L_4, and L_5. Source-degeneration resistors R_1 and R_2 help keep the oscillation undistorted at the output port when there is a strong oscillation at the input port of the buffer.

Table 12.4 lists the parameters of the designed CMOS buffer. Figure 12.37 shows the calculated output return loss of the buffer with respect to 50-Ω impedance.

Figure 12.38 shows the microphotograph of the complementary cross-coupled LC VCO with output buffer, where one of the outputs of the buffer is terminated with an on-chip 50-Ω resistor.

TABLE 12.2. Summary of the Designed CMOS Cross-Coupled VCO's Parameters

Circuit element	Element value
M_1, M_2	120-μm gate width 0.25-μm gate length
M_3, M_4	200-μm gate width 0.25-μm gate length
M_5	50-μm gate width 0.75-μm gate length
M_6	500-μm gate width 0.75-μm gate length
C_1, C_2	PMOS, 180-μm gate width 0.75-μm gate length
L_1	1 nH
C_3, C_4	MIM capacitor, 200 fF
C_5, C_6	MIM capacitor , 20 pF
V_{dd}	1.8 V
I_{tail}	4 mA

TABLE 12.3. Phase Noise and Noise Contributions (V^2_{rms}/Hz) at 5.8 GHz with 4-mA Bias Current

Circuit element	Noise contribution	Percentage
R_p	6.05×10^{-15}	42.9%
M_1, M_2	2.98×10^{-15}	21.2%
M_3, M_4	5.06×10^{-15}	35.9%
M_6	N/A	N/A
Peak–peak voltage amplitude V_A	980 mV	
Phase noise at 1 MHz	−119 dBc/Hz	

TABLE 12.4. Summary of the CMOS Buffer's Designed Parameters

Circuit element	Element value
M_7, M_8	20-μm gate width 0.25-μm gate length
M_9	50-μm gate width 0.75-μm gate length
M_{10}	2500-μm gate width 0.75-μm gate length
R_1, R_2	Poly resistor 12 Ω
C_7, C_8	MIM, 120 fF
L_2, L_3	2.3 nH
L_4, L_5	3.7 nH
C_6, C_{10}	MIM, 2 pF
V_{dd}	1.8 V
I_{tail}	Tunable

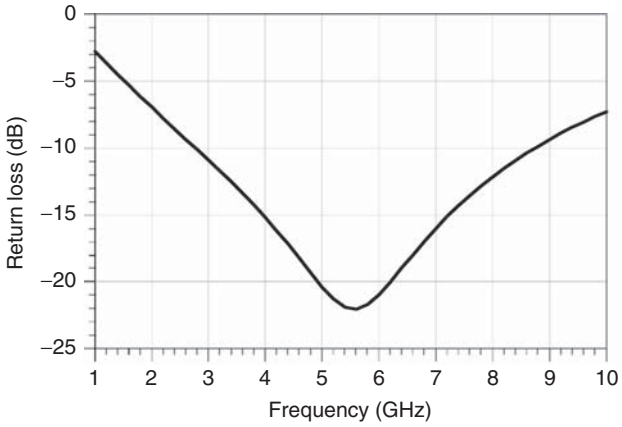


Figure 12.37. Calculated return loss at the output port of the buffer.

Performance. The fabricated cross-coupled VCO was measured on-wafer. Figure 12.39 shows the measured output spectrum. Figure 12.40 shows the measured phase noise at 5.8-GHz. The phase noise at 1-MHz offset is −117 dBc/Hz, 2 dB lower than the calculation. The $1/f^3$ corner frequency is at around 500 KHz. Figure 12.41 presents the measured frequency tuning range versus the control voltage. The measured center frequency of the VCO is 5.2 GHz, 600 MHz lower than the simulation. The primary reason for the difference between the measured and calculated results is the extra inductances caused by the connecting tracks between the spiral inductor and two varactors in the LC tank were not extracted precisely by Cadence at high frequencies, resulting in a lower resonance frequency than expected. This Cadence-related issue may be

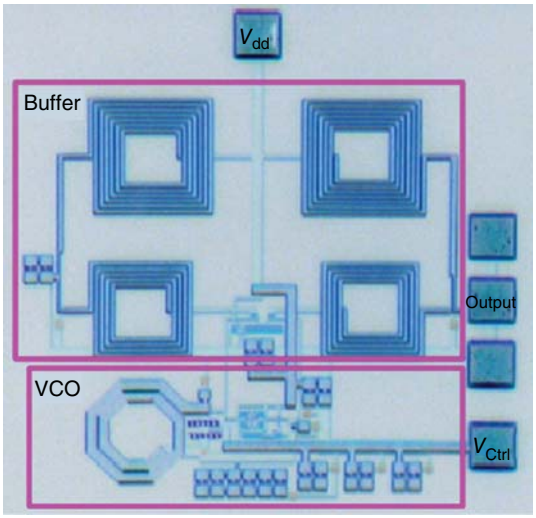


Figure 12.38. Microphotograph of the complementary cross-coupled VCO with buffer. The die area without pads is 320 μm by 300 μm.

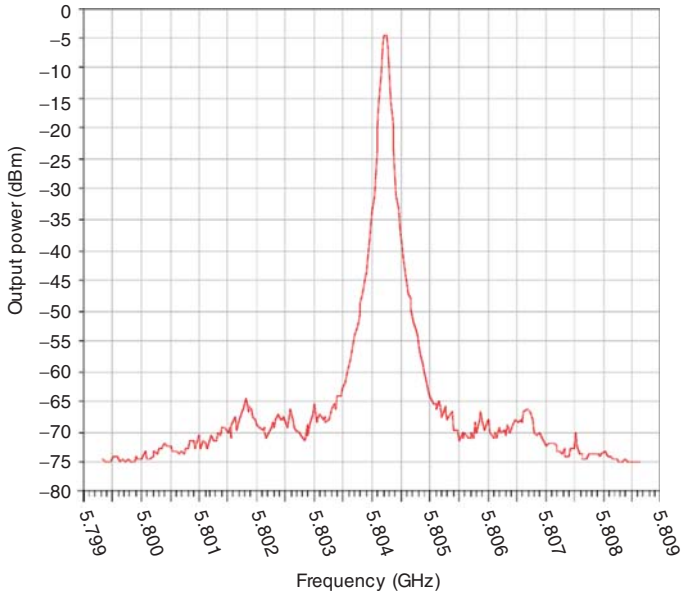


Figure 12.39. Measured output spectrum with $V_{dd} = 1.8$ V and $V_{ctrl} = 1.75$ V.

overcome by extensive EM simulations or by giving 10% margin to the Cadence calculation. The measured tuning range is 25% of the center frequency, covering 4.6 to 5.9 GHz for control voltage from 0 to 2 V.

12.4.2 Distributed Oscillators

Distributed oscillators are based on the concepts of feedback oscillators and distributed amplifiers described earlier in this chapter and in Chapter 11, respectively. Distributed oscillators were first proposed by Skvor et al. [10]. Numerous studies have been conducted into their operation, design and analysis of important circuit parameters, for example, Kleveland et al. [11] and Wu et al. [12]. The distributed topology presents a good opportunity to realize oscillators at high frequencies in CMOS where device technology limitations may limit the performance of traditional LC-oscillators. Distributed topology implemented for VCOs exhibits wide tuning ranges, which are attractive for broadband wireless applications. Distributed oscillators, however,

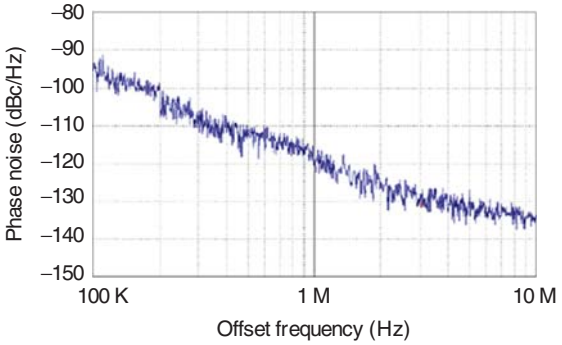


Figure 12.40. Measured phase noise at 5.8-GHz.

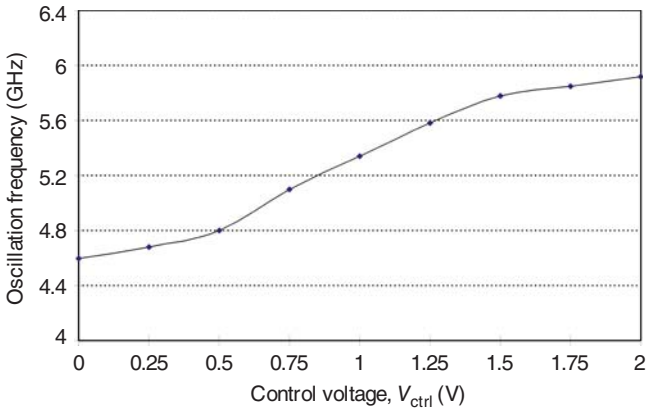


Figure 12.41. Measured VCO tuning range versus control voltage.

suffer from poor properties of large size and power consumption, just like other active distributed circuits such as amplifiers and mixers.

12.4.2.1 Theory. Distributed oscillators can be realized as a feedback oscillator and can be implemented by connecting the output of a distributed amplifier to its input, so as to form a positive feedback loop which sustains oscillation. The oscillation, as discussed earlier for the feedback oscillators, would then be dependent upon the total gain and phase of the closed loop.

We consider general distributed oscillators as shown in Figure 12.42 having a feedback between the input and output of a distributed amplifier. The gain of the constituent distributed amplifier is given by (11.243) in Chapter 11 as

$$G = g_m^2 Z_g Z_d \left| \frac{\overline{\gamma}_g \left(e^{-\overline{\gamma}_g N} - e^{-\overline{\gamma}_d N} \right)}{\overline{\gamma}_g^2 - \overline{\gamma}_d^2} \right|^2 \tag{12.123}$$

where Z_g and Z_d are the characteristic impedances of the gate and drain lines, respectively; $\overline{\gamma}_{g(d)} \equiv \gamma_{g(d)} \ell_{g(d)} = \overline{\alpha}_{g(d)} + j\overline{\beta}_{g(od)}$ is the (total) propagation constant of each actual or synthetic gate (drain) line section with $\overline{\alpha}_{g(d)}$ and $\overline{\beta}_{g(d)}$ being the corresponding (total) attenuation constant (in neper) and phase constant (in radian) of each actual or synthetic transmission-line section; N is the number of cells or stages; and g_m is the small-signal transconductance of each gain cell which is the transistor in Figure 12.42. This gain can be rewritten following [12] as

$$G = -G_m Z_{gd} e^{-(\overline{\gamma}_d + \overline{\gamma}_g)/2} \frac{e^{-N\overline{\gamma}_d} - e^{-N\overline{\gamma}_g}}{e^{-\overline{\gamma}_d} - e^{-\overline{\gamma}_g}} \tag{12.124}$$

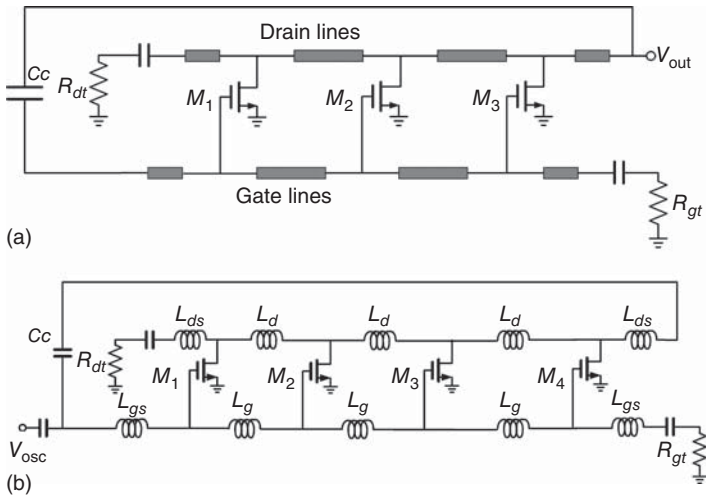


Figure 12.42. General distributed oscillator based on feedback mechanism using transmission lines (a) and inductors (b).

where $Z_{gd} = Z_g \| Z_d$ and G_m is the large-signal transconductance of each transistor. In typical distributed-amplifier design, the propagation constants for each section of the gate and drain lines are approximately equal ($\overline{\gamma}_g \simeq \overline{\gamma}_d \equiv \overline{\gamma} = \overline{\alpha} + j\overline{\beta}$). Under this condition, the gain in (12.124) can be simplified as

$$G = -NG_m Z_{gd} e^{-N\overline{\gamma}} = -NG_m Z_{gd} e^{-N\overline{\alpha}} e^{-jN\overline{\beta}} \tag{12.125}$$

Now recalling the feedback oscillator's condition stated in (12.28), or (12.29) and (12.30), for oscillations to occur, we must have

$$NG_m Z_{gd} e^{-N\overline{\alpha}} e^{-jN\overline{\beta}} = -1 \tag{12.126}$$

In order to satisfy this condition, the imaginary part of the left-hand-side of (12.126) must be equal to zero. This implies, assuming lossless gate and drain lines (i.e., Z_g and Z_d are real), that

$$N\beta\ell = \pi \tag{12.127}$$

Substituting $\beta = 2\pi f / v_p$, where v_p is the phase velocity of the gate and drain lines, in (12.127), we then obtain an expression for the frequency of oscillation as

$$f_o = \frac{v_p}{2N\ell} \tag{12.128}$$

Equation (12.128) implies that, in order to obtain maximum frequency of oscillation, the number of stages as well as segment length must be minimized. Equation (12.128) can also be rewritten as

$$f_o = \frac{1}{\sqrt{L_{gd} C_{gd}}} \tag{12.129}$$

where $L_{gd} = 2N\ell L$ and $C_{gd} = 2N\ell C$, with L and C being the inductance and capacitance per unit length of the transmission lines, represent the total inductance and capacitance of the combined gate and drain lines, respectively.[9] Equation (12.129) shows that using smaller transistors, shorter segment and greater number of stages would increase the oscillation frequency.

[9]These inductance and capacitance are the total inductance and capacitance according to the transmission-line equivalent circuit based on the per-unit-length inductance and capacitance for an infinitesimal length. They are not the inductance and capacitance exhibited by the two transmission lines.

The number of stages of the distributed oscillator can be determined from (12.126) as

$$N = \frac{-1}{G_m Z_{gd} e^{-N\overline{\gamma}}} \tag{12.130}$$

Approximating $e^{-N\overline{\gamma}}$ as $(1 - N\overline{\gamma}/2)$ for small $N\overline{\gamma}$, ignoring higher order terms, we get

$$N = \frac{-1}{G_m Z_{gd}(1 - N\overline{\gamma}/2)} \tag{12.131}$$

Solving for N by taking quadratic roots, we obtain

$$N = \frac{1 + mod(\sqrt{1 + 2\overline{\gamma} G_m Z_{gd}})}{\overline{\gamma}} \tag{12.132}$$

Making a reasonable assumption that the negative solution is not practically relevant, and expanding the square-root term and considering only the real values, we obtain

$$N = mod\left(\frac{2}{\overline{\gamma}} + \frac{1}{G_m Z_{gd}}\right) \tag{12.133}$$

It is noted that the number of stages of a distributed VCO affects the amplitude and phase noise of the output signal. A greater number of stages increase the losses through the actual transmission lines or inductors making up the synthetic transmission lines as well as the noise associated with the transistors.

Distributed oscillators can also be implemented as VCOs. There are several viable options in the frequency-tuning operation of distributed VCOs. Using external varactors proves detrimental to the overall operation as they tend to degrade the oscillation frequency by increasing the total capacitance. An efficient way to tune the oscillator is to tune its intrinsic parasitics at the drain-source and gate-source nodes. However, doing so might alter the operating point of the transistors. In order to prevent this, the transistors could be biased separately with current sources while DC voltages could be applied which could tune the transistor parasitics appropriately. Tuning ranges up to 14% have been obtained by just following this method [12].

12.4.2.2 Distributed Oscillators Versus Cross-Coupled Oscillators. Cross-coupled oscillators described in Section 12.4.1 and distributed oscillators have their own advantages and disadvantages, and their choice depends on various design and operating factors. The cross-coupled oscillators provide superior phase noise at lower power dissipation as compared to distributed oscillators. However, they have several shortcomings. The main drawbacks are limited frequency tuning range and excessive dependence on the transistor's speed. Also, for multi-gigahertz applications, the smaller inductor values of the tank necessitated by the oscillator that controls the frequency of operation lead to higher power dissipation as the inductor operates in a parallel LC tank. On the contrary, distributed oscillators offer the possibility of wideband tuning range at a moderate level of power consumption (for proper design). They also facilitate the possibility of multiple-phase signals using different stages that eliminate the need for power consuming divide-by-2 digital circuits or lossy poly phase filters. Also, the operating frequencies of the distributed oscillators depend on the round trip time delay of a signal; they do not depend on the speed of the transistor alone and can be fine tuned by the additional degree of design freedom. However, their bulky size, owing to their dependence on transmission lines or inductors and usage of many gain cells, and large power consumption coming from many transistors are the two main aspects that need to be addressed in order to make them compatible with cross-coupled oscillators.

12.4.2.3 Design Example. In this section, a 15-GHz distributed VCO with a wide tuning range is designed using a 0.18-μm CMOS process and presented as a design example.

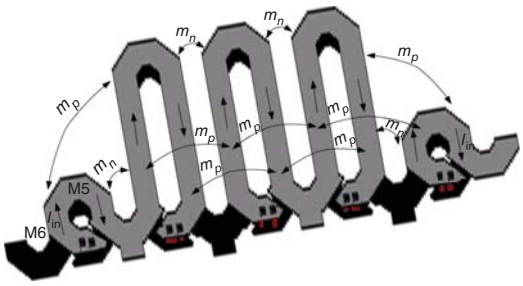


Figure 12.43. Integrated inductor segment using multi-layered inductors

Design Principles

The constituent distributed amplifiers of distributed oscillators can be implemented using either actual transmission lines or inductors, in conjunction with the transistors' capacitors, to realize synthetic transmission lines. Utilizing inductors, instead of actual transmission lines, is a good alternative as they tend to provide greater inductance per unit area than transmission lines, hence conserving chip area. However, in most modern CMOS technologies, single-layered inductors tend to occupy large chip areas since even most single-layered inductors also suffer from poor chip area consumption. Using an integrated multilayered inductor approach seems to be the most economical solution. Some important properties of these multilayered inductors and their use in distributed amplifiers are discussed in Section An Ultra-Compact Distributed Amplifier of Chapter 11. The designed distributed VCO employs the multi-layered inductors instead of transmission lines in order to significantly minimize the circuit dimensions.

Figure 12.43 shows the integrated multilayered inductor segment. This inductor is first analyzed in order to find out the impact of negative mutual coupling on the overall performance. It is analytically calculated that the impact of the first-order positive mutual coupling overwhelms the net negative coupling, thereby causing a slight increase in the overall inductance. Full wave EM simulations are then performed using IE3D [13] to carefully optimize and design this structure. The inductance is intentionally kept low as it entails several advantages apart from the obvious size reduction. First, lower inductance makes it easier to enhance the quality factor as well as self-resonant frequency of the inductor, independent of the CMOS technology of implementation and without any further trade-off. Second, it makes it possible to use transistors of smaller W/L aspect ratios which enable lower power consumption. The terminal inductors are kept at 0.3 times the value of the central inductors for matching purposes. The larger inductances are 0.75 nH while the smaller ones are just 0.25 nH. The oscillation frequency depends on the round trip delay time of the drain and gate inductor segments and is given by (12.128). The gate capacitances are controlled by an external bias, which provides the tuning mechanism for this oscillator. A concern here is that any strong variation in the gate voltage to get additional tuning might displace the transistors from their DC operating points. This could be averted by fixing the drain current with constant current sources.

The VCO is primarily designed as a high gain, broadband, unstable distributed amplifier. A four-stage implementation is implemented as the transistor aspect ratios are kept small for power consumption purposes. The W/L ratio of each transistor that satisfied the power and matching requirements is 128 μm/0.18 μm. In order to further minimize the dimensions, the lengthy drain to gate feedback line is replaced by transistor gate-to-gate inductance segments, which are reverse connected as shown in the schematic of Figure 12.44. Successive EM iterations of the multi-layer structure are imported into Agilent's ADS [14] to optimize the transistor gain blocks of the circuit. A simple common-source transistor is used as the transistor gain cell. The high-frequency transistor models are obtained from the design kit provided by JAZZ CA18HR 0.18-μm CMOS process [15].

Implementation and Results

After the multilayered inductor segment is laid out in Cadence, post-layout simulations are carried out by re-performing simulations in IE3D and imported to ADS after their layout in Cadence. The total chip area is

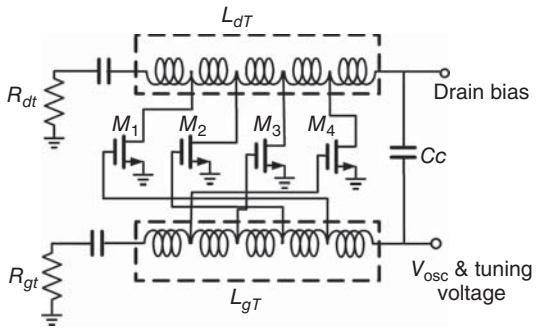


Figure 12.44. Schematic of the multilayered inductor-based distributed VCO.

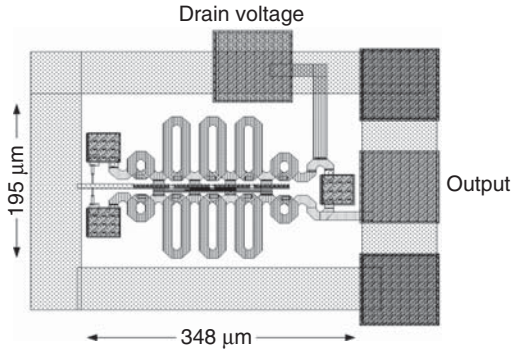


Figure 12.45. Layout of the multilayered inductor-based DVCO.

about 0.08 mm^2, including the RF pads, while the core occupies only 0.06 mm^2 as seen in Figure 12.45. The post-layout simulated results for the VCO spectrum, tuning and phase noise are shown in Figure 12.46. The VCO generates a 12% tuning range from 14.1 to 15.8-GHz, centered around 14.9 GHz, within reasonable limits of current consumption. The output spectrum shows a peak harmonic at 14.9 GHz, while the phase noise at 1-MHz offset is −100.2 dBc/Hz. This performance is facilitated by a current consumption of 19 mA from a 1.8-V voltage source.

12.4.3 Push-Push Oscillators

Typical oscillators including the feedback, cross-coupled and distributed oscillators discussed previously are designed to operate at fundamental frequencies. Push-push oscillators, on the other hand, are designed to produce signals at even harmonics, most often second harmonic, of the fundamental frequencies. Push-push oscillators are attractive for various applications, particularly in the high RF range where typical oscillators cannot generate sufficiently high (fundamental) frequencies due to the limited operating frequencies of active devices dictated by the frequency of unity current-gain f_T. Other interesting properties of push-push oscillators include the use of resonators operating at a half of the operating frequency, hence having increased quality factor, and lower phase noise as compared to their constituent oscillators. Moreover, although push-push oscillators are physically and electrically balanced, they operate as a single-ended circuit with only one output port instead of two differential ports, and are thus convenient in certain circuit environments and usage.

12.4.3.1 Analysis. A push-push oscillator consists of two identical oscillators. Each constituent oscillator or one half of a push-push oscillator can be a simple transistor or a more complicated oscillator circuit such as a cross-coupled or feedback oscillator, and can be designed for either narrow or wideband operation with or without frequency-tuning capability. Figure 12.47 shows a general push-push oscillator comprising two identical oscillators connected to a single load Z_L via an output network. For the convenience of analysis

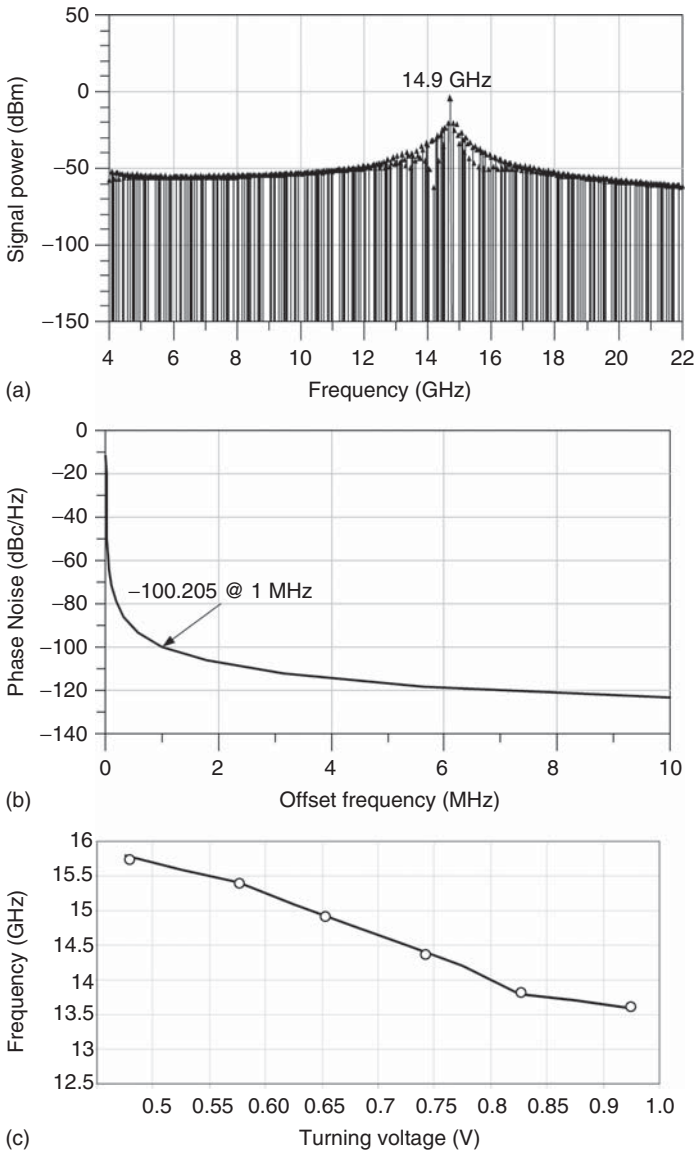


Figure 12.46. Simulation results of the multilayered inductor-based distributed VCO: (a) output spectrum, (b) phase noise and (c) tuning range.

without loss of generality, we consider a push-push oscillator consisting of two identical oscillators as shown in Figure 12.48(a), in which each oscillator is represented by the same circuit as shown in Figure 12.5(a). We assume that the output network consists of two symmetrical identical "half output networks" and the load impedance Z_L is replaced with two parallel impedances of $2Z_L$ each. The push-push oscillator is thus a perfectly balanced structure with respect to the symmetry line. It is noted that a non-resonant terminating network can also be used in lieu of the resonator at the gate and the matching network can be a combination of a resonator and matching network, which may also include a phasing circuitry to enable proper phase for the fundamental signals. The two individual oscillators' output signals, V_{o1} and V_{o2}, are combined through an output network, which may represent a power combiner or a direct connection through a resonator or joining transmission lines or elements.

The push-push oscillator, as can be seen in Figure 12.48(a), has several unique characteristics as compared to its constituent oscillator elements operating alone due to its balanced nature. It possesses inherent rejection or immunity from "signal disturbances" common to both constituent oscillators (common-mode signals) such

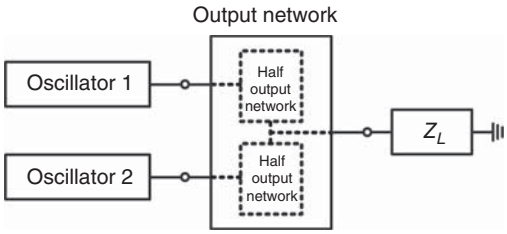


Figure 12.47. A general push–push oscillator topology.

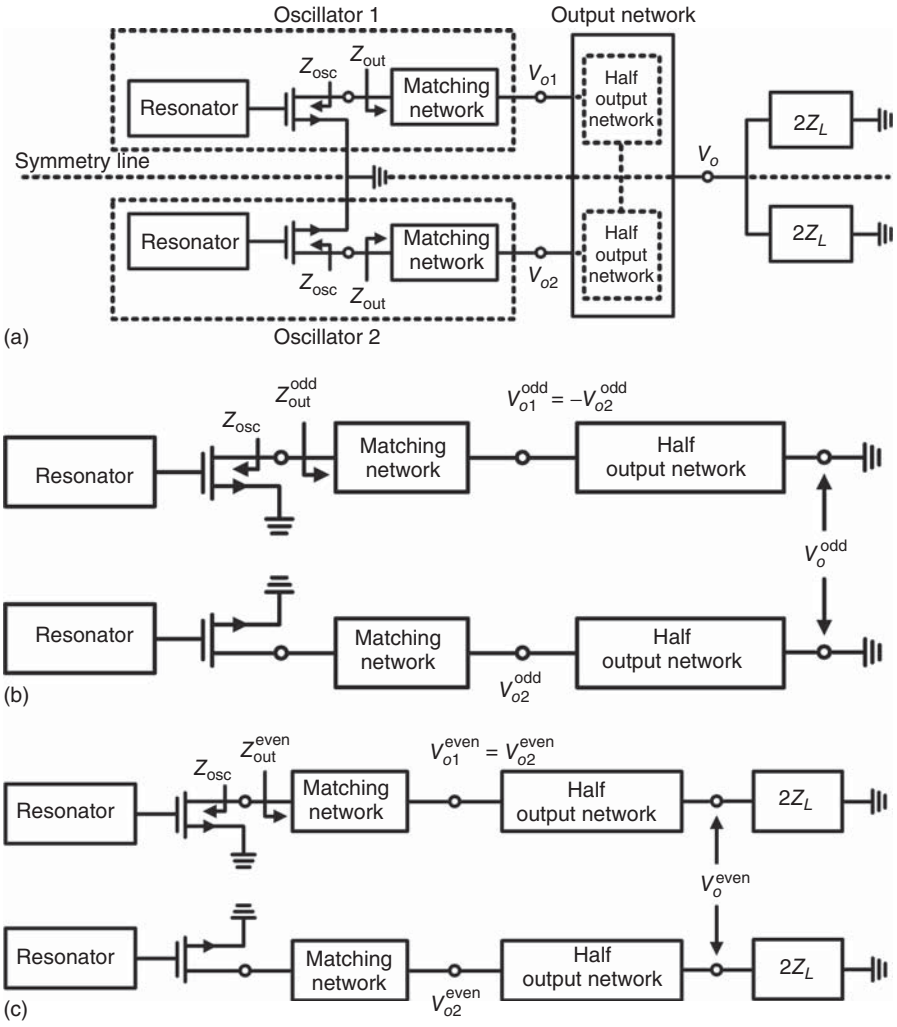


Figure 12.48. A push-push oscillator (a) and its equivalent odd-mode (b) and even-mode (c) circuits. $Z_{osc} = R_{osc} + jX_{osc}$ is the same for the even and odd modes. $Z_{out}^{odd(even)} = R_{out}^{odd(even)} + jX_{out}^{odd(even)}$.

as common noise from power supplies or unwanted couplings from nearby circuits. This feature is very desirable in practical RFICs, especially those integrated directly with digital circuits on the same chip. It should be noted, however, that, even for ideal or perfectly balanced push-push oscillators, perfect common-mode rejection occurs only when the perturbed signals appear equally on both output signals. In order for this to happen, the physical structures delivering the perturbed signals must be perfectly symmetrical with respect to the two output ports.

We assume a perfectly balanced push-push oscillator with identical transistors and associated networks. We also assume that the two transistors are operated exactly under the same condition, so that all their electrical parameters are identical, and the oscillator's layout is perfectly symmetrical with respect to the symmetry line between the two halves. Under these conditions, the considered push-push oscillator has perfect symmetry – both electrically and physically. This of course occurs only in ideal situations.

Due to the push-push oscillator's perfect symmetry, there exist two possible operating modes for the oscillator, namely even and odd modes. These operating modes are independent of each other and the oscillation in each of these modes is made possible only under certain conditions dictated by the impedances presented to the unstable transistors according to the necessary conditions for oscillation described earlier. Particularly, a push-push oscillator is designed to operate in the odd mode while suppressing the even-mode oscillation. Following an analysis similar to that discussed for the differential amplifiers in Section 11.5.1 of Chapter 11 or symmetrical passive microwave components described in Chapter 8 (RF Passive Components), for instance, parallel-coupled directional couplers in Section 8.2.2.2, we can assume an electrical wall (or short circuit) and a magnetic wall (or open circuit) along the symmetry line of the oscillator for the odd mode and even mode, respectively. Accordingly, the push-push oscillator in Figure 12.48(a) can be decomposed into two separate identical independently operating subcircuits for each operating mode as shown in Figure 12.48(b) and (c).

Odd Mode

Under the assumption of odd-mode oscillation, the output voltages of the two constituent oscillators 1 and 2 can be expressed via power-series expansion as

$$V_{o1}^{\text{odd}} = \sum_{n=1}^{N} A_n \cos(n\omega t) \tag{12.134}$$

$$V_{o2}^{\text{odd}} = \sum_{n=1}^{N} (-1)^n A_n \cos(n\omega t) \tag{12.135}$$

respectively, where A's are constant coefficients, which show that the fundamental and odd-harmonic components of the signals produced by the two oscillators have equal amplitude but 180° out of phase. The corresponding output voltage V_o of the push-push oscillator (without the load connected) is obtained as the sum of the individual output voltages as

$$V_{o}^{\text{odd}} = V_{o1}^{\text{odd}} + V_{o2}^{\text{odd}} = 2 \sum_{n=2,4,6,\cdots}^{N} A_n \cos(n\omega t) \tag{12.136}$$

which contains only even-order terms. Push-push oscillators operating under the odd mode thus inherently suppress the fundamental signal and all of the spurious signals of odd orders, leaving the even-harmonics as the signal sources as compared to unbalanced oscillators which contain all fundamental and harmonic signals. Typically, the second harmonic is selected as the output signal due to its significantly higher amplitude than other even harmonics, which can be further suppressed using a filter. Other even harmonics instead of second harmonic can also be chosen using a proper filter at the output of the push-push oscillator. The output voltage with no fundamental and odd harmonic components implies that push-push oscillators have a virtual RF ground along the symmetry line. It should be noted that these ideal results are obtained assuming two identical constituent oscillators. Otherwise, the fundamental and odd-harmonic signals will appear, causing degradation to circuit performance. For good design, however, well balance between the two halves is maintained and so the fundamental signal is significantly low.

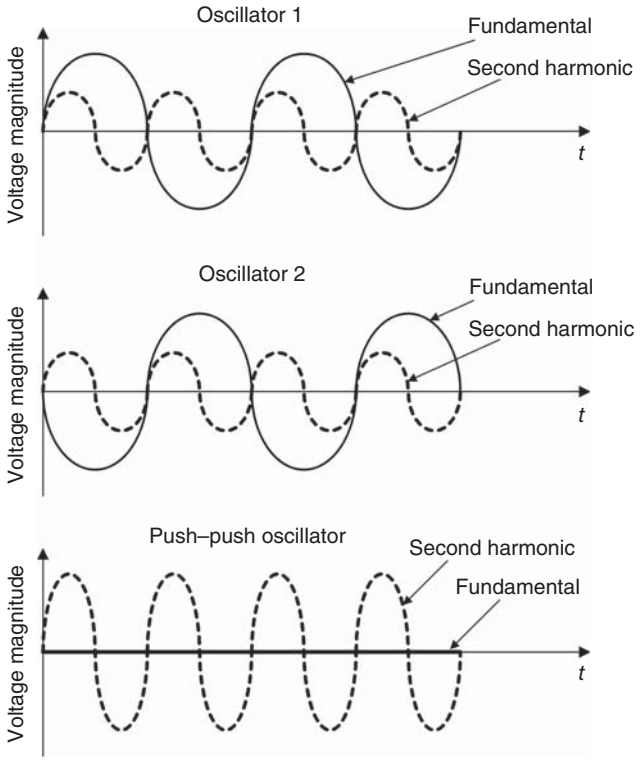


Figure 12.49. Fundamental and second-harmonic signals.

In order to sustain the odd-mode steady-state oscillation, the following conditions for oscillation as given in (12.14) must be satisfied at the steady-state voltage V_o for the odd-mode equivalent circuit of Figure 12.49(b):

$$R_{\text{out}}^{\text{odd}}(\omega_o) + R_{\text{osc}}(V_o, \omega_o) = 0$$
$$X_{\text{out}}^{\text{odd}}(\omega_o) + X_{\text{osc}}(V_o, \omega_o) = 0$$
$$\left.\frac{\partial R_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dX_{\text{out}}^{\text{odd}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} - \left.\frac{\partial X_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dR_{\text{out}}^{\text{odd}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} > 0 \tag{12.137}$$

The odd-mode circuit would begin an oscillation and produce a signal of voltage V at frequency ω_o when the following conditions are met:

$$|R_{\text{osc}}(V, \omega_o)| > R_{\text{out}}^{\text{odd}}(\omega_o)$$
$$|X_{\text{osc}}(V, \omega_o)| = -X_{\text{out}}^{\text{odd}}(\omega_o) \tag{12.138}$$

Under conditions (12.138), each constituent oscillator would oscillate and force the (composite) push-push oscillator to function in the odd mode, having opposite fundamental and equal second-harmonic signals for each individual oscillator. These fundamental and second-harmonic signals thus cancel and combine at the output of the push-push oscillator through the output network, respectively. Figure 12.49 shows the sketch of the output signals at the fundamental and second-harmonic frequencies.

Even Mode

Under the even-mode operation, the two constituent oscillators 1 and 2 oscillate in-phase and their two fundamental signals have equal amplitude and phase. As a result, these signals add constructively when are combined. This operating mode generates only fundamental signals and the (composite) push-push oscillator

can be considered a parallel combination of two individual oscillators. It is indeed not desired for push-push oscillators and need to be suppressed. That means the even-mode oscillation must be prevented and, in order to do that, the following conditions must be enforced:

$$|R_{\text{osc}}(V,\omega_o)| < R_{\text{out}}^{\text{even}}(\omega_o) \tag{12.139}$$

12.4.3.2 Design. The design of push-push oscillators typically starts with the design of their constituent oscillators and the design principle lies in the enforcement of the odd-mode oscillation while avoiding the even-mode oscillation. Therefore, the circuit needs to be designed to meet the odd-mode oscillation and even-mode non-oscillation conditions stated in (12.138) and (12.139), respectively, following the oscillator design procedure described earlier. Several particular design notes, however, need to be exercised. They are listed as follows.

Design of Constituent Oscillators

Each constituent oscillator occupies one half of the push-push oscillator. Although the load for the push-push oscillator can be anything depending on subsequent component or load, we assume a standard 50-Ω load for the push-push oscillator. The load for each constituent oscillator is thus 100 Ω since these individual oscillators are considered connecting in parallel at their outputs. The matching network of each individual oscillator, as shown in Figure 12.48(a), should provide a good match to the 100-Ω load at the second-harmonic frequency, not at the fundamental frequency. Compromise may need to be made for R_{osc} to achieve both desired oscillation at the fundamental frequency and good match at the second-harmonic frequency which is the desired operating frequency. The second-harmonic content of the output signals of the constituent oscillators should be enhanced as much as possible for power consideration. Although the second-harmonic signal is desired, this signal must come from the signals produced by the oscillation at the fundamental frequency only, not from a second-harmonic oscillation. Therefore, oscillation must be avoided at the second harmonic. In fact, oscillation should be avoided at all frequencies except fundamental frequency. To that end, the circuit must have, considering only frequencies higher than the fundamental frequency:

$$|R_{\text{osc}}(V,\omega)| < R_{\text{out}}(\omega), \quad \omega > \omega_o \tag{12.140}$$

where ω includes all harmonic frequencies and R_{out} indicates either $R_{\text{out}}^{\text{odd}}$ or $R_{\text{out}}^{\text{even}}$. Ideally, the circuit should not have negative resistance at frequencies other than the fundamental frequency to avoid instability causing spurious oscillations. That is, $R_{\text{osc}}(V,\omega) > 0$ for $\omega \neq \omega_o$. From (12.138)–(12.140), the following conditions must be achieved to produce oscillation at the fundamental frequency only:

$$\begin{aligned} &R_{\text{out}}^{\text{even}}(\omega_o) > |R_{\text{osc}}(V,\omega_o)| > R_{\text{out}}^{\text{odd}}(\omega_o) \\ &|X_{\text{osc}}(V,\omega_o)| = -X_{\text{out}}^{\text{odd}}(\omega_o) \\ &|R_{\text{osc}}(V,\omega)| < R_{\text{out}}^{\text{odd/even}}(\omega), \quad \omega > \omega_o \end{aligned} \tag{12.141}$$

Design of Push-Push Oscillator

The output ports of the two constituent oscillators are combined using an output network. The main function of this output network is to combine symmetrically (in-phase) the output signals of the two constituent oscillators at a common port at both the fundamental and second-harmonic frequencies. The fundamental energy is inherently suppressed by the circuit balance of the push-push oscillator. It should be noted that the matching to the load Z_L at the second-harmonic frequency, not the fundamental frequency, is concerned here; the output network should be optimized to deliver maximum power to the load at the second-harmonic frequency. The signal combination can be done by a direct connection of the two outputs; that is, the output network simply connects two individual outputs together. A power combiner can also be used for the output network to combine the two outputs. Using a properly designed power combiner avoids the potential problem of signal from one oscillator leaking on to the other oscillator. Another form of power combiner

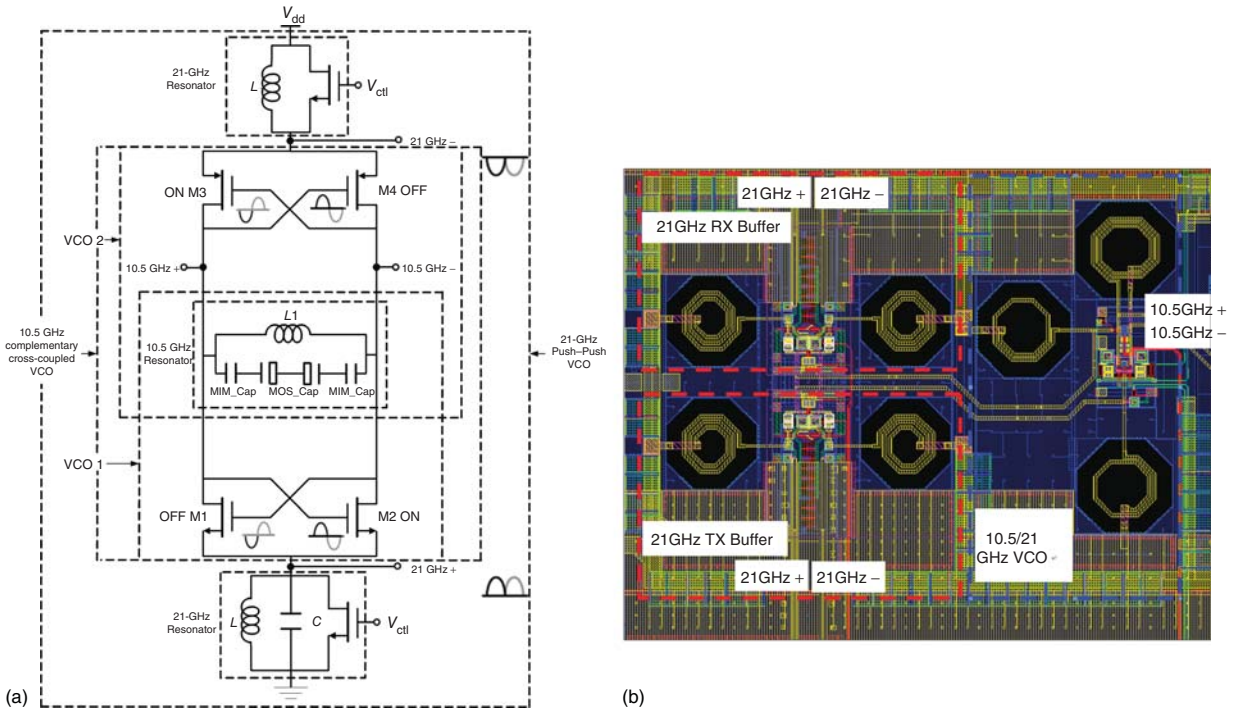


Figure 12.50. Schematic (a) and layout (b) of 10.5/21-GHz VCO. There are two 21-GHz buffers in the layout: 21 GHz TX buffer and 21 GHz RX buffer.

can be implemented using a distributed structure such as a resonator that has a common point representing a short- and open-circuit at the fundamental and second-harmonic frequencies, respectively. This particular structure allows the two out-of-phase fundamental signals to be canceled and the second-harmonic signals to be combined constructively at the common point. Such a circuit was implemented using a half-wavelength open-circuited resonator operating at the fundamental frequency [16]. This combining technique serves to provide a common resonator for both constituent oscillators while facilitating the signal combination through a direct combination. Although the fundamental signal is theoretically eliminated by the circuit balance, a filter may be needed at the push-push oscillator's output to further suppress it to a desired level. It is particularly noted that the two constituent oscillators need to oscillate simultaneously; otherwise, the fundamental voltages are not 180-deg out of phase and hence they do not cancel each other at the combining point.

Design Example

Figure 12.50 shows the schematic and layout of a 10.5/21-GHz VCO designed using a 0.18-μm CMOS process. The layout also includes buffers for the 21-GHz signals. This circuit consists of two VCO's: one is a 10.5-GHz complementary cross-coupled VCO and another is a 21-GHz push-push VCO. This circuit hence serves as examples for push-push as well as complementary cross-coupled oscillator designs. The cross-coupled and push-push oscillators produce output signals separately. We will describe the 10.5-GHz complementary VCO first and then the 21-GHz push-push VCO.

10.5-GHz Complementary Cross-Coupled VCO

The 10.5-GHz complementary cross-coupled VCO employs two cross-coupled NFET (M_1/M_2) and PFET (M_3/M_4) transistor pairs, and a 10.5-GHz resonator to produce two differential 10.5-GHz output signals at Port 10.5 GHz+ and 10.5 GHz−. The resonator is composed of a spiral inductor (L_1), MIM capacitors, and MOS capacitors. The MOS capacitors are voltage-controlled to enable frequency tuning. As noted in Figure 12.50(a), the transistors in each pair are turned on and off alternately according to the voltages driving the gates, which is a basic operation of the complementary cross-coupled oscillators.

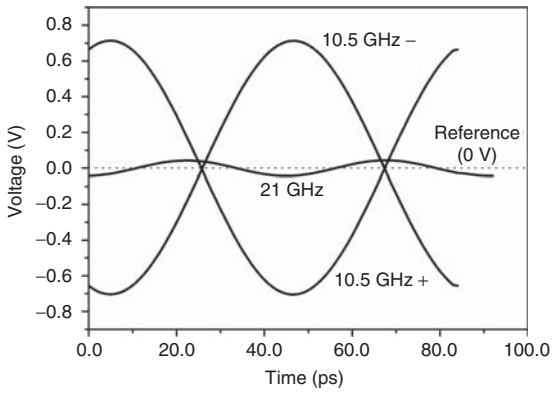


Figure 12.51. Simulated output voltage waveforms 10 GHz+ and 10.5 GHz– of the 10.5-GHz complementary VCO at ports 10.5 GHz+ and 10.5 GHz–, respectively. The 21-GHz signal (second harmonic) at each of the ports 10.5 GHz+ and 10.5 GHz– is also included.

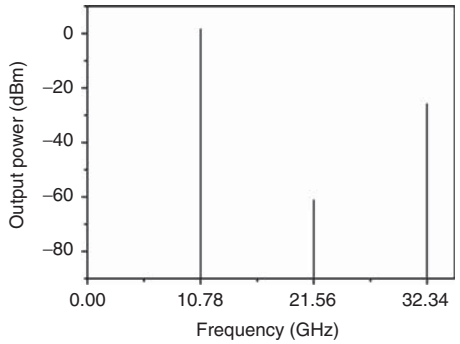


Figure 12.52. Simulated output powers of the 10.5-GHz complementary VCO.

Figure 12.51 shows the simulated output voltages (10.5 GHz+ and 10.5 GHz–) at ports 10.5 GHz+ and 10.5 GHz–, respectively. These voltages are referenced[10] to 0 V. As can be seen, the 10.5 GHz+ and 10.5 GHz– signals are equal in amplitude and 180° out of phase as expected. Figure 12.52 shows the simulated output powers at the 10.5-GHz output port. The output power at 10.5 GHz is 1.67 dBm. The second and third harmonic rejections are 59.5 and 25.8 dBc, respectively. In the power simulations, the differential ports are connected to a balun and the powers are obtained at the balun's output port. Figure 12.53 shows the calculated phase noise at 10.5 GHz (10.5 GHz+ or 10.5 GHz– signal). The phase noises of the 10.5-GHz signal is −102.05 dBc/Hz at 1-MHz offset.

21-GHz Push-Push VCO

The 10.5-GHz complementary cross-coupled VCO produces two differential 10.5-GHz signals as shown in Figure 12.51. This VCO can hence be used to form a 21-GHz push-push VCO by combining the 10.5 GHz+ and 10.5 GHz– signals in-phase (at port 21 GHz+ or 21 GHz–) as seen in Figure 12.50(a) according to the push-push oscillator theory discussed previously. Specifically, the 21-GHz push-push VCO can be viewed as consisting of two constituent 10.5-GHz oscillators: one is formed by the NFET transistor pair M_1/M_2 (VCO 1) and another by the PFET transistor pair M_3/M_4 (VCO 2) sharing the same 10.5-GHz resonator. As can be seen in Figure 12.51, the 10.5-GHz fundamental signals and the 21-GHz second-harmonic signals at ports 10.5 GHz+ and 10.5 GHz– are equal in amplitude but 180° out-of-phase and in-phase, respectively. We then

[10]It is noted that for the applied DC bias voltage V_{dd} of 1.8 V, the DC voltages at ports 10 GHz+, 10.5 GHz–, 21 GHz+ and 21 GHz– are 0.9 V, 0.9 V, 1.8 V and 0 V, respectively. Therefore, the actual waveforms have DC-offset and hence are shifted up by that amount – for instance, 0.9 – V DC-offset for the 10.5 GHz+ and 10.5 GHz– signals (shifted up 0.9 V).

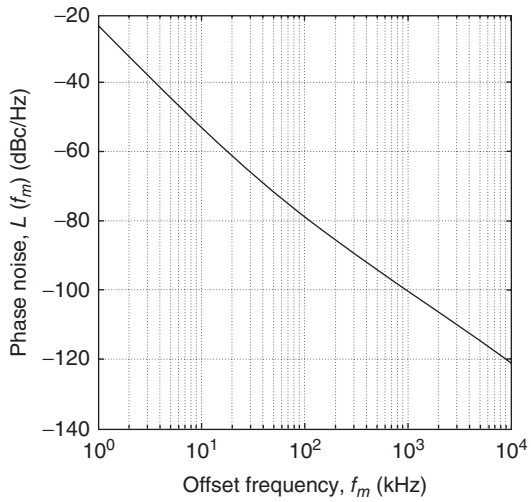


Figure 12.53. Calculated phase noise of the 10.5-GHz complementary VCO at 10.5 GHz.

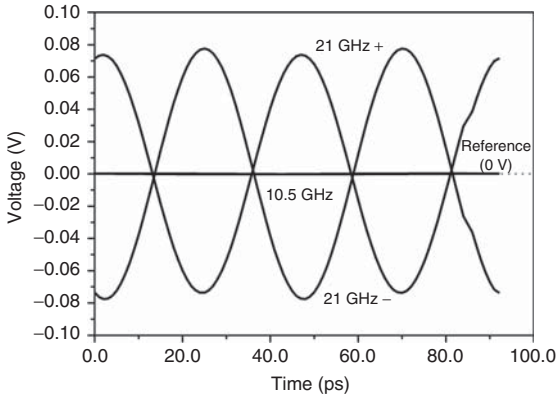


Figure 12.54. Simulated output voltage waveforms 21 GHz+ and 21 GHz− of the 21-GHz push−push VCO at ports 21 GHz+ and 21 GHz−, respectively. The 10.5-GHz signals are cancelled out.

expect that, when they are combined in-phase at a common port (21 GHz+ or 21 GHz−), the fundamental signals would be cancelled and the second-harmonic signals would be added.

Figure 12.54 shows the simulated 21-GHz output signals and 10.5-GHz signal at port 21 GHz+ (21 GHz+ signal) and port 21 GHz− (21 GHz−signal) with reference[11] to 0 V. It is noted that the 21 GHz+ signal is 180° out-of-phase with respect to the 21 GHz−signal. Figure 12.55 shows the simulated output powers at the 21-GHz output port. The output power at 21 GHz is −2.4 dBm. The rejections for the fundamental and 3rd harmonic signals are 79.37 and 58.1 dBc, respectively. In the power simulations, the differential ports are connected to a balun and the powers are obtained at the balun's output port. Figure 12.56 shows the calculated phase noise at 21 GHz (21 GHz+ or 21 GHz− signal). The phase noises of the 21-GHz signal is −99.05 dBc/Hz at 1-MHz offset, which is 3-dB worse than that of the 10.5-GHz signal.

The two additional resonators, labeled 21-GHz resonator, are used as a filter to reject the 10.5-GHz fundamental (if any) and its harmonics, except the second harmonic. These resonators include the capacitances at the sources of the cross-coupled NFET and PFET pairs. Additionally, the lower 21-GHz resonator also includes an MIM capacitor (C). This extra MIM capacitor is needed to make the two resonators electrically equal since the capacitance of the NFET pair is lower than that of the PFET pair. It is noted that the 21-GHz

[11]The actual waveforms are not symmetrical with respect to 0 V; they have DC voltage-offset due to V_{dd}.

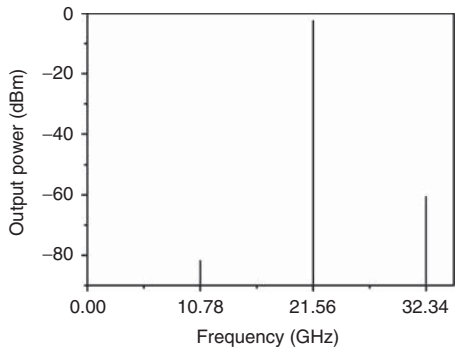


Figure 12.55. Simulated output powers of the 21-GHz push–push VCO.

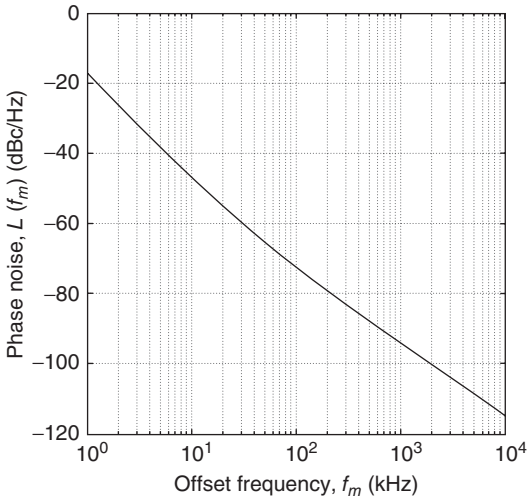


Figure 12.56. Calculated phase noise of the 21-GHz push–push VCO at 21 GHz.

signals are generated from the 10.5 GHz signals as a characteristic of push-push oscillators. These 21-GHz signals are not produced by oscillation; that is, there is no negative resistance generated at 21 GHz.

REFERENCES

1. K. Kurokawa, "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," *Bell Syst. Tech. J.*, pp. 1937–1955, July–August 1969.
2. D. M. Pozar, *Microwave and RF Design of Wireless Systems*, John Wiley & Sons, Inc., New York, NY, 2001.
3. U. L. Rhode, *Digital PLL Frequency Synthesizers, Theory and Design*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1983.
4. D. Leeson, "A Simple Model of Feedback Oscillator Noise Spectrum," *IEEE Proceedings*, Vol. 54, pp. 329–330, Feb. 1966.
5. D. Ham and A. Hajimiri, "Concepts and Methods in Optimization of Integrated LC VCOs," *IEEE J. Solid-State Circuits*, Vol. 36, pp. 896–909, June 2001.
6. Cadence, *Cadence Design Systems, Inc.*, San Jose, CA, USA.
7. A. Hajimiri and T. H. Lee, "Design Issues in CMOS Differential LC Oscillators," *IEEE J. Solid-State Circuits*, Vol. 34, pp. 717–724, May 1999.
8. S. Gierkink et al., "Intrinsic $1/f$ Device Noise Reduction and Its Effect on Phase Noise in CMOS Ring Oscillators," *IEEE J. Solid-State Circuits*, Vol. 34, pp. 1022–1025, July 1999.

9. TSMC 0.25-μm CMOS Process, MOSIS, Marina del Rev, California, USA.
10. Z. Skvor, S. R. Saunders, and C. S. Aitchison, "Novel Electronically Tunable Microwave Oscillator Based on the Distributed Amplifier," *Electron. Lett.*, Vol. 28, No. 17, pp. 1647–1648, Aug. 1992.
11. B. Kleveland et al., "Monolithic CMOS Distributed Amplifier and Oscillator," *Proc. of IEEE ISSCC Dig. Tech. Papers*, pp. 70–71, Feb. 1999.
12. H. Wu and A. Hajimiri, "Silicon-Based Distributed Voltage-Controlled Oscillators," *IEEE J. Solid State Circuits,* Vol. 36, pp. 493–502, Mar 2001.
13. IE3D Program, *Zeland Software, Inc.*, Fremont, CA, USA.
14. Advanced Design System (ADS), *Agilent Technologies, Inc.,* Santa Clara, CA, USA.
15. CA18HR 0.18-μm CMOS Process, *Jazz Semiconductor*, Newport Beach, CA, USA.
16. H. Xiao, T. Tanaka, and M. Aikawa, "Push-Push Oscillator with Simplified Circuit Structure," *Electron. Lett.*, Vol. 38, No. 24, pp. 1545–1547, 21st Nov. 2002.

PROBLEMS

12.1 Consider a basic oscillator as shown in Figure 12.5(a) along with its equivalent circuits shown in Figure 12.5(b) and (c). Assuming the equivalent circuit shown in Figure 12.5(c) is oscillated, prove that the circuit shown in Figure 12.5(a) or (b) is also oscillated as well.

12.2 Consider a MOSFET having a large-signal S-parameters at 10 GHz of $S_{11} = 0.9085\angle -51.23^{o}$ $S_{12} = 0.1193\angle 53.69^{o}$ $S_{21} = 2.6812\angle 138.06^{o}$ $S_{22} = 0.6428\angle -42.06^{o}$
Design and describe in details a single-ended oscillator at 10 GHz using this transistor and lumped elements in a 50-Ω system, including the stability circle and the designed oscillator's schematic. Any circuit topology including that shown in Figure 12.5(a) may be used for this oscillator.

12.3 Repeat Problem 12.2 using microstrip lines. The microstrip lines are the same as that given in Problem 4.25 of Chapter 4 with SiO_2 thickness of 8 μm.

12.4 Consider a 0.18-μm NMOS FET whose small-signal S-parameters from 0.1-20 GHz are given in Problem 9.1 of Chapter 9. Choose a frequency at which the device is potentially unstable and design an oscillator operating at that frequency using lumped elements. Describe your design in details and include the stability circuit and the circuit schematic.

12.5 Design a 10-GHz oscillator using lumped elements and any available 0.18-μm CMOS process. There are no required specifications for this oscillator. However, you are required to try your best to achieve best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, and power consumption.

12.6 Draw a common-drain Colpitts oscillator with a three-branch feedback network between the gate and source, and derive the relations between the feedback elements and the MOSFET's parameters, assuming a simple unilateral device model as shown in Figure 12.11.

12.7 Consider a parallel RLC resonator having a resonant frequency ω_o. Derive its transfer function $H(j\omega)$ for a narrow bandwidth around ω_o specified by $\omega_o \pm \Delta\omega$ in terms of ω_o, $\Delta\omega$ and the resonator quality factor Q.

12.8 Repeat Problem 12.7 for a series RLC resonator.

12.9 Consider signal given as $v(t) = V_s \cos[2\pi f_o + \phi(t)]$ where $\phi(t)$ represents the phase fluctuation or noise. Assume that the signal's frequency is modulated sinusoidally around f_o as $f = f_o + \Delta f \cos(2\pi f_m)$. This frequency modulation shows a variation of the signal $v(t)$ in terms of the frequency fluctuation or noise. Derive this $v(t)$.

12.10 Consider Figure 12.25. Prove that the power transfer function $|V_o/V_i|^2$ given in (12.88) represents the ratio between the PSD of the output noise $S_\phi(\omega)$ and the PSD of the input noise $S_{\phi i}(\omega)$.

12.11 Consider a 24-GHz oscillator with $P_{\text{avs}} = 5$ dBm, $F = 5$ dB and $f_c = 200$ KHz. Compute the phase noise at 100 KHz and 1 MHz offset frequency for Q_L of 10, 20, and 50. Comment on the results with respect to f_c, $f_o/2Q_L$ and Q_L.

12.12 Consider a 24-GHz feedback CMOS oscillator employing an amplifier having $F = 4$ dB, $P_{\text{avs}} = 1$ dBm using a MOSFET with $f_c = 250$ KHz. Calculated and plot the phase noise (in dBc/Hz) using Eq. (12.98) as a function of the offset frequency f_m (in Hz) from 1 KHz to 10 MHz for resonator Q_L from 5 to 30 in steps of 5. Comment on the results.

12.13 Consider a 35-GHz feedback CMOS oscillator employing an amplifier having $F = 5$ dB, $P_{\text{avs}} = 0$ dBm and a feedback resonator having $Q_L = 8$. Calculated and plot the phase noise (in dBc/Hz) using Eq. (12.98) as a function of the offset frequency f_m (in Hz) from 1 KHz to 10 MHz for flicker frequency $f_c = 1,\ 5,\ 25,\ 125,\ 625$ KHz. Comment on the results.

12.14 Consider a 35-GHz feedback CMOS oscillator employing an amplifier having $P_{\text{avs}} = 0$ dBm using a MOSFET with $f_c = 100$ KHz and a feedback resonator having $Q_L = 8$. Calculated and plot the phase noise (in dBc/Hz) using Eq. (12.98) as a function of offset frequency f_m (in Hz) from 1 KHz to 10 MHz for noise figure $F = 3,\ 5,\ 7,\ 10$, and 12 dB. Comment on the results.

12.15 Consider a 35-GHz feedback CMOS oscillator employing an amplifier having $F = 5$ dB using a MOSFET with $f_c = 100$ KHz and a feedback resonator having $Q_L = 8$. Calculated and plot the phase noise (in dBc/Hz) using Eq. (12.98) as a function of offset frequency f_m (in Hz) from 1 KHz to 10 MHz for the amplifier's input power $P_{\text{avs}} = -10,\ -5,\ 0,\ 5,\ 10$ and 15 dBm. Comment on the results.

12.16 The phase noise of feedback oscillators is given in Eq. (12.98) as

$$\mathscr{L}(f_m) = \frac{FkT}{2P_{\text{avs}}}\left[f_c\left(\frac{f_o}{2Q_L}\right)^2\left(\frac{1}{f_m^3}\right) + \left(\frac{f_o}{2Q_L}\right)^2\left(\frac{1}{f_m^2}\right) + f_c\left(\frac{1}{f_m}\right) + 1\right]$$

The unloaded Q_U and loaded Q_L of the resonator used in the feedback loop can be determined based on Chapter 5 as

$$Q_U = \frac{\omega_o W}{P_{L,\text{res}}}$$

$$Q_U = \frac{\omega_o W}{P_{L,\text{res}}}$$

$$Q_L = \frac{\omega_o W}{P_L} = \frac{\omega_o W}{P_{\text{avs}} + P_{L,\text{res}} - P_o}$$

where W is the stored energy, P_L is the total power loss including coupling loss, $P_{L,\text{res}}$ is the power loss in the resonator, P_{avs} is the input power entering the amplifier, and P_o is the oscillator's output power (i.e., the power taken out from the feedback network.) Derive the following expression for $\mathscr{L}(f_m)$:

$$\mathscr{L}(f_m) = \frac{FkT}{2P_{\text{avs}}}\left[1 + \frac{f_o^2}{4f_m^2}\left(\frac{P_{\text{avs}}}{2\pi f_o W} + \frac{1}{Q_U} + \frac{P_o}{2\pi f_o W}\right)^2\right]\left(1 + \frac{f_c}{f_m}\right)$$

This equation describes in detail the influence of several design parameters on the oscillator's phase noise. Based on this equation, provide your design guidelines to achieve minimum phase noise.

12.17 Although phase noise can be measured directly by a spectrum analyzer, there are two possible problems limiting such measurement. One is the limited dynamic range of a spectrum analyzer preventing it from measurement of very low phase noise below its own phase noise. The other problem is an analyzer measures both amplitude- and phase-modulation of a signal simultaneously. Assume a spectrum analyzer can be used for measuring an oscillator's phase noise. Describe a measurement procedure and show an example to demonstrate it – for instance, an oscillator's output signal spectrum or any other data that allow you to demonstrate your method of determining the phase noise.

12.18 Doppler radars are used to measure speeds of targets by measuring the (Doppler) frequency shift of the targets' return signals. In operations, the return signals can include those from desired targets as well as those from undesired objects such as (stationary) buildings or trees. Describe with rationale possible effects of the phase noise of the receiver's LO on the operation of a Doppler radar.

12.19 We model a sinusoidal signal corrupted by phase noise as a linear frequency-modulation process as

$$V(t) = A\cos[2\pi f_o t + m(f_m)\sin(2\pi f_m t)]$$

where A and f_o are the carrier's amplitude and frequency, respectively, f_m is the modulating (or offset) frequency, and $m(f_m)$ is the modulation index defined by

$$m(f_m) = \frac{\delta f}{f_m}$$

with δf being the peak frequency deviation at the offset frequency f_m due to the frequency instability of the signal source. Derive the following expression for the single-sideband phase noise:

$$\mathscr{L}(f_m) \equiv \frac{P_{\text{noise}}(f_m)}{P_{\text{signal}}} = \frac{m_{\text{rms}}^2(f_m)}{2}$$

where P_{signal} is the carrier's power and m_{rms} is the root mean square of the modulation index.

12.20 The (statistical average) variances of the phase noise and frequency noise at offset frequency f_m over a bandwidth for f_m from f_L and f_H can be expressed respectively as

$$\overline{\sigma_\phi{}^2} = 2\int_{f_L}^{f_H} L(f_m)df_m \quad rad^2(dB)$$

$$\overline{\sigma_f{}^2} = 2\int_{f_L}^{f_H} L(f_m)f_m^2 df_m \quad Hz^2(dB)$$

If we assume, as a special case, the phase noise is white over the considered bandwidth, with the frequency noise spectrum having a rising slope of 20 dB/decade with increasing frequency, derive the following equations:

$$\overline{\sigma_\phi{}^2} = 2[L(f_m)(f_H - f_L)] \quad rad^2(dB)$$

$$\overline{\sigma_f{}^2} = 2\left[L(f_m)\left(\frac{f_H^3 - f_L^3}{3}\right)\right] \quad Hz^2(dB)$$

These equations enable the determination of the rms phase and frequency error originated from the instability of RF sources, respectively, assuming the phase noise is white noise across the bandwidth.

12.21 Derive Equation (12.118).

12.22 Derive Equation (12.120).

12.23 Design a 10-GHz lumped-element cross-coupled oscillator using only NMOSFETs in any available 0.18-μm CMOS process. There are no required specifications for this oscillator. However, you are required to try your best to achieve best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, and power consumption.

12.24 Repeat Problem 12.23 using only PMOS transistors.

12.25 Derive Equation (12.124).

12.26 Derive Equation (12.133).

12.27 Design a 4-stage distributed oscillator operating at 10 GHz using any available 0.18-μm CMOS process. Either on-chip spiral inductors or transmission lines can be used for the synthetic transmission lines. There are no required specifications for this oscillator. However, you are required to try your best to achieve the best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, and power consumption.

12.28 Design a 20-GHz lumped-element push-push oscillator using any available 0.18-μm CMOS process using a conventional architecture as shown in Figure 12.48. There are no required specifications for this oscillator. However, you are required to try to the maximum level to achieve the best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, spectrum and power consumption. Also include in your design the following and discussion of the results: simulated voltage waveforms of the fundamental and second harmonic signals at the outputs of the constituent oscillators and push-push oscillator.

12.29 Design a 20-GHz push-push oscillator using any available 0.18-μm CMOS process based on Reference [16] with a half-wavelength open-circuited CPW or a microstrip resonator at 10 GHz. There are no required specifications for this oscillator. However, you are required to try your best to achieve the best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, spectrum, and power consumption. Also include in your design the following and discussion of the results: simulated voltage waveforms of the fundamental and second harmonic signals at the outputs of the constituent oscillators and push-push oscillator. If you also designed the oscillator described in Problem 12.28, then compare the performance and describe the advantages and disadvantages of the oscillators in Problems 12.24 and 12.25.

12.30 Consider the oscillator topology shown in Figure 12.50, which can generate both fundamental and second-harmonic differential signals simultaneously. Design a lumped-element oscillator based on that topology using any available 0.18-μm CMOS process to produce 7.5-GHz and 15-GHz differential signals. There are no required specifications for this oscillator. However, you are required to try your best to achieve the best possible output power, phase noise, match, and power consumption. Draw the schematic with all element values. Prepare the layout of the oscillator ready for tape-out for fabrication. Perform post-layout simulations. Plot and discuss the performance including output power, phase noise, return loss, and power consumption corresponding to 7.5- and 15-GHz signals, and spectrums. Also include in your design the following and discussion of the results: simulated voltage

waveforms of the fundamental 7.5-GHz and second-harmonic 15-GHz signals at the 7.5-GHz and 15-GHz differential ports.

12.31 Extend the design in Problem 12.30 to include a (passive or active) balun at the 7.5-GHz ports to combine the output signals and another balun at the 15-GHz ports to convert differential ports into a single-ended port. Describe the design and performance of these baluns. Simulated and plot the voltage waveforms of the 7.5-GHz signal and its second harmonic at the 7.5-GHz single-ended port and the 15-GHz signal at the 15-GHz single-ended port. Discuss the results.

12.32 One of the requirements for the constituent oscillators of push-push oscillators is that they are to be designed to produce oscillation only at fundamental frequency, not at both fundamental and second-harmonic frequencies. Provide your rationale of why only fundamental-frequency oscillation is wanted.

12.33 Consider the push-push oscillator shown in Figure 12.48(a), either one of the following criteria would result in no second-harmonic oscillation:

a) $|\Gamma_{\text{osc}}(2\omega_o)| < 1$ where $\Gamma_{\text{osc}}(2\omega_o)$ is the reflection coefficient corresponding to $Z_{\text{osc}}(2\omega_o)$, or

b) If $R_{\text{osc}}(2\omega_o) < 0$ then $R_{\text{out}}(2\omega_o)$ must be greater than $|R_{\text{osc}}(2\omega_o)|$, or

c) If $R_{\text{osc}}(2\omega_o) < 0$ and $R_{\text{out}}(2\omega_o) < |R_{\text{osc}}(2\omega_o)|$, then $X_{\text{out}}(2\omega_o)$ must be different from $-X_{\text{osc}}(2\omega_o)$. From a fundamental oscillator design principle of out-of-band stability, criterion (a) is the most desirable condition. Conditions (b) and (c) are acceptable but not desirable. Provide your rationale to support this statement. If you feel otherwise, then provide your reasoning accordingly.

12.34 Consider a VCO block diagram as shown in Figure P12.1, where K_1, K_2 are the coupling coefficients of the corresponding coupling networks and Q_V, Q_{TL}, Q_R are the varactor unloaded Q, transmission line unloaded Q, resonator loaded Q, respectively. Derive the following expression for Q_V:

$$Q_V = \frac{(1+K_1)(1+K_2)Q_R Q_{TL}}{Q_{TL} - (1+K_2)Q_R}$$

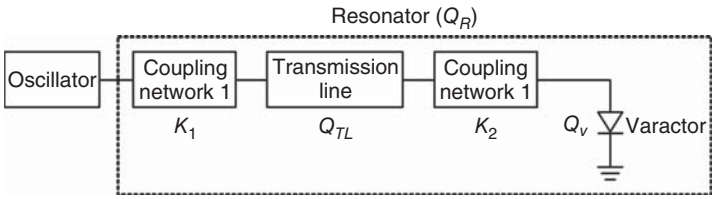


Figure P12.1.

12.35 Consider the VCO described in Problem 12.34. Assume the VCO operates at 35 GHz and requires a minimum loaded Q of 5 for the resonator, the (power) coupling coefficients K_1 and K_2 are 0 and −10 dB, respectively, and the transmission line unloaded Q is 20. Determine the required minimum unloaded Q for the varactor at 35 GHz.

12.36 Consider a resonant structure for a VCO consisting of a microstrip line and varactor as shown in Figure P12.2. Derive an expression for the total unloaded Q of the resonator in terms of the unloaded Q of the microstrip line and varactor.

12.37 Derive the following equation for stable oscillation given in Eq. (12.14):

$$\left.\frac{\partial R_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dX_{\text{out}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} - \left.\frac{\partial X_{\text{osc}}(V,\omega)}{\partial V}\right|_{V=V_o} \left.\frac{dR_{\text{out}}(\omega)}{\partial \omega}\right|_{\omega=\omega_o} > 0$$

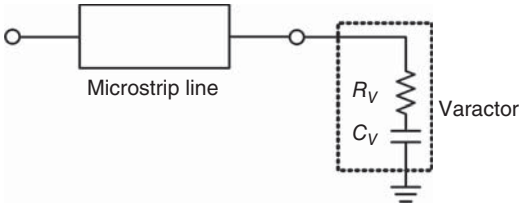


Figure P12.2.

12.38 From Fourier series, a pulse signal can be expressed as consisting of multiple continuous-wave sinusoidal signals, which sets a basic design approach for pulse generators. Assume we want to produce a rectangular pulse or impulse signal with a 3-dB pulse width of 100 ps by combining the sinusoidal signal outputs of multiple oscillators. Determine the frequencies, amplitudes, and phases of the sinusoidal signals you need to generate such a pulse. You can use just enough frequencies (not too many or too little). Use a commercially available circuit analysis program to combine these oscillators to simulate the resultant pulse. Comment on the results.

12.39 Repeat Problem 12.34 for a 200-ps monocycle pulse.

12.40 Figure of Merit (FOM) of VCOs is commonly used as a means to evaluate the overall performance of VCOs. The higher the value of FOM, the better the VCO design. One equation characterizing FOM is

$$\text{FOM} = -\mathscr{L}_{\text{dB}}\{f_m\} + 10\log\left[\frac{1}{P_{\text{dc}}}\left(\frac{f0}{f_m}\right)^2\right]$$

where f_o is the oscillation frequency, f_m is the offset frequency, $\mathscr{L}_{\text{dB}}\{f_m\}$ is the phase noise at f_m in dB, and P_{dc} is the DC power consumption in megawatt. The FOM as given in the equation normalizes the phase noise to the oscillation frequency and offset frequency and indicates how efficient the consumed power is used to achieve a given phase noise performance. This FOM does not completely characterize VCOs. Describe any drawbacks of this FOM and, if possible, modify it to overcome the shortcomings that you describe. Search the literature to find out the minimum and maximum values for FOM of recently developed fully integrated CMOS RF VCOs.

CHAPTER 13

MIXERS

Mixers are among the most important components in RF systems. They are typically used as the second component after a low-noise amplifier (LNA) in receivers or before a power amplifier (PA) in transmitters. Their performance is crucial to the operation of RF systems. Mixers can perform two functions: down-conversion and up-conversion. Down-conversion mixers, or down-converters, are the heart of RF receivers that convert an RF input signal, through mixing with a local oscillator (LO) signal in nonlinear mixing elements, to a signal at the difference or intermediate frequency (IF). Up-conversion mixers, or up-converters, on the other hand, convert a signal into a higher-frequency signal through mixing with an LO signal and find applications in transmitters. The main function of mixers is basically converting an input signal (RF) being modulated by another signal (LO) into another signal of different frequency (IF), while keeping the amplitude and phase linearly scaled with respect to those of the original RF signal. Mixers thus work essentially as a linear component but with nonlinear solid-state devices. Linear function is in fact desired for mixers. In contrast with other frequency-translation components such as frequency multipliers, mixers preserve the amplitude and phase of incoming signals and hence are used as the main component to detect and process received signals. There are a number of different topologies that can be used for RF mixers, ranging from those employing only a single transistor to those having four transistors or more, serving different subsystem or system requirements. Common requirements for mixers are high gain and low noise figure. Depending on applications and operating environments, however, other specifications such as linearity may become more critical. This chapter discusses the fundamentals of mixers, their topologies, analysis, and design for RFICs.

13.1 FUNDAMENTALS OF MIXERS

13.1.1 Mixing Principle

As we know from Fourier transform theory, the Fourier transform of the product $x(t)e^{j2\pi f_o t}$ between a signal $x(t)$, whose Fourier transform is $X(f)$, and a complex function $e^{j2\pi f_o t}$ is $X(f - f_o)$. The resultant spectrum of the product indicates that the frequency f of a signal can be transformed into another frequency $(f - f_o)$ by multiplying the signal with a pure complex (exponential) signal of frequency f_o. Theoretically, this mathematical

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

model can be realized electrically using a (signal) multiplier device whose inputs are $x(t)$ and $e^{j2\pi f_o t}$. In practice, however, the signals we are dealing with, such as RF signals received by antennas, are real-valued signals such as sinusoidal signals, not complex signals. As a result, a single-frequency translation, which transfers frequency f into another frequency $(f - f_o)$, does not happen; instead, f is translated into multiple frequencies as will be seen later. Nevertheless, the overall principle of "frequency translation" by "signal multiplication" from Fourier transform theory remains the same. To illustrate this operation, we consider two real signals with zero-phase (chosen for simplicity) represented by their voltages $v_{\mathrm{RF}}(t) = A_{\mathrm{RF}} \cos(2\pi f_{\mathrm{RF}} t)$ and $v_{\mathrm{LO}}(t) = A_{\mathrm{LO}} \cos(2\pi f_{\mathrm{LO}} t)$ with maximum amplitudes A_{RF} and A_{LO}, respectively, and obtain their product

$$\begin{aligned} v_{\mathrm{IF}}(t) &= v_{\mathrm{RF}}(t) \times v_{\mathrm{LO}}(t) = A_{\mathrm{RF}} A_{\mathrm{LO}} \cos(2\pi f_{\mathrm{RF}} t) \cos(2\pi f_{\mathrm{LO}} t) \\ &= \frac{A_{\mathrm{RF}} A_{\mathrm{LO}}}{2} \{\cos[2\pi(f_{\mathrm{RF}} - f_{\mathrm{RF}})t] + \cos[2\pi(f_{\mathrm{RF}} + f_{\mathrm{RF}})t]\} \end{aligned} \tag{13.1}$$

Examination of (13.1) shows that two frequency translations, one to $(f - f_o)$ and another to $(f + f_o)$, are obtained, which set the basis for mixers. From these very fundamental mathematical concept and signal-multiplying device, we can see that a mixer can be formed using an active device that is capable of multiplication and operation with two input signals to produce signals of different frequencies. Figure 13.1 shows a general mixer block, which contains three ports: RF, LO, and IF. The RF and LO ports are used to feed signals at RF (f_{RF}) and LO (f_{LO}) frequencies, respectively, to the active devices, herein assumed to be MOSFETs, in the mixer, and the IF output port is used to extract the output signal at IF frequency. The mixing in MOSFET mixers occurs due to the nonlinearity in the MOSFETs, whose dominant nonlinear elements are the gate–source capacitance (C_{gs}), gate–drain capacitance (C_{dg}), transconductance (g_m), and (output) drain–source conductance (g_{ds}).

The mixing concept as described above considers ideal mixing and is relatively simple. Practical mixers, however, are not ideal and do not produce only two IF signals at $f_{\mathrm{RF}} \pm f_{\mathrm{LO}}$, but additional signals at other frequencies. These resulting output signals are due to the nonlinearity of the employed MOSFETs. Typical mixers employ MOSFETs operating as nonlinear devices and, hence, are in general considered nonlinear component working essentially as a device to perform a multiplication of signals. We assume that the MOSFET has a nonlinear response where the resultant current $i(t)$ can be expressed as an infinite power series of the driving voltage $v(t)$ as

$$i(t) = \sum_{n=0}^{\infty} a_n v^n(t) \tag{13.2}$$

where a_n are amplitude coefficients. We consider the same RF and LO signal as described before and express the (total) input signal voltage as

$$v(t) = v_{\mathrm{RF}}(t) + v_{\mathrm{LO}}(t) = A_{\mathrm{RF}} \cos(2\pi f_{\mathrm{RF}} t) + A_{\mathrm{LO}} \cos(2\pi f_{\mathrm{LO}} t) \tag{13.3}$$

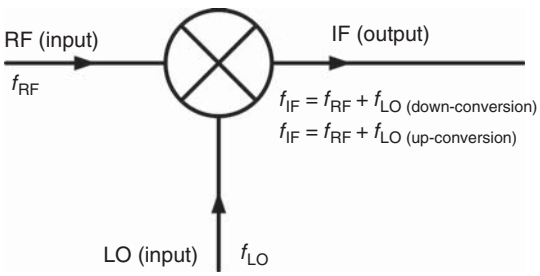


Figure 13.1. A general mixer block.

Substituting (13.3) into (13.2) gives

$$\begin{aligned} i(t) =& a_o + a_1[A_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + A_{\mathrm{LO}}\cos(2\pi f_{\mathrm{LO}}t)]+ \\ & a_2[A_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + A_{\mathrm{LO}}\cos(2\pi f_{\mathrm{LO}}t)]^2+ \\ & a_3[A_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + A_{\mathrm{LO}}\cos(2\pi f_{\mathrm{LO}}t)]^3 + \cdots \end{aligned} \tag{13.4}$$

which becomes, after applying trigonometric identities,

$$\begin{aligned} i(t) =& \left[a_o + \frac{1}{2}a_2\left(A_{\mathrm{RF}}^2 + A_{\mathrm{LO}}^2\right)\right] + \\ & a_1[A_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + A_{\mathrm{LO}}\cos(2\pi f_{\mathrm{LO}}t)]+ \\ & \frac{1}{2}a_2[A_{\mathrm{RF}}^2\cos(4\pi f_{\mathrm{RF}}t) + A_{\mathrm{LO}}^2\cos(4\pi f_{\mathrm{LO}}t)]+ \\ & A_{\mathrm{RF}}A_{\mathrm{LO}}\{\cos[2\pi(f_{\mathrm{RF}} - f_{\mathrm{LO}})t] + \cos[2\pi(f_{\mathrm{RF}} + f_{\mathrm{LO}})t]\}+ \\ & a_3\left\{\frac{A_{\mathrm{RF}}^3}{4}\left[\cos(6\pi f_{\mathrm{RF}}t) + \cos(2\pi f_{\mathrm{RF}}t)\right] + \frac{A_{\mathrm{LO}}^3}{4}[\cos(6\pi f_{\mathrm{LO}}t) + \cos(2\pi f_{\mathrm{LO}}t)]\right\} + \\ & 3a_3A_{\mathrm{LO}}A_{\mathrm{RF}}^2\left\{\frac{1}{2}\cos\left(2\pi f_{\mathrm{LO}}t\right) + \frac{1}{4}[\cos[(4\pi f_{\mathrm{RF}} - 2\pi f_{\mathrm{LO}})t] + \cos[(4\pi f_{\mathrm{RF}} + 2\pi f_{\mathrm{LO}})t]]\right\} + \\ & 3a_3A_{\mathrm{RF}}A_{\mathrm{LO}}^2\left\{\frac{1}{2}\cos(2\pi f_{\mathrm{RF}}t) + \frac{1}{4}[\cos[(4\pi f_{\mathrm{LO}} - 2\pi f_{\mathrm{RF}})t] + \cos[(4\pi f_{\mathrm{LO}} + 2\pi f_{\mathrm{RF}})t]]\right\} + \cdots \end{aligned} \tag{13.5}$$

Equation (13.5) shows that the mixer would produce, from the input RF signal at frequency f_{RF} and LO signal at frequency f_{LO}, signals at DC and frequencies of $f_{\mathrm{RF}} - f_{\mathrm{LO}}$,[1] $f_{\mathrm{RF}} + f_{\mathrm{LO}}$, f_{RF}, f_{LO}, $2f_{\mathrm{RF}}$, $2f_{\mathrm{LO}}$, $3f_{\mathrm{RF}}$, $3f_{\mathrm{LO}}$, $2f_{\mathrm{RF}} - f_{\mathrm{LO}}$, $2f_{\mathrm{RF}} + f_{\mathrm{LO}}$, $2f_{\mathrm{LO}} - f_{\mathrm{RF}}$, $2f_{\mathrm{LO}} + f_{\mathrm{RF}}$, and so on. These frequencies can be expressed generally as

$$f_o = mf_{\mathrm{RF}} + nf_{\mathrm{LO}} \tag{13.6}$$

where m and n can be any positive and negative integers including zero ($\pm 0, \pm 1, \pm 2, \ldots$). Of these mixing products, the ones at the IF frequency, $f_{\mathrm{RF}} - f_{\mathrm{LO}}$, or $f_{\mathrm{LO}} - f_{\mathrm{RF}}$ when $f_{\mathrm{RF}} < f_{\mathrm{LO}}$, and $f_{\mathrm{RF}} + f_{\mathrm{LO}}$ are often most interested for down-conversion and up-conversion mixers, respectively. The other products are (unwanted) spurious signals, typically referred to as "spurs," occurring due to the harmonic generation of the input signals and the mixing of these harmonics in the mixer. Since the desired function of mixers is to convert an incoming RF signal to an IF signal as much as possible, the existence of these undesired signals reduces the amount of an RF signal that can be converted into an IF signal. Ideally, the generation of these spurious signals needs to be as small as possible and those generated should be suppressed as much as possible to minimize the degradation of the IF signal. Since these mixing products are generated within the MOSFET, they can travel toward any port of the mixer, and so proper filtering needs to be included to suppress these undesired signals. For instance, the IF circuit typically includes a low-pass filter (LPF) to reject all mixing products except the desired IF signal. Spurs happen at various frequencies and, while those falling out of the IF band can be easily filtered out, those occurring next or within the IF band, such as the image-converted IF signal, are troublesome, as it is extremely difficult, if not impossible, to remove these spurs. The levels of spurs are important in mixer design. There are several spurious products that, if not properly suppressed, can degrade the performance of the mixer. The most significant of these is the signal at $2f_{\mathrm{LO}} - f_{\mathrm{RF}}$ or $f_{\mathrm{LO}} - f_{\mathrm{IF}}$ symmetrically opposite to the RF signal with respect to the LO signal. This signal is known as the image signal and can be converted to a signal having the same frequency of the desired IF signal through a mixing with the

[1]This happens when $f_{\mathrm{RF}} > f_{\mathrm{LO}}$. For $f_{\mathrm{RF}} < f_{\mathrm{LO}}$, the frequency is $f_{\mathrm{LO}} - f_{\mathrm{RF}}$.

LO signal. It is noted that, not only that the image signal is not the desired RF signal (and hence the resulting IF signal) to be detected, but also that the inherent noise of the image signal, and hence of the undesired down-converted IF signal, may severely degrade the mixer performance. The image signal can also come from other sources, externally or internally such as that coming from an adjacent channel in multichannel receivers. The image-converted IF signal is undesirable since it comes from a signal of no interest and would contaminate the desired IF signal, hence degrading the system performance. The image signal thus needs to be suppressed, most often using an image-reject filter or image-reject mixer. It is noted again that spurs are unwanted signals and need to be suppressed as much as possible. Conventionally, filters can be used to achieve this, which, however, may not be possible for certain mixing frequencies. Properly configured mixer topologies may help in reducing the spur levels and overcoming possible filtering problems. Figure 13.2 shows the several components of the output spectrum of mixers, assuming $f_{\mathrm{RF}} > f_{\mathrm{LO}}$.

13.1.2 Mixer Parameters

Compared to other RFICs, mixers are characterized by more parameters. These parameters are described in this section.

13.1.2.1 Conversion Gain. As for other RFICs, such as amplifiers, conversion gain is defined as the ratio between the IF output signal level (voltage or power) and the RF input signal level (voltage and power). Conversion gain measures the strength of an output signal with respect to an input signal, essentially characterizing the mixer's efficiency in signal conversion. Conversion gain is typically achieved in active MOSFET mixers as opposed to conversion loss obtained for passive MOSFET (and diode) mixers. Conversion gain for mixers is needed not only for achieving efficient signal conversion, but also for overcoming possible high noise of components following the mixers. In general, both conversion gain and low noise figure for mixers are desired. This is especially important in the millimeter-wave regime, where high gain may not be possible or difficult to obtain for an LNA preceding the mixer.

13.1.2.2 Noise Figure. Noise figure of a mixer characterizes the "noise" of the mixer by itself regardless of any noise that can possibly enter it. The concept of noise figure for mixers is essentially the same as that for amplifiers as discussed in Section 11.2.1. The noise figure of mixer is defined as the ratio between the signal-to-noise (S/N) ratio at the RF input and the S/N ratio at the IF output using the same equation (12.92) or (12.90) for amplifiers as

$$F = \frac{P_{\mathrm{Si}}/P_{\mathrm{Ni}}}{P_{\mathrm{So}}/P_{\mathrm{No}}} = \frac{\mathrm{S/N\ \ at\ \ RF\ \ input}}{\mathrm{S/N\ \ at\ \ IF\ \ output}} \tag{13.7}$$

or

$$F = \frac{P_{\mathrm{No}}}{P_{\mathrm{Ni}}G} \tag{13.8}$$

respectively, where P_{Si} and P_{So} are the available signal power at the input and output of the mixer, respectively, P_{Ni} and P_{No} are the input and output noise power of the mixer, respectively, and G is the conversion

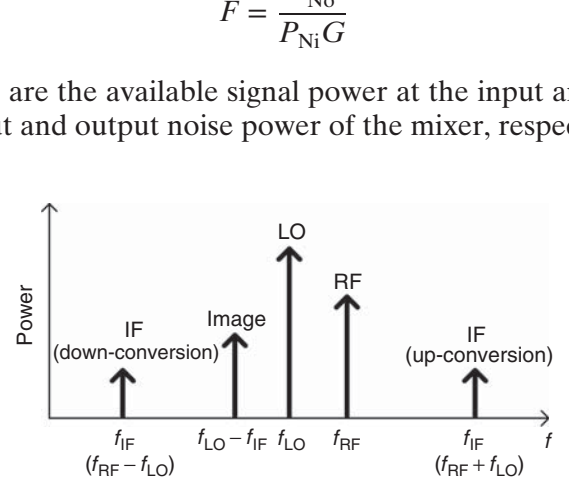


Figure 13.2. Mixer's output spectrum with $f_{\mathrm{RF}} > f_{LO}$.

gain of the mixer. For passive MOSFET mixers, the noise figure is approximately equal to the conversion loss. As a mixer in a receiver front-end is typically used after an LNA, its noise figure does not play a significant role in the receiver's noise performance as that of the LNA, provided that the LNA's gain is sufficiently high according to (11.108). However, since a mixer is typically the second component of a receiver front-end, it is still desirable to keep the mixer's noise figure as low as possible in order to maintain an overall low noise figure for the receiver, particularly when the LNA does not have sufficient gain. For other mixers used after the main receiver mixer, their noise figures are generally not a concern. There is one important note, however, for front-end mixers operating in the millimeter-wave range: it may be difficult to achieve high gain for millimeter-wave LNA's and, under this situation, low noise figure for these mixers is critical for achieving an overall low noise figure for receivers. For mixers with sufficiently high frequency for the IF signal, the $1/f$ or flicker noise does not contribute to the overall mixer's noise figure. However, for direct-conversion mixers whose IF signal is in the base-band, the $1/f$ noise affects the noise performance and hence degrading the mixer performance.

Due to the possible existence of the image signal located at $2f_{LO} - f_{RF}$ or $f_{LO} - f_{IF}$ (for $f_{RF} > f_{LO}$), there exist two possible noise figures: single-sideband (SSB) and double-sideband (DSB) noise figures. For SSB noise figure, we assume that no image signal is present. However, since white noise inherently exists every where, the white noise at the image frequency can convert into noise at IF just like the white noise at RF. As a result, while there is only a single down-converted IF signal, there are two equal down-converted noise signals at IF. Accordingly, the noise figure is obtained from (13.7) as

$$\begin{aligned} F_{SSB} &= \frac{P_{Si}/P_{Ni}}{P_{So}/P_{No}} = \frac{P_{Si}/P_{Ni}}{P_{Si}G/(2P_{Ni}G + P_{Nm})} \\ &= 2 + \frac{P_{Nm}}{P_{Ni}G} \end{aligned} \tag{13.9}$$

where P_{Nm} is the (intrinsic) noise power produced by the mixer itself. For DSB noise figure, we consider an image signal occurring concurrently with the desired RF signal. One example is image-recovery mixers in which the image signal is used in conjunction with the RF signal. Consequently, there are two equal real signals and two equal noise signals at RF and image frequency. These signals are all down-converted into signals at IF and, hence, the noise figure can be determined from (13.7) accordingly as

$$\begin{aligned} F_{DSB} &= \frac{P_{Si}/P_{Ni}}{P_{So}/P_{No}} = \frac{P_{Si}/P_{Ni}}{2P_{Si}G/(2P_{Ni}G + P_{Nm})} \\ &= 1 + \frac{P_{Nm}}{2P_{Ni}G} = \frac{1}{2}F_{SSB} \end{aligned} \tag{13.10}$$

upon comparison with (13.9). The DSB noise figure is thus a half of or 3-dB lower than the SSB noise figure. This is expected since there is twice as much signal power in the DSB case as compared to the SSB case while the noise power is the same for both cases. The SSB noise figure is normally considered in typically mixer design. It is noted that when an image signal is not present but filters, whose stop-bands encompass the image frequency, are used in mixers, then the noise at the image frequency will be reduced due to filtering. As a result, the SSB noise figure will be less than that without filters and hence the difference between the SSB and DSB noise figures is less than 3 dB.

13.1.2.3 Isolation. Isolation in mixers measures how well the mixer's ports are isolated from each other and is defined as the ratio between the power or voltage at one port and that at the other port at the same frequency. Ideally, it is desired to have no leakage between any two of the three mixer ports; that is, at the RF, LO, and IF port, it is desired to have only the RF, LO, and IF signals, respectively. In principle, four kind of isolation are often used in mixers: LO – RF isolation, LO – IF isolation, RF – LO isolation, and RF – IF isolation and, among them, the LO – RF and LO – IF isolations are the most important ones primarily due to the typically large LO signal used. The LO – RF isolation characterizes the feed-through of the LO signal

from the LO port onto the RF port and is defined as the ratio between the power at the RF port at the LO frequency to that entering the LO port at the same frequency. This isolation is more important than the RF – LO isolation since the LO signal is typically much larger than the RF signal and its leakage to the RF port can cause various detrimental effects such as unwanted radiation through the receive antenna, interference in other components preceding the mixer, and additional spurious responses in the mixer due to possible mixing of the LO leakage reflected back to the mixer with other signals including the original LO signal. Particularly, the mixing between the reflected LO leakage with the original LO signal produces DC, which results in a DC offset in the mixer. The DC-offset level can be amplified along with the IF input signal in a subsequent IF amplifier, causing possible problems in the IF amplifier's operation and/or subsequent components, such as degradation in linearity and hence system performance, and may even potentially damage these circuits if the DC level is sufficiently large. This undesired effect is more pronounced in direct-conversion mixers with IF signal being in the baseband. In some mixer design, the LO and RF frequencies are close to each other, making it difficult to reject the LO signal leaking onto the RF port using filters and hence poor isolation. High LO – IF and RF – IF isolation are easy to achieve using an LPF at the IF port due to a (typically) wide separation from the LO and RF frequencies to the IF frequency. Due to the fact that the LO signal is much stronger than the RF and IF signals, the LO – IF isolation is more concerned than the RF – IF isolation. Nevertheless, high RF – IF isolation is still desired for well-designed mixers. For the same reason, the RF – LO isolation is less significant than the LO – RF isolation. The level of port-to-port isolation is determined by a mixer type such as single-ended, single-balanced, double-balanced, and some mixer types can provide inherent isolation or enhanced isolation as compared to others. Additionally, the isolation can also be increased using filters provided that the frequencies of the interested and noninterested signals are sufficiently apart.

13.1.2.4 Linearity. Although mixers are operated as a nonlinear device producing additional signals with frequency translation through a (nonlinear) mixing process, their linearity is an important figure of merit. It is important to note the fundamental difference between nonlinearity and linearity for mixers. Mixers need to operate in their "nonlinear" state in order to conduct a mixing function. On the other hand, their performance needs to be as "linear" as possible in order to produce high-quality signal conversion. That is, mixers need to operate nonlinearly but perform linearly. The linearity measures one of the qualities of mixers in handling receiving RF signals, just like for LNA. There are two important aspects in signal handling: one is the maximum tolerable level of the input RF signal that is converted into the desired output IF signal and another one is the maximum level of the input RF signals that are converted into undesired signals near or within the IF signal band. The first one is described by a power compression magnitude, typically 1 dB, and the latter is characterized by the intermodulation (IM) levels, mostly the third-order intermodulation (IM3) point.

1-dB Power Compression

Mixers display linear performance within its linear region in which the RF input signal level is relatively small. In this linear region, the conversion gain of mixers is approximately constant and hence the output IF signal increases linearly with respect to the input RF signal. However, as the RF signal is increased beyond the linear range, reaching a large-signal level, the conversion gain is no longer constant and the output power begins to saturate and reduces as the input power is increased, causing mixers to compress. The 1-dB power compression point ($P_{1\,\mathrm{dB}}$) measures the departure of mixers from their linear operation in which the output IF signal power increases linearly with the input RF signal power. The 1-dB input or output power compression is the respective input RF or output IF signal level at which the actual output IF power is 1 dB less than the linearly increased output IF level; they are related as

$$P_{1\,\mathrm{dB}}^{\mathrm{in}} = P_{1\,\mathrm{dB}}^{\mathrm{out}} - G(\mathrm{dB}) - 1\,\mathrm{dB} \tag{13.11}$$

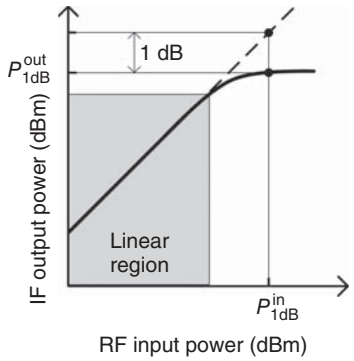


Figure 13.3. Input and output 1-dB power compression points. The dashed line is an extrapolation of the linear line.

where $P_{1\,dB}^{in(out)}$ and G(dB) are the input (output) $P_{1\,dB}$ (in decibel-milliwatt) and (positive) conversion gain (in decibel), respectively. Figure 13.3 illustrates the 1-dB power compression points. The 1-dB power compression point is the point at which the actual power (solid line) is 1 dB lower than the theoretical power assuming linear performance (dashed line). In the linear region, the output power increases linearly with the input power according to the mixer's conversion gain. The slope of the line in the linear region is 1 due to the fact that a 1-dBm increase in the input RF power causes a 1-dBm increase in the IF output power. The 1-dB input power compression point is typically used as the maximum input power that a mixer can handle or considered the upper limit of the mixer's dynamic range.[2] Beyond this power level, a portion of the RF input power is not converted into the desired IF signal; instead, it converts into heat and IM products, which may obscure the desired IF signal. Similar definition is used for other power compression-point levels such as 0.5 or 3 dB. It is noted that the definition for the power compression point is also the same as for other RFICs such as amplifiers.

Third-Order Intermodulation and Intercept Point

IM in mixers is caused by the mixing of the RF signals or their harmonics with the LO signal or its harmonics and can be characterized by different signals (tones) and orders, depending on the number and harmonics of the involved RF signals. The existence of concurrent multiple RF signals is possible in practical operations of mixers, particularly in multichannel receivers. IM products are undesired signals and need to be kept as small as possible. There are two kinds of IM products: single-tone and multi-tone. The single-tone IM products are generated when there is only one RF input signal and are spurious signals formed by the mixing of the harmonics of that RF signal with the harmonics of an LO signal. Their frequencies are given in (13.6). Multitone IM products involve two or more RF signals occurring simultaneously and are the mixing products of these RF signals, their harmonics, and an LO signal and its harmonics.

Intercept points of mixers, typically given in decibel-milliwatt, measure the suppression of the mixers' IM products. Intercept point is normally specified as either input intercept point (IIP) or output intercept point (OIP). The IIP is the RF input power at which the desired IF output power is equal to the undesired IM product. The IIP is thus the point that the IF and IM power curves intercept. The OIP is the corresponding output power level, which is theoretically equal to the IIP plus the mixer's conversion gain according to mixer operation. Practical mixers are compressed under certain input RF power levels and, if this occurs, then the IF and/or IM power curves should be linearly extrapolated pass the 1-dB power compression point until they intersect each other. An intercept point is indeed related to its corresponding IM product and hence, as the IM products are desired to be as low as possible, it is desired to have the intercept points as high as possible. A mixer with high intercept points implies that it can handle large RF input signals before producing IM products that obscure the desired IF signals.

[2]The mixer's upper dynamic range is actually set by the smaller of the input P_{1dB} and the input third-order intercept point described later. The lower limit of a mixer's dynamic range is the minimum RF input signal level that the mixer can detect, which must be at least above the mixer's noise floor.

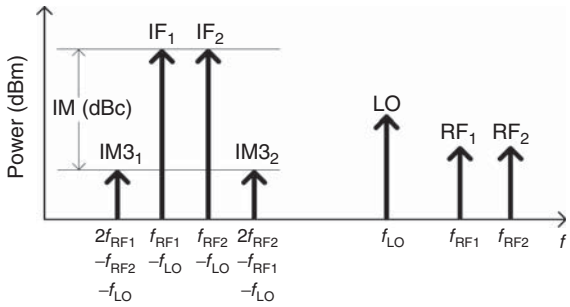


Figure 13.4. IM3 signals with respect to desired IF signals.

Among the different IMs, the two-tone IM, obtained when there are two RF signals at f_{RF1} and f_{RF2} entering a mixer simultaneously, is most interested in mixer design and performance. The two-tone IM products of $(m_1 + m_2)^{th}$ order occur at frequencies $(m_1 f_{RF1} + m_2 f_{RF2}) + nf_{LO}$, where m_1, m_2, and n are positive and negative integers including zero. The IM3 products,[3] located at frequencies $(2f_{RF1} \pm f_{RF2}) \pm f_{LO}$ and $(2f_{RF2} \pm f_{RF1}) \pm f_{LO}$, with f_{RF1} and f_{RF2} being very close to each other, are generally the most concerned ones since they are very close to the desired output signals and hence can potentially disrupt the desired signals. For down-conversion mixers with $f_{RF} > f_{LO}$, the IM3 products occur at $(2f_{RF1} - f_{RF2}) - f_{LO}$ and $(2f_{RF2} - f_{RF1}) - f_{LO}$, hence are very close to or fall within the desired IF band, making it very difficult or impossible to filter out. These IM products would obscure the desired IF signal. Figure 13.4 illustrates the third-order intermodulation signals ($IM3_1$ and $IM3_2$) located near the desired IF signals (IF1 and IF2) for down-conversion mixers. We assume the RF input signals (RF1 and RF2) at f_{RF1} and f_{RF2} have an equal amplitude. Since these IM3 products are spurious signals, it is desired that they are kept as small as possible as compared to the desired IF signals at any given input RF signal level. As the input RF signal level is increased, both the desired IF and undesired IM3 signals increase. Since the IF and IM3 signals increase at different rates due to different conversion rates, they become equal at a certain RF signal level. If the mixer does not compress at this RF level, then this equality corresponds to the actual IF and IM3 signal magnitude. Otherwise, the equal point is obtained through extrapolation of the IF and IM3 signals. This RF signal magnitude and the corresponding (equal) IF and IM3 signal level are referred to as the input (IIP3) and output (OIP3) third-order intermodulation or intercept point,[4] respectively. The IP3 of a mixer depends on the frequencies of the two RF tones and LO power level. However, under a certain operating LO power with two RF tones within the mixer's operating bandwidth, the IP3 should maintain the same level. IIP3 specifies the level of the RF signal that produces equal IM3 and IF signal, and OIP3 is the actual IM3 power. Under this condition, the (output) IM3 signal is generally too large to be handled with respect to the desired IF signal, and the mixer is considered ineffective. The input third-order intercept point hence indicates the maximum input RF signal that the mixer should be subjected to in addition to the maximum input RF power that causes the 1-dB power compression discussed previously. The smaller of this RF signal level and the RF 1-dB power compression level actually sets the mixer's upper dynamic range. As for the 1-dB power compression point, the IP3 points characterize the linearity of mixers in handling RF signals. Typically, the IIP3, just like the input 1-dB power compression point, is preferred over the OIP3 in describing the mixer's linearity. High intercept point, in general, and third-order intercept point, in particular, for a mixer implies high mixer's linearity and that the mixer can handle strong RF input signals before causing large IM signals that degrade the mixer's performance. Figure 13.5 plots the magnitude of the mixer output IM3 and IF signals as a function of the input RF signal power, showing the intercept points IIP3 and OIP3. As IM3 signals are spurious, it is desired that their magnitudes be as small as possible or the IP3 level as high as possible. The 1-dB compression and IP3 points are obtained with respect to the RF input power as can be seen in Figures 13.3 and 13.5, respectively; they are

[3]The third order is obtained according to $m_1 + m_2 = 3$ from the second harmonic of RF1 (or RF2) and the fundamental RF2 (or RF1).
[4]IP3 point is always defined where the IF and IM3 signals are equal, which is very convenience in mathematical description as well as in graphical display. However, other IP3 points, such as those corresponding to the IM3 signal being a half of the IF signal, can also be defined.

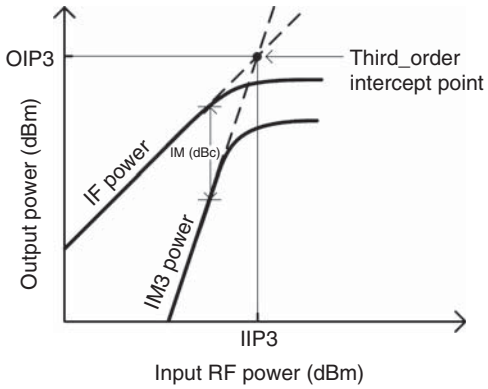


Figure 13.5. IP3 obtained at the intercept point between two lines representing the output IF and IM3 signals versus the RF input signal.

indeed related and hence together characterizing the mixer's linearity and upper dynamic range. The IIP3 (in decibel-milliwatt) can be approximately determined as

$$\text{IIP3(dBm)} = \frac{\text{IM(dBc)}}{m_1 + m_2 - 1} + \text{RF(dBm)} \tag{13.12}$$

where IM(dBc) is the intermodulation suppression or the intermodulation signal's power with reference to the IF power, that is, IM(dBc) = IF(dBm) − IM(dBm), as seen in Figure 13.4, RF(dBm) is the power of the input RF signal in decibel-milliwatt corresponding to the IM suppression level, and $(m_1 + m_2)$ specifies the IM order. The OIP3 (in decibel-milliwatt) can be obtained from (13.12) as

$$\text{OIP3(dBm)} = \text{IIP3(dBm)} + G\text{(dB)} = \frac{\text{IM(dBc)}}{m_1 + m_2 - 1} + \text{RF(dBm)} + G\text{(dB)} \tag{13.13}$$

13.2 MIXER TYPES

There are various MOSFET mixer configurations, typically realized as unbalanced and balanced types. The unbalanced type is known as single-ended mixer. The balanced type is classified into several different topologies: single-balanced, double-balanced, and doubly double-balanced, etc. Balanced mixers can be generally considered consisting of multiple single-ended mixers connected in a particular fashion, which dictates the balanced type. Physically, the unbalanced and balanced mixer are different in the number of transistors used and how the transistors are fed (single-ended or differential feeding) while, electrically, they distinguish from each other through their unique performance characteristics. Balanced mixers provide better performance than single-ended mixers. They have better isolation between ports and inherent rejection of the amplitude-modulation (AM) noise from LO source, spurs, and IM products, of which the levels of isolation and rejection depend on the balanced types. The isolation and signal rejection are obtained without using filters and primarily determined by the phase relationship between the RF, LO, and IF voltages at the devices and mixer ports provided by the balanced structures. High-level balanced structures such as the double-balanced mixer particularly alleviate the isolation problems of RF and LO signals whose frequencies are so close to each other that cannot practically, or even theoretically, be separated using filters. Balanced mixers can handle more RF power and hence have higher linearity than their single-ended counterparts due to the use of multiple MOSFETs. They, however, are more complicated and require stronger LO power drives. For a given mixer type, such as single-balanced, the distributed technique employed for distributed amplifiers discussed in Chapter 11 can also be implemented to combine several mixer cells to realize distributed mixers, which are extremely wide-band. Other types of mixers that are unique in different aspects,

such as passive, image-reject, quadrature, and sampling mixers, also exist; they can be designed to meet certain system requirements that may not be fulfilled using other mixer types. These mixers will be discussed in this section. Table 13.1 provides a comparison between single-ended, single-balanced, and double-balanced mixer cores (without filters and matching networks). Since it is difficult to quantify the performance difference, the comparison is done relatively to the single-ended mixer whose parameters are designated as "0" with "+" means "better" and "++" implies "much better" with respect to the performance of a single-ended mixer.

13.2.1 Single-Ended Mixer

Figure 13.6 shows single-ended mixer topologies employing single-and dual-gate MOSFET. Single-ended mixer is the simplest mixer type and consists of only one MOSFET. It requires the least LO power drive. A diplexer, which typically consists of an RF band-pass filter (BPF) and an LO BPF to feed the RF and LO signals to the MOSFET via its gate terminal, respectively, is needed as shown in Figure 13.6(a). Two separate BPFs can also be used in lieu of a diplexer when the RF and LO feeding points are different as shown in Figure 13.6(b). A LPF is normally used at the IF port to suppress the feed-through RF and LO signals. The simplicity of the single-ended mixer, however, comes with a price in performance. AM noise from the LO source can get into the mixer and degrade the mixer's noise performance. The single-ended mixer has poor inter-port isolation unless proper filters, such as those in Figure 13.6, are used and the operating frequencies are well separated to enable the filters to function properly. Particularly, when the LO signal is fed to the gate terminal, it can be amplified by the MOSFET, possibly resulting in substantial leaking signal at the IF port and hence increasing the LO leakage level at the IF port, even with the existence of a good LO–IF isolation. Better isolation between the RF and LO port can be obtained by feeding the RF and LO signals at separate ports – for instance, gate and source, respectively, as shown in Figure 13.6(b). Improved isolation between the RF and LO ports can be achieved by using a dual-gate MOSFET, as shown in Figure 13.6(c), in which the RF and LO signals are fed via different gates. Two single-gate MOSFETs, with the source of one transistor connecting to the drain of another, can also be used in lieu of a dual-gate MOSFET to enhance the RF–LO and LO–RF isolation. Another problem in single-ended mixers is their inability to suppress harmonic signals intrinsically without using filters, which hampers its function. Single-ended mixers are not typically used in systems due to their poor performance, except for systems requiring only minimal performance with simplest architecture and lowest cost.

13.2.2 Single-Balanced Mixer

In general, a single-balanced mixer can be considered consisting of two single-ended mixers, whose RF and LO signals are fed to the single-ended mixers with certain phases to achieve particular characteristics such as inherent isolation between ports, rejection of spurious responses, and IF signals with desired phase. Figure 13.7 shows a single-ended mixer topology with the RF and LO signals driving two single-ended mixers in-phase and 180° out-of-phase, respectively. This mixer generates two 180° out-of-phase IF signals, which are combined through a 180° out-of-phase (or subtraction) circuitry to produce a single IF signal. The RF and LO signals can also feed the two constituent mixers 180° out-of-phase and in-phase, respectively, to produce two out-of-phase IF signals.

Figure 13.8 shows two block diagrams of single-balanced mixers derived from Figure 13.7 that can be realized in practice. They consist of two MOSFETs, each representing a single-ended mixer, interconnected through a 180° hybrid (also known as balun or transformer) or a 90° hybrid operating at RF and LO frequencies and a 180° hybrid or balun working at IF. The RF/LO balun essentially provides differential RF or LO signals to the MOSFETs, while the IF balun converts a differential into a single-ended IF signal – or, in other words, combining two differential IF signals into a single-ended IF signal. The use of differential ports or the generation of differential signals enables signal cancellation to happen, resulting in inherent isolation. In the mixer topology of Figure 13.8(a), the RF and LO signals are applied to the sum Σ (or difference Δ) and difference Δ (or sum Σ) ports of the RF/LO balun, respectively, and the IF signal is extracted from the difference

TABLE 13.1. Comparison of Mixer Cores

Mixer type	Port type	Port isolation			$P_{1\,dB}$	IIP3	IM suppression	Linearity	LO AM suppression	LO power	DC power consumption	Design complexity	Size
		LO–RF	LO–IF	RF–IF									
Single-ended	Single-ended RF, LO, IF	0	0	0	0	0	0	0	0	0	0	0	0
Single-balanced	Single-ended RF (or LO), differential LO (or RF), IF	+	+	+	+	+	+ (even-order IM canceled)	+	+	+	+	+	+
Double-balanced	Differential RF, LO, IF	++	++	++	++	++	++ (even-order IM and odd-order (with odd LO harmonics) canceled	++	+	++	++	++	++

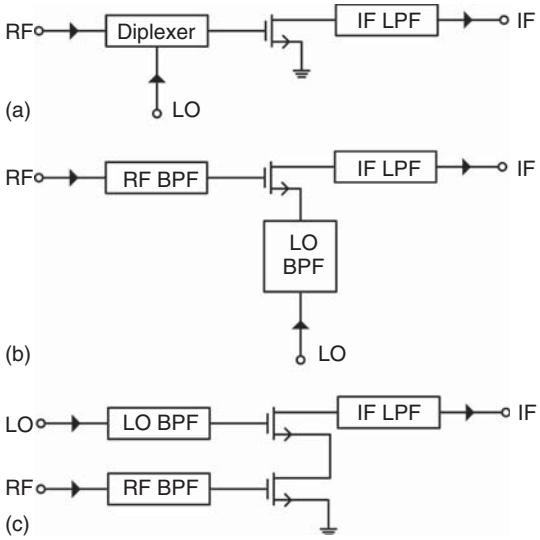


Figure 13.6. Single-ended mixer using single-gate (a), (b), and dual-gate (c) MOSFET. Matching circuits are omitted for simplicity.

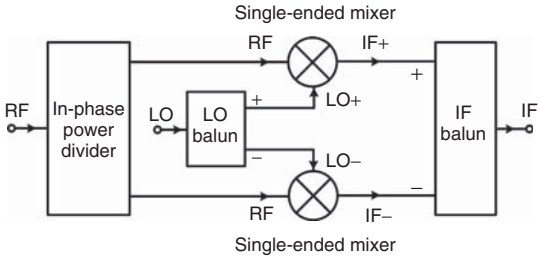


Figure 13.7. Single-balanced mixer consisting of two single-ended mixers.

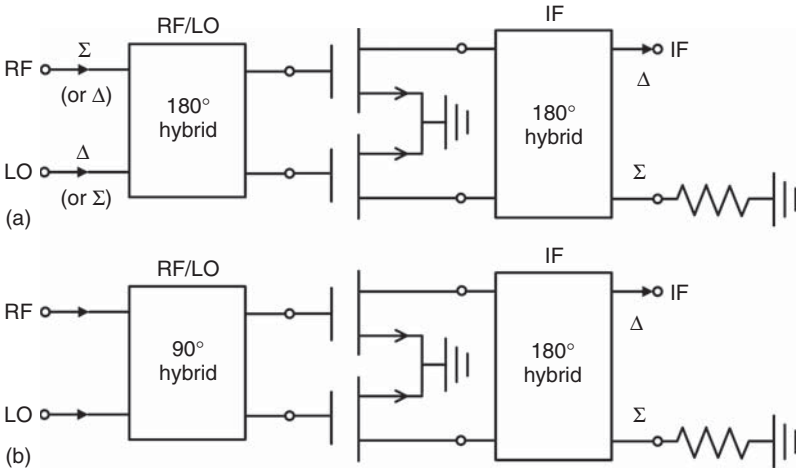


Figure 13.8. (a,b) Single-balanced mixers employing two MOSFETs. Matching circuits are omitted for simplicity.

Δ port of the IF balun. A LPF is normally used at the IF port to suppress unwanted signals. When the RF and LO signals are applied to the Σ and Δ ports, respectively, they split equally in-phase and 180° out-of-phase into the two MOSFETs, respectively. As a result, the IF generated are always 180° out-of-phase and hence are combined at the Δ port of the IF balun. For a perfectly balanced IF balun, there is no IF signal appearing at the Σ port; however, a matched terminating resistor is always used at this port in practice to absorb any possible leaking IF signal. In the mixer topology described in Figure 13.8(b), the 90° hybrid allows the RF and

LO signals each reaching the two MOSFETs 90° out-of-phase, hence generating 180° out-of-phase IF signals, which are then added constructively at the IF balun's Δ port. In contrast to its single-ended counterpart, the isolation between the RF and LO ports using an RF/LO balun is inherent due to signal cancellation resulted from the 180° out-of-phase provided by the balun as can be inferred from Figure 13.8(a). Therefore, the RF – LO and LO – RF isolation is completely determined by the design of the RF/LO balun. On the other hand, the isolation between the RF and LO ports in a single-balanced mixer employing an RF/LO 90° hybrid is not completely inherent and is not determined by the hybrid itself, but rather by the matching between the MOSFETs and hybrid. As can be seen in Figure 13.8(b), the LO (RF) signals reflected from the MOSFETs due to mismatch are canceled at the LO (RF) port while added constructively at the RF (LO) port. Therefore, unless the MOSFETs are perfectly matched to the hybrid, there are always LO and RF signal leakages at the RF and LO ports, respectively, thereby degrading the inherent isolation produced by the hybrid and resulting in poor isolation between the RF and LO ports. The LO – RF and RF – LO isolation is in fact theoretically equal to the return loss at the connecting point between the individual MOSFET and the RF/LO 90° hybrid at the LO and RF frequencies, respectively. The matching between the MOSFETs and hybrid is thus critical not only for the conversion gain but also for the isolation between the RF and LO ports. Since the IF balun is typically designed at IF only, the isolation from the RF and LO port to the IF port in single-balanced mixers is primarily determined by the suppression of the RF and LO signals by the IF balun's out-of-band rejection, not by its balun's anti-phase cancellation characteristic, as well as by an IF LPF. It is noted that, if the IF balun is designed to work at the RF/LO frequencies as well, then the RF and LO signal feed-through to the IF port will be canceled and added, respectively; this design, however, is not typically considered. The single-balanced mixer also has inherent suppression of the AM noise coming from the LO.

Figure 13.9 shows circuit topologies for single-balanced mixers based on the concept described in Figure 13.8(a), in which the LO signal is applied to an LO balun and the IF signal is extracted from an IF balun. The RF signal is fed directly to the sources of the MOSFETs (Figure 13.9(a)) or via another common-source MOSFET acting as an RF driver stage (Figure 13.9(b)). The single-balanced mixer in Figure 13.9(b) is essentially a single-balanced version of the single-ended mixer shown in Figure 13.6(c), whose upper MOSFET where LO is fed to is replaced with two MOSFETs simulating a single-balanced topology. Even though the LO and RF signals are not applied to an RF/LO balun as seen in Figure 13.8(a), the LO signal feed-through to the RF port is also rejected due to its differential feeding. In integrated-circuit

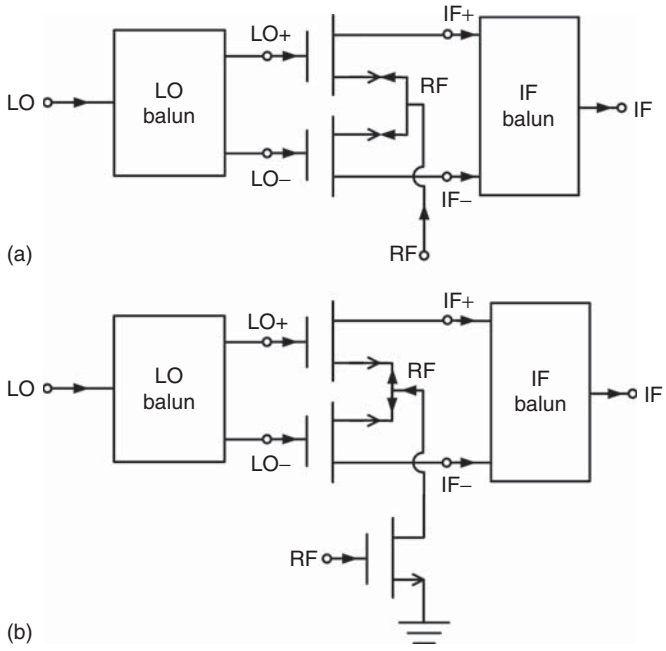


Figure 13.9. Single-balanced mixers with RF signal feeding MOSFETs directly (a) and via another MOSFET (b).

environments, such as a fully integrated CMOS RF receiver, the LO signal may already be available in the differential form and hence the LO balun may be removed, allowing a simpler circuit with better performance. Similarly, the IF balun is not needed when differential IF signals are desired. With the availability of commercial differential probes, it is possible to measure devices with differential ports; therefore, single-balanced mixers can be designed without LO and/or IF balun. It is noted, however, that, when a mixer is designed to produce differential IF signals (i.e., no IF balun), the RF and LO signal feed-through is not subjected to suppression by an IF balun, hence degrading the isolation between the RF and LO port to the (differential) IF port. This poor isolation is expected since, in this case, each mixer-half carrying the RF and LO+ (or LO−) signals basically acts as a single-ended mixer. It is particularly noted that it is important to consider the design of a single-balanced mixer together with other components preceding it (e.g., LNA) and following it (e.g., IF amplifier) so that proper port configurations can be used for both the mixer and other components. When the LO balun is not used, the isolation between the LO and RF is not determined by the LO balun but by the circuit preceding the LO port. Nevertheless, the preceding circuit provides differential LO signal, and so good isolation between the RF and LO ports is still obtained due to the anti-phase signal cancelation of that circuit.

As compared to the single-ended mixers, the single-balanced mixers have several advantages including inherent isolation between the RF and LO ports, better isolation from the RF and LO ports to the IF port, inherent suppression of LO AM noise, rejection of certain spurious responses, and higher linearity. Particularly, the single-balanced mixers suppress the even-order[5] of the IM products more than the odd-order. In the ideal case of perfect mixer balance (identical circuit elements including MOSFETs and perfectly symmetrical layout), the even-order products are completely canceled. That is, half of the possible IM products that would otherwise occur in single-ended mixers do not exist in single-balanced mixers. The single-balanced mixers, however, consume more DC power (approximately twice as much) and require larger LO drive (roughly twice), more complex design, and differential ports at the LO and IF ports. Differential ports require baluns to convert into single-ended ports for the LO and IF signals. Inclusion of baluns not only makes the design more complicated, but also potentially reduces the bandwidth and linearity (if an active balun is used for the RF signal) of a mixer. Perhaps, most importantly with respect to size and cost, baluns can increase the mixer die size, and hence cost, significantly due to their typically large size. The drawbacks of using baluns, however, are avoided if the LO and IF single-ended ports are not needed – for instance, when differential LO signals are available from a preceding LO or differential IF signals are needed for a subsequent component. If differential ports are desired for the LO and IF signals in integration, then they add to the advantages of the single-balanced mixers. Under this situation, however, differential probes need to be used for on-wafer measurement of mixers.

13.2.3 Double-Balanced Mixer

Double-balanced mixers can be considered comprising two single-balanced mixers or four single-ended mixers with proper phases for the RF and LO signals, as shown in Figure 13.10. The RF signal feeds single-ended mixers 1 and 2 in-phase and 3 and 4 180° out-of-phase via an RF balun, while the LO signal drives mixers 1 and 4 in-phase and 2 and 3 180° out-of-phase through an LO balun. The IF signals generated from mixers 1 and 3 are in-phase while those from 2 and 4 are out-of-phase. These IF+ signals and IF− signals are combined respectively and the final differential IF+ and IF− signals are subtracted through an IF balun to produce a single-ended IF signal. Other phases for the RF and LO signals can also be arranged to produce desired IF signals. For instance, the RF signal can be fed in-phase to the first and third single-ended mixer and 180° out-of-phase to the second and fourth mixer, while keeping the LO signal the same as in Figure 13.10. This arrangement, however, produces the IF+ and IF− signals from the first–second mixer pair and the third–fourth mixer pair, respectively. The IF signal is hence obtained as the difference between the IF signals coming from these corresponding mixer pairs.

[5]The order refers to the order of RF tones, that is, m in $mf_{RF} + nf_{LO}$ for single-tone RF and $(m_1 + m_2)$ in $(m_1 f_{RF1} + m_2 f_{RF2}) + nf_{LO}$ for two-tone RF.

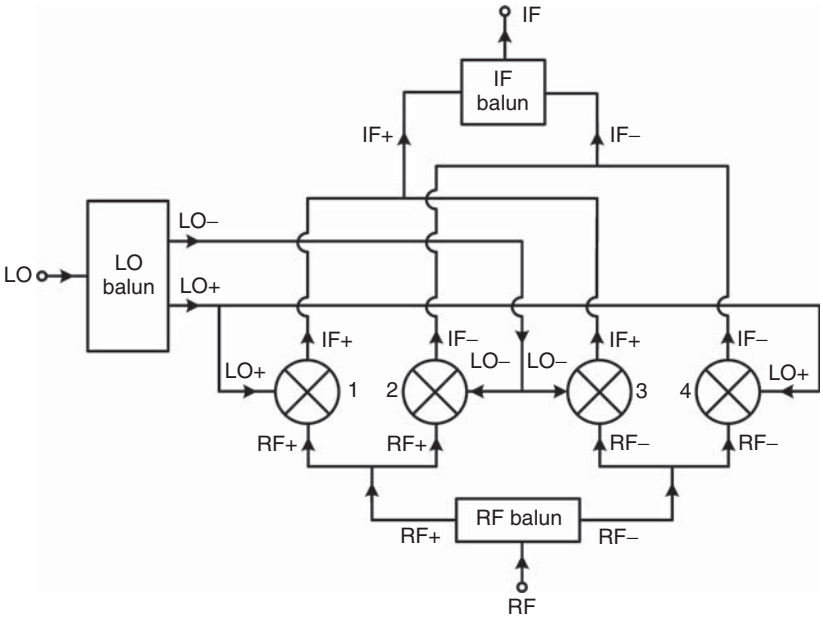


Figure 13.10. Double-balanced mixer consisting of four single-ended mixers.

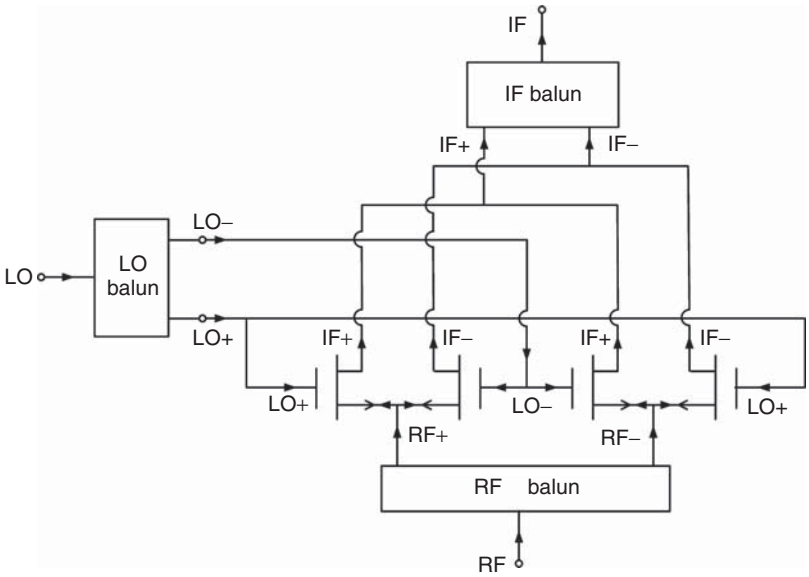


Figure 13.11. A MOSFET double-balanced mixer.

Figure 13.11 shows a double-balanced mixer consisting of four MOSFETs, each representing a single-ended mixer as seen in Figure 13.10, and RF, LO, and IF baluns working at RF, LO, and IF frequencies, respectively. The RF and LO signals are applied to the RF and LO baluns, respectively, to provide in-phase and 180° out-of-phase signals to the MOSFETs, and the IF signal is extracted from the difference Δ port of the IF balun. The RF and LO ports are electrically isolated from each other due to the LO and RF baluns provided that both of these baluns work at the RF and LO frequencies. For instance, the RF+ or RF− signals leaking toward the LO port will arrive at the LO balun's anti-phase ports and hence cancel at the LO port. The RF+ and RF− signals (or LO+ and LO− signal) arrive at each of the two combining points for IF+ signals and IF− signals 180° out-of-phase and hence cancel each other, resulting in inherent RF − IF isolation (or LO−IF isolation). Perfect cancellation of the RF and LO leaking signals at the combining

points happens only when there is perfect symmetry of the RF and LO paths with respect to the combining points. In reality, this would never happen and so the cancellation depends on the electrical and physical symmetry of the design. Care must be exercised to ensure that all electrical elements including MOSFETs in the RF + (LO+) and RF – (LO–) paths are matched to each other and the layouts of these paths are symmetrical as much as possible. Note that the cancellation of the RF and LO signal feed-through occurs before reaching the input ports IF+ and IF– of the IF balun. Additional suppression of the RF and LO signal feed-through is also provided by the IF balun's out-of-band performance as well as by an IF LPF.

Figure 13.12 shows a MOSFET double-balanced mixer topology based on the concept described in Figure 13.11 with additional common-source MOSFET driver stages for the RF signal. This mixer is commonly referred to as the Gilbert mixer [1] and is among the most widely used mixers in CMOS RFIC. It essentially combines two single-balanced mixers described in Figure 13.9(b). In fully integrated subsystems or systems, the RF and/or LO signal may be available in a differential form, and/or a differential IF signal may be desired; in this case, the RF, LO, or IF balun is not needed. As mentioned for single-balanced mixers, it is important to consider the design of a double-balanced mixer together with other components preceding it (e.g., LNA) and following it (e.g., IF amplifier) so that proper port configurations can be decided for both the mixer and other components. It is noted that, when an RF or LO balun is not used, the LO – RF or RF – LO isolation is not determined by the respective balun but rather than by the circuit preceding the RF or LO port. Nevertheless, these preceding circuits provide differential RF and LO signals and so good isolation between RF and LO ports is still obtained due to their anti-phase signal cancelation.

As compared to the single-balance mixers, the double-balanced mixers have better isolation between the RF, LO, and IF ports, higher linearity, high intercept points, more rejection of spurious responses, and better suppression of external noise or signal (e.g., noise from power supplies or signal from nearby circuits). In fact, the RF, LO, and IF ports are inherently isolated from each other. If the mixer balance is perfect (identical elements including MOSFETs and perfectly symmetrical layout), three quarters of the possible IM products would not appear. Specifically, the even-order IM products (including those associated with the even or odd LO harmonics) and the odd-order containing the even LO harmonics are completely suppressed, leaving the odd-order IM products with odd LO harmonics – for example, $3f_{RF} + f_{LO}$ for single-tone RF or

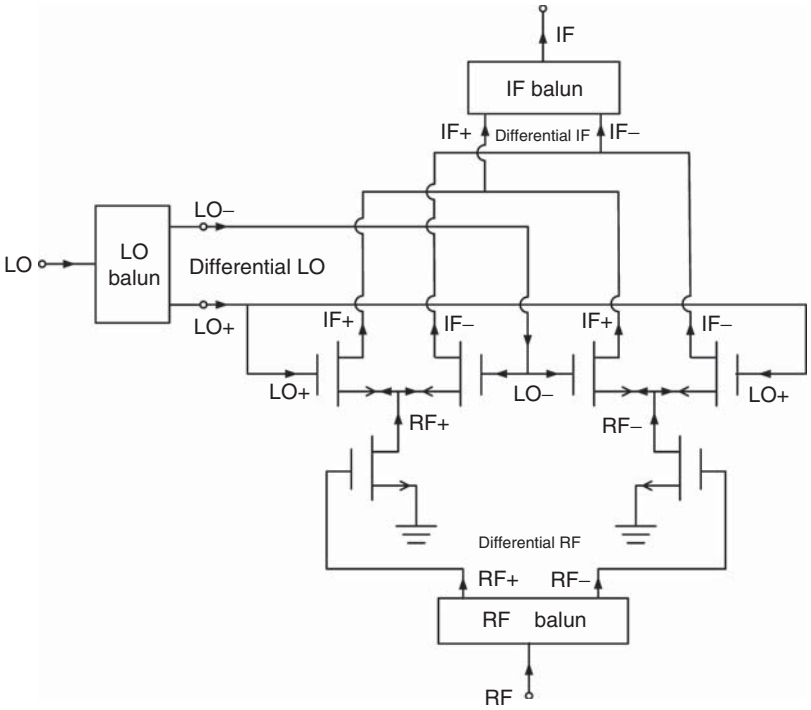


Figure 13.12. Gilbert MOSFET double-balanced mixer.

$(f_{RF1} + 2f_{RF2}) + f_{LO}$ for two-tone RF – as the only products appearing at the IF port. Practical mixers, however, are not perfectly balanced, and hence the even-order IM products are not totally rejected. Nevertheless, the suppression of the even-order IM products is higher than that of the odd-order IM products. The suppression of external noise and signal occurs at the combining RF, LO, and IF points due to signal common to both RF+ and RF− paths, common to both LO+ and LO− paths, and common to both IF+ and IF paths. This feature is very desirable in practical RFICs, especially those integrated directly with digital circuits on the same chip. The double-balanced mixers, however, consume more DC power and require larger LO drive (about twice as much DC power and LO drive), more complex design, larger die size, and differential port for the RF signal. The need of additional differential port for the RF signal causes additional design problems, particularly size and cost due to an additional balun at the RF port, similar to what discussed previously for the single-balanced mixers. The availability of a differential RF signal, however, is considered an advantage if it, instead of a single-ended RF signal, is desired. Even though double-balanced mixers have several drawbacks, their superior performance in many aspects as compared to single-ended and single-balanced mixers makes them the preferred choice for many systems and applications.

13.2.4 Doubly Double-Balanced Mixer

Figure 13.13 shows a doubly double-balanced or triple-balanced mixer, which integrates two double-balanced mixers described in Figure 13.12. This mixer, as can be imagined, is much more complicated than the commonly used doubled-balanced Gilbert mixer. It is extremely difficult, if not impossible, to fabricate this mixer in nonmonolithic RF techniques. In CMOS RFICs, however, it is possible to implement such a mixer. Similar to the Gilbert mixer and as noted in Figure 13.13, the RF and LO signals from the RF and LO baluns, respectively, provide in-phase and 180° out-of-phase signals to the MOSFETs and the IF signal is derived from the IF balun. This mixer also has inherent isolation between the RF, LO, and ports. Moreover, it has a greater dynamic range and IM suppression, as well as higher linearity and intercept points than the double-balanced mixer. However, it has larger size and requires more DC power, LO drive, and complicated design, and hence

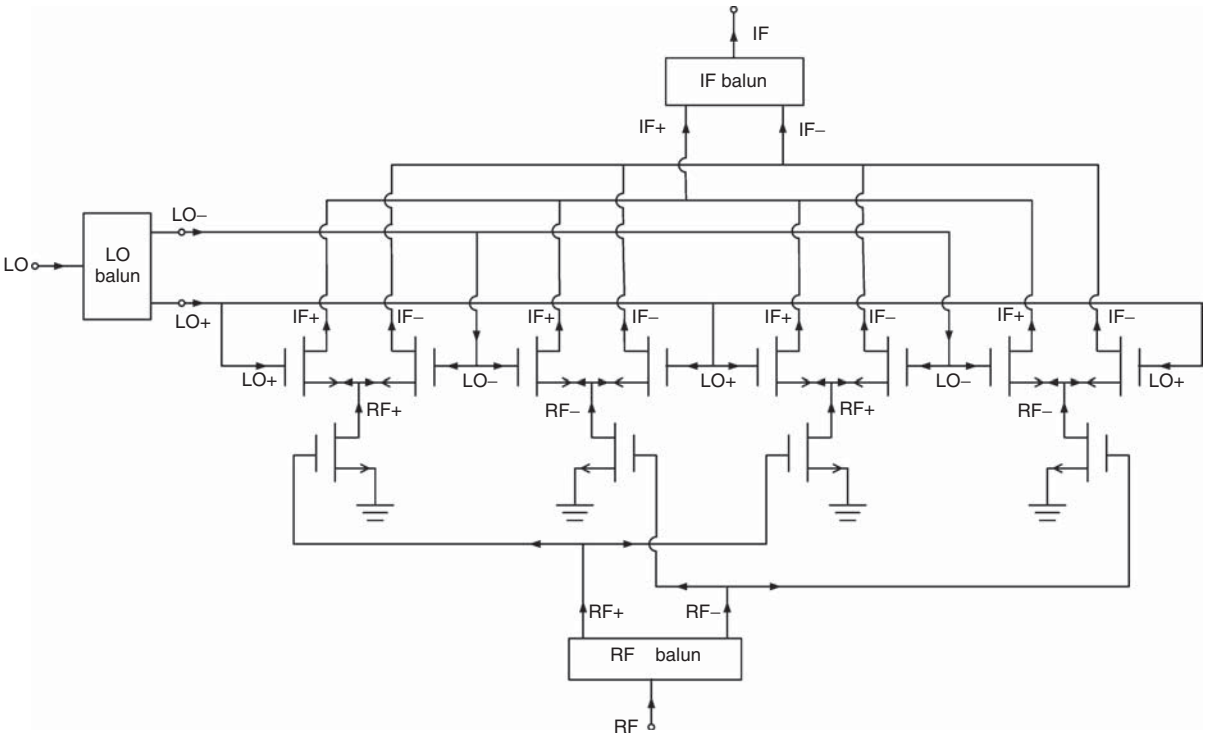


Figure 13.13. A doubly double-balanced MOSFET mixer.

is not very common for implementation in systems. However, due to their exceptional RF input power handling capability, doubly double-balanced mixers may be useful as up-converters for transmitters where they can accommodate large RF input signal levels, particularly for millimeter-wave systems where generation of high power is more convenient and less expensive at lower frequencies (before signal upconversion) than at millimeter-wave frequencies using PAs.

Although other more complicated higher-level mixers can be designed in CMOS RFIC, their advantages in system implementation do not outweigh their disadvantages to warranty their use in systems.

13.3 OTHER MIXERS

The forgoing discussed mixers define different mixer types for active mixers whose conversion efficiency is positive – that is, possessing conversion gain. By combining several mixer cores of a mixer type, for example, single-balanced, various other mixer topologies can be formed. This section discusses some important mixers based on this concept.

13.3.1 Passive Mixer

Passive MOSFET mixers can be realized based on similar topologies for active mixers such as the single-ended, single-balanced, double-balanced, and doubly doubled-balanced described earlier. The main difference between passive and active mixers is passive mixers use MOSFETs operating in the passive (unbiased) mode, where (mostly) only gate bias is applied, to conduct the signal conversion. One of the first passive FET mixers implementing passive MESFETs was reported in 1987 [2]. Passive mixers basically operate as a switch represented by the unbiased transistors modulated by a LO. An unbiased transistor used in passive mixers works as a modulated switch based upon the transistor's channel resistance which varies linearly as a function of the gate bias voltage with respect to time. This linear time-varying voltage-controlled resistance produces zero IM in signal conversion under an ideal condition. In order for passive mixers to work well, the time-varying channel resistance must be linear; this is the primary operating principle that passive mixers must adhere to. In practice, however, unbiased CMOS transistors, just like those used in CMOS switches discussed in Chapter 14, do not behave linearly at large RF input power, hence limiting the mixer's linearity and degrading the overall performance. As discussed in the MOSFET's operation in Chapter 9, a MOSFET exhibits a linear behavior in the linear region under which it functions as a (gate) voltage-controlled resistor. This behavior is exploited to design switches as mentioned in Chapters 9 and 14, and switches, as indicated early, form the basic element for passive mixers. As compared to active mixers, passive mixers have several advantages including higher linearity, better IM, higher power compression, wider dynamic range, and no DC power consumption. Passive mixers, however, have conversion loss (instead of gain), higher noise figure, and require larger LO drive. In practice, active mixers are typically preferred than passive mixers in most systems, primarily due to their conversion gain and lower noise figure. However, in systems where a substantial gain is needed for or provided by the LNA that precedes a mixer, the mixer would need a very high linearity, making a passive mixer an attractive candidate.

Figure 13.14 shows a passive CMOS doubled-balanced mixer employing four MOSFETs connected in a ring, much like a double-balanced diode mixer, and three baluns (RF, LO, and IF) [3]. The LO source drives the quad-MOSFET mixer core with equal amplitude but 180° out-of-phase at the junctions between transistor pairs M1/M3 and M3/M4, while the RF signal feeds M1/M4 and M2/M3 with 180° out-of-phase. As can be seen in Figure 13.14, the two output ports of the LO balun is connected with a pair of identical MOSFETs (M1/M4 pair and M3/M2 pair). Therefore, from the operating principle of baluns, the common connection in each pair (A, B, C, and D) represents a virtual ground for the LO signal. Similarly, since the identical transistors M1 and M2 (or M4 and M3) are connected to the output ports A and C of the RF balun, their common connection point B (or D) is a virtual ground for the RF signal. Furthermore, we can also see that the output ports A and C of the RF balun are virtual grounds of the IF signal and the gate of each transistor is a virtual ground for the RF and IF signals. It is noted that the gates of the transistors are not virtual grounds

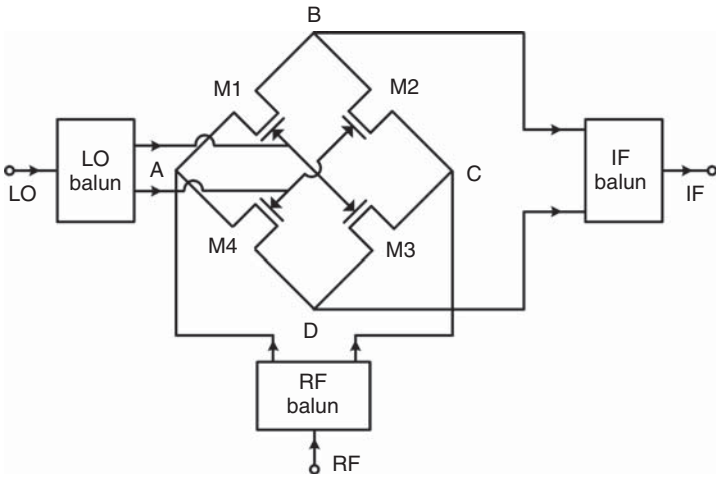


Figure 13.14. Passive CMOS double-balance mixer employing a quad-MOSFET ring.

for the LO signal due to the fact that M1/M4 and M3/M2 are not interconnected at their gates. Due to these virtual grounds, the RF, LO, and IF ports of a double-balanced FET ring mixer are inherently isolated.

The LO drive is used to turn on the switches represented by M1, M2, M3, and M4 for the mixer to operate, thereby requiring significantly high LO power. The LO signal turns on a transistor pair (e.g., M1/M3) and turns off the other pair (e.g., M2/M4) during each of its positive half-cycles. On the other hand, during each of the LO's negative half-cycles, the transistor pairs that were turned off and on during the LO's positive half-cycle are switched on and off, respectively. When a transistor pair is turned on, the output ports of the RF balun are connected to the input of the IF balun via these on-transistors. An examination of the mixer operation reveals that the RF signals arriving at the input of the IF balun corresponding to M1/M3 on, M2/M4 off and M1/M3 off, M2/M4 on are 180° out-of-phase or have opposite polarity. As the RF and LO signals are applied to the MOSFET ring, the switching occurs at the LO frequency according to the LO signal modulation, consequently performing a signal conversion from the RF to IF signals similar to the mixing concept based on switching function described for active mixers in Section 13.4.

13.3.2 Image-Reject Mixer

As discussed earlier, the RF image signal located symmetrically opposite to the desired RF signal with respect to the LO signal – for example, at image frequency $f_{IM} = f_{LO} - f_{IF}$ for $f_{RF} > f_{LO}$ – is also, along with the desired RF signal, converted to a signal at IF by mixing with the LO signal. This RF image signal and its converted IF signal can be real or noise signal which degrades the mixer, and hence system, performance in both signal detection and noise. As an example of noise degradation, the noise contribution of an LNA preceding the mixer at the image frequency would increase the (SSB) noise figure of the mixer caused by the LNA's noise at the RF frequency by 3 dB due to the conversion of the added RF image noise. The image signal thus needs to be suppressed as much as possible, and image-reject mixers are a typical component used for this purpose without employing image-reject filters, which are either difficult or sometimes impossible to implement in practice.

Figure 13.15 shows a block diagram of image-reject mixers consisting of two identical mixers, which can be of any type (e.g., single-balanced mixer), an RF 90° hybrid, an LO in-phase power divider, and an IF 90° hybrid. For the sake of discussion, we assume that $f_{RF} > f_{LO}$. It is noted that the phase angles of signals involved in signal transformation in mixers are theoretically preserved. This signal preservation is essential in order to make use of the measurement or function conducted by systems employing mixers – for instance, to reconstruct an incoming RF signals from its down-converted IF signal. As such, the phase of the desired IF signal at $f_{IF} = f_{RF} - f_{LO}$ is equal to the difference $\phi_{RF} - \phi_{LO}$ between the phase of the RF signal (ϕ_{RF}) and that of the LO signal (ϕ_{RF}). First, let us consider the desired RF signal feeding mixers 1 and 2 with equal magnitude

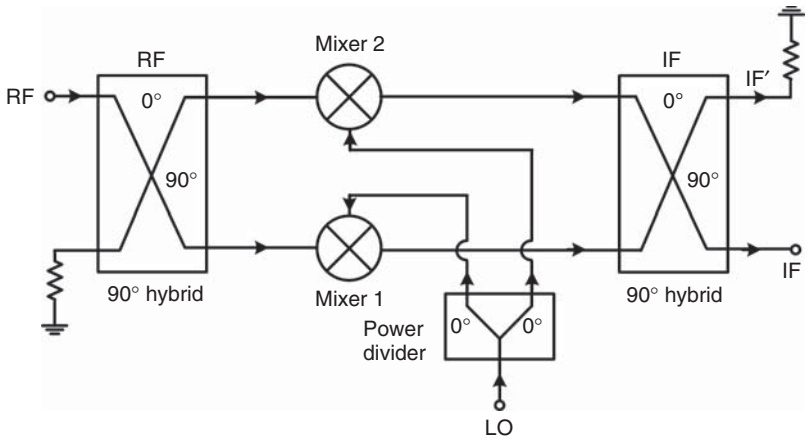


Figure 13.15. Block diagram of image-reject mixers.

and 90° out of phase. The desired IF signal produced by mixer 1 has a phase delay of 90° with respect to that produced by mixer 2 due to the 90° phase difference of the RF signals arriving at these mixers. These desired IF signals are then combined and canceled at the IF and IF′ output port of the IF hybrid, respectively, due to the 90° phase delay of the IF hybrid. The desired IF signal therefore appears at port IF only. Now we consider the undesired RF image signal (real signal or noise) that produces the undesired IF signal at $f_{IF} = f_{LO} - f_{IM}$. The undesired IF signal produced by mixer 2 is 90° delayed with respect to that produced by mixer 1. These undesired signals are then added and canceled at ports IF′ and IF, respectively, due to the IF hybrid. The undesired IF signal hence appears at port IF′ only. A matched resistor is used to terminate port IF′ to absorb the undesired IF signal. The addition and cancelation of the desired and undesired signals are affected by the difference of the RF signals reaching the constituent mixers, the difference of the LO signal driving the mixers, and the difference of the IF signals produced by the mixers. Therefore, not only the phase and amplitude of the RF and IF hybrid and the LO divider need to be carefully considered, but also the match between the two constituent mixers and the symmetry of all the components including interconnects need to be taken care of in the layout in order to produce a well-working image-reject mixer.

13.3.3 Quadrature Mixer

Quadrature mixers, also known as quadrature detectors or quadrature demodulators, convert an RF input signal into two IF signals that are orthogonal to each other. These IF signals are known as in-phase (I) and quadrature (Q) signals. Quadrature mixers are particularly used for measuring both the amplitude and phase of a received RF signal relative to a transmitted RF signal based on the two orthogonal down-converted IF components. These mixers are described in Section 17.2.3.4.

13.3.4 Distributed Mixer

Distributed mixers are based on the transmission-line concept similar to distributed amplifiers discussed in Chapter 11. Similar to distributed amplifiers, distributed mixers can operate over extremely wide frequency ranges and have high linearity, but they suffer from large circuit size, low conversion gain, and high power consumption. The concept of distributed circuit design, analysis, and design equations is already discussed through distributed amplifiers; these can be applied to distributed mixers. Specifically, this concept utilizes inductors or actual transmission lines in conjunction with the parasitic capacitors produced by MOSFET mixer cells to form synthetic transmission lines for the input (RF and LO) and output (IF) signals. This design concept along with a proper mixer-cell topology, such as the Gilbert double-balanced mixer core, can be implemented for designing distributed mixers. A mixer cell may consist of a single or multiple transistors forming a single-ended, single-balanced, or double-balanced mixer core – for instance, the Gilbert mixer

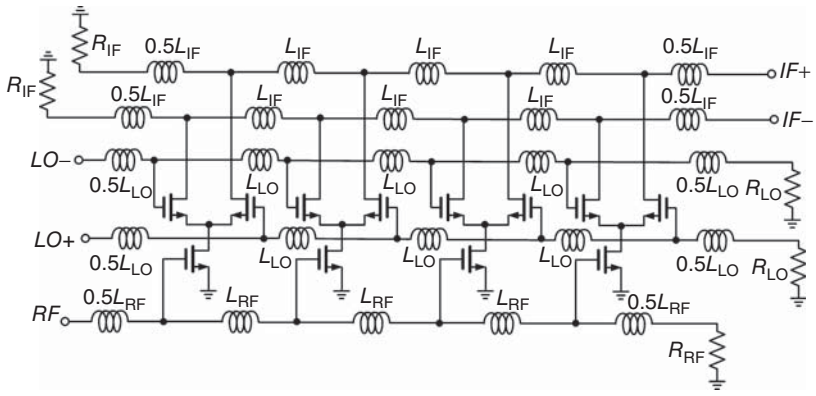


Figure 13.16. Schematic of a MOSFET distributed mixer.

core shown in Figure 13.12 (without the RF, LO, and IF baluns) – and it is this mixer cell that primarily sets a particular performance for the resulting mixer, hence distinguishing one distributed mixer from another. A well-conceived mixer cell can achieve good performance such as high conversion gain, low noise figure, high linearity, and low power consumption. The configuration of a mixer cell and its design are thus the most crucial task in distributed mixer design and should be approached with consideration in gain, noise figure, linearity, and power consumption.

Figure 13.16 shows the schematic of a MOSFET distributed mixer employing multiple single-balanced mixer cells shown in Figure 13.9(b) (i.e., without the LO and IF baluns) along with inductors to form synthetic transmission lines at the RF, LO, and IF ports. Matched terminating resistors are used at one end of the synthetic transmission lines to absorb possible incoming signals. The design of these synthetic transmission lines follows essentially the same technique for distributed amplifiers. As can be seen in Figure 13.16 and, in fact, expected from the single-balanced mixer core in Figure 13.9(b) which the distributed mixer is based on, while the RF port is single-ended, the LO and IF ports are all differential which can be advantageous or disadvantageous depending on the design constraints. If differential ports are desired for the LO and/or IF signal then having such inherent differential ports is most convenient – for instance, a differential LO signal is available only through a differential output oscillator and/or the IF signal is needed as a differential input to a following IF amplifier. On the other hand, if a single-ended port is required for the LO and/or IF signal, then one or two baluns are needed to convert differential into single-ended ports, thereby increasing the circuit complexity, design, size, and possibly degrading the circuit performance and bandwidth. The mixer circuit in Figure 13.6, although can produce extremely wide bandwidth, has low conversion gain, high power consumption (due to many transistors), and large size (due to many on-chip inductors). These are the basic problems facing the design of distributed mixers and, in fact, to all RFIC design based on the distributed-circuit concept like distributed amplifiers. They are particularly concerned for commercial wireless CMOS devices. These issues may be alleviated by employing a proper mixer core as will be described in the following.

Figure 13.17 shows the schematic of a distributed mixer [4] that lends itself to low power consumption and high conversion gain based on the same concept for the distributed amplifier that achieves high gain and low power consumption described in section "Design of Low-Power-Consumption and High-Gain CMOS Distributed Amplifiers". This distributed mixer, while still implementing synthetic transmission lines at the RF, LO, and IF ports as in usual distributed circuits for wideband performance, has only single-ended RF, LO, and IF ports and, particularly, employs a mixer core that results in high gain with low power consumption. It is noted that, as compared to the conventional distributed mixer as shown in Figure 13.16, which requires five synthetic transmission lines, this mixer needs only three synthetic transmission lines, hence reducing the circuit complexity and size significantly. The improved mixer core, as shown in Figure 13.18(a), basically integrates a cascode mixer cell shown in Figure 13.18(b) together with a cascade common-source gain cell shown in Figure 13.18(c), which is described in Figure 11.57(d) and used for the distributed amplifier discussed in

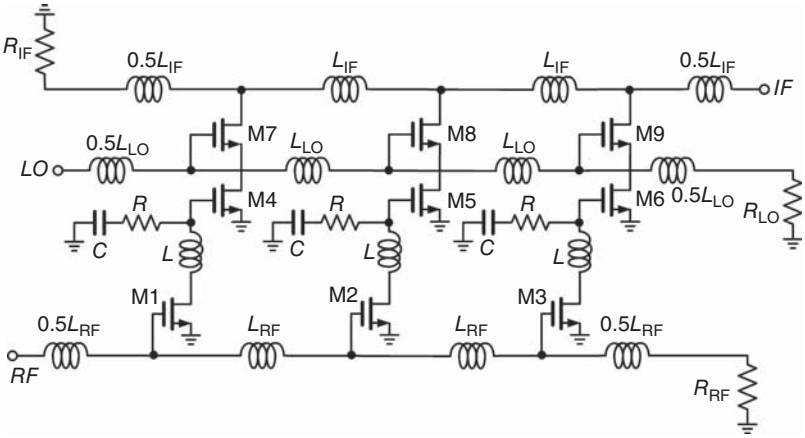


Figure 13.17. Schematic of the distributed mixer employing mixer cores combining cascode and cascade transistors.

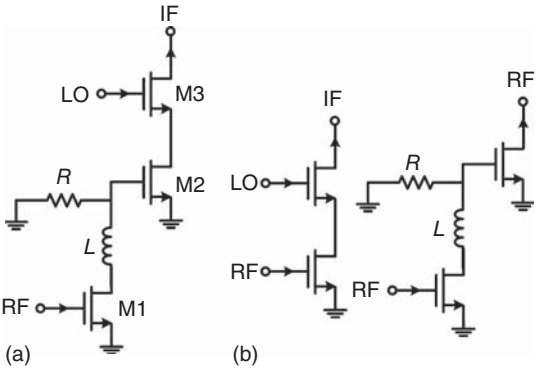


Figure 13.18. Improved mixer core (a) integrating a cascode mixer cell (b) and a cascade gain cell (c).

section "Design of Low-Power-Consumption and High-Gain CMOS Distributed Amplifiers". It is noted that the cascode mixer cell in Figure 13.18(b) is essentially the single-ended mixer cell described in Figure 13.6(c) where the RF and LO inputs are single-ended signals applied to different gates. The combination of a conventional cascode mixer cell, which has small conversion gain, and a cascade gain cell produces an improved mixer cell that can perform signal conversion (through the cascode cell) with high gain (through the cascade cell).

In operation, the cascade gain cell of the mixer core (Figure 13.18(a)) is used for gain enhancement. The common-source transistor (M1) of the gain cell is mainly used for low-noise amplification, providing gain and lessening the noise effect from the following stages. The upper transistor (M2) of the gain cell, which is also the lower transistor in the cascode mixer cell, also contributes some gain. The upper transistor (M3) of the cascode mixer cell works as a switch, performing the frequency translation by turning the current on and off at the frequency of the LO signal as for other mixers as described in Section 13.4. Hence, the RF input signal is amplified by the transistors M1 and M2, and the amplified RF signal is mixed with LO signal at the transistor M3 to produce an IF signal.

Compared with a conventional distributed mixer as shown in Figure 13.16, the improved distributed mixer can provide much higher gain and lower noise figure because of the existence of a cascade gain cell at the RF input. Also, since only signal-ended RF and LO inputs are required, only three synthetic transmission lines are needed, instead of five as in the conventional distributed mixer described in Figure 13.16. The less number of synthetic transmission lines leads to the less number of inductors and hence small chip area. However, this mixer also has several drawbacks. Since only single-ended RF and LO signals are used at the input, they will inevitably appear at the IF output. It is recalled that, for double balanced mixers, the RF and LO leaking

signals appear at the IF port as common-mode signals and therefore can be canceled by the differential IF configuration. However, although the leakages from the RF and LO to IF port can be huge, they can be easily eliminated by a LPF, as either RF or LO signal is typically at a much higher frequency than the IF signal.

The conversion gain of the improved distributed mixer can be approximated as the same as the gain of the distributed amplifier implementing the cascade gain cells in section "Design of Low-Power-Consumption and High-Gain CMOS Distributed Amplifiers", assuming the upper transistors (M7, M8, M9) of the cascode mixer cells act as perfect switches. In practice, however, these cascode transistors do not function like perfect switches, hence lowering the conversion gain. As for other mixers, we can approximate the IF output voltage as the product of the RF and LO signals:

$$V_{\mathrm{IF}} = AV_{\mathrm{RF}} \cos(\omega_{\mathrm{RF}} t + \phi) \cdot V_{\mathrm{LO}} \cos(\omega_{\mathrm{LO}} t) \tag{13.14}$$

where V_{RF} and V_{LO} are the RF and LO voltage amplitudes, respectively, ω_{RF} and ω_{LO} are the RF and LO frequencies, respectively, and A is the gain from the RF port to the IF port, which is normally proportional to the transconductance and load resistance of the mixer. It is noted that the conversion gain of the distributed mixer is also affected by the voltage swing of the LO signal like in other mixers.

The RF bandwidth of the improved distributed mixer is about the same as that of the distributed amplifier discussed in section "Design of Low-Power-Consumption and High-Gain CMOS Distributed Amplifiers", since the dominant elements to determine the bandwidth are the R, L, and C in the cascade gain cells (i.e., between M1 and M4, M2 and M5, M3 and M6) instead of the cut-off frequencies of the synthetic transmission lines.

Figure 13.19 shows the layout of the improved distributed mixer designed using a 0.18-μm CMOS process [5]. All inductors were designed and simulated using the EM simulator IE3D. The circuit simulation was made using Agilent ADS Program. All the simulations were made at an LO power of 5 dBm and current consumption of 170 mA. Figure 13.20 shows the simulated conversion gain as a function of RF input frequency. The IF frequency is 500 MHz and the LO frequency is kept at 500 MHz lower than the RF frequency. A

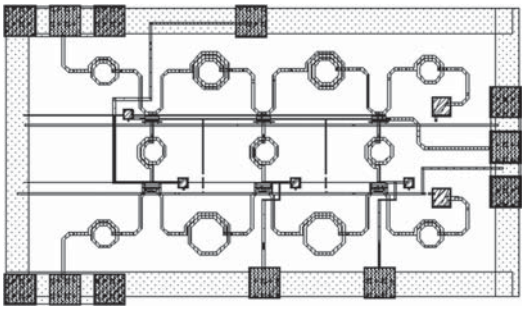


Figure 13.19. Layout of the improved distributed mixer.

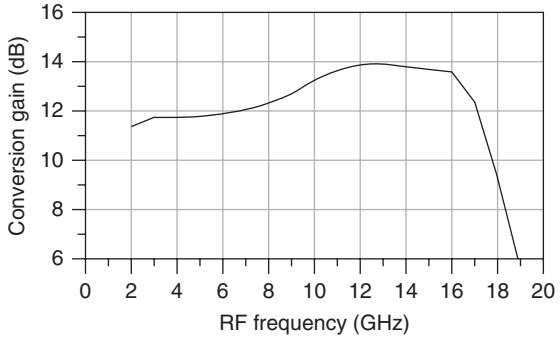


Figure 13.20. Conversion gain of the improved distributed mixer.

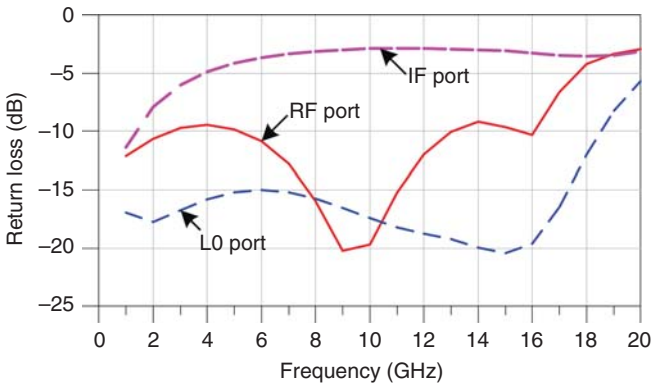


Figure 13.21. Return loss of the improved distributed mixer.

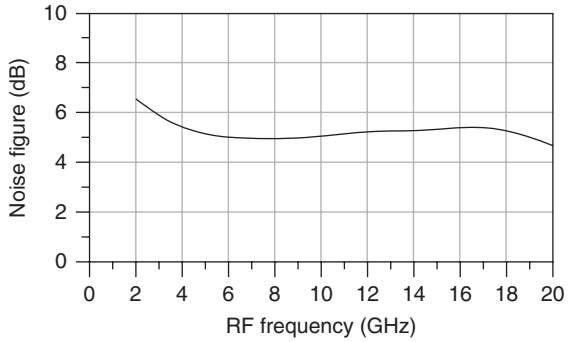


Figure 13.22. Noise figure of the improved distributed mixer.

relatively flat gain around 12–14 dB can be observed from 2 to 17 GHz. Figure 13.21 shows the matching performance at the RF, LO, and IF ports. As can be seen, like other distributed circuits, wideband matching can be easily acquired. The results show that the return loss at both the LO and RF port is less than −9 dB across 1–16 GHz. The IF port has a return loss less than −10 dB below 1 GHz. Since the IF port only handles low frequencies, its matching at high frequencies can be sacrificed for the gain. The calculated noise figure of the improved distributed mixer is shown in Figure 13.22. Around 5–6 dB noise figure can be achieved from 3 to 20 GHz.

13.4 MIXER ANALYSIS AND DESIGN

Mixers, in general, function based on the mixing of RF and LO signals. Specifically, mixers can be considered a mixing device or a switch. The switch-based mixer, from a general point of view, works as a sampling mixer described in Section 13.5, in which the RF input signal is switched on and off periodically by a switch controlled by a pulse signal or a sinusoidal signal functioning as an LO signal. The main difference between a switched-based mixer and a sampling mixer is the use of a resistive (or active) load for the earlier and a sampling capacitor for the latter. This section presents analyses for the single-ended, single-balanced, and double-balanced MOSFET mixers based on the switching concept, which essentially produces the same results as using the mixing approach.

13.4.1 Switching Mixer Fundamental

Figure 13.23 shows a basic switching-based mixer. The signal controlling the switch can be a rectangular pulse or a sinusoidal signal. In practice, these control signals are nonideal – for instance, there is always a rising

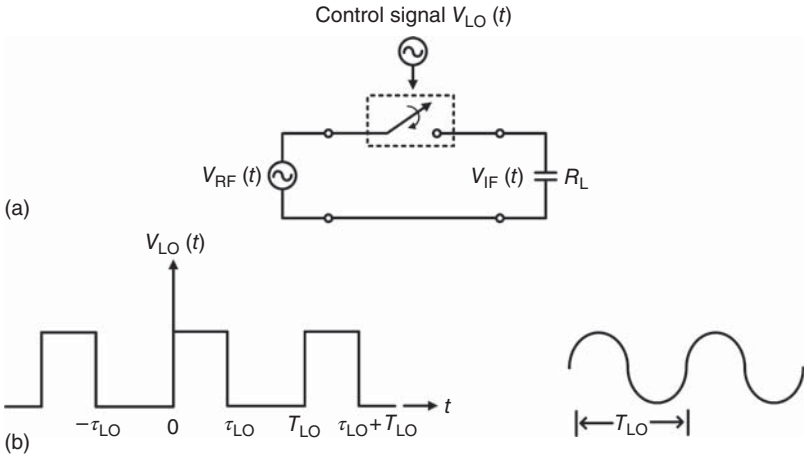


Figure 13.23. Switching mixer (a) and its ideal switch-controlled square-pulse and sinusoidal signals (b). R_L can also be complex. The impedance presented by the RF signal is assumed to be zero. In reality, if an LNA is placed in front of the mixer, then this impedance is the output impedance of the LNA.

time and a falling time for a practical pulse. Also, in typical mixer operation, the control signal is a sine wave, which is the LO signal. For simplicity without loss of generality, we assume that the switch is ideal – that is, it behaves as perfect open and short during on- and off-state, respectively, which is controlled by the control signal (e.g., "on" in a half cycle and "off" in a subsequent half cycle of a sinusoidal LO signal). The ideal control rectangular pulse signal can be described as

$$V_{\mathrm{LO}}(t) = \sum_{n=-\infty}^{n=\infty} V_{\mathrm{LO}} \prod \left(\frac{t - nT_{\mathrm{LO}} - \frac{1}{2}\tau_{\mathrm{LO}}}{\tau_{\mathrm{LO}}} \right) \tag{13.15}$$

where

$$\prod(t) = \begin{cases} 1, & |t| < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{13.16}$$

$T_{\mathrm{LO}} = 1/f_{\mathrm{LO}}$ and τ_{LO} are the period and pulse width, respectively. The (periodic) rectangular-pulse signal can be described using Fourier series as

$$V_{\mathrm{LO}}(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos \omega_n t + b_n \sin \omega_n t) \tag{13.17}$$

where $\omega_1 \equiv \omega_{\mathrm{LO}}$ is the (fundamental) LO frequency, $\omega_n = n\omega_1 (n = 2, 3, \ldots)$ are the LO harmonic frequencies, and a_o, a_n, b_n are the Fourier coefficients given as

$$a_o = \frac{1}{T} \int_{t_o}^{t_o + T_{\mathrm{LO}}} V_{\mathrm{LO}}(t) dt = \frac{V_{\mathrm{LO}} \tau_{\mathrm{LO}}}{T_{\mathrm{LO}}} \tag{13.18}$$

$$a_n = \frac{2}{T} \int_{t_o}^{t_o + T_{\mathrm{LO}}} V_{\mathrm{LO}}(t) \cos \omega_n t dt = 2V_{\mathrm{LO}} \frac{\tau_{\mathrm{LO}}}{T_{\mathrm{LO}}} \frac{\sin \omega_n \tau_{\mathrm{LO}}}{\omega_n \tau_{\mathrm{LO}}} \tag{13.19}$$

$$b_n = \frac{2}{T} \int_{t_o}^{t_o + T_{\mathrm{LO}}} V_{\mathrm{LO}}(t) \sin \omega_n t dt = \frac{2V_{\mathrm{LO}}}{T_{\mathrm{LO}} \omega_n} (1 - \cos \omega_n \tau_{\mathrm{LO}}) \tag{13.20}$$

where t_o is arbitrary (normally chosen as 0 or $-T_{\mathrm{LO}}/2$). The LO pulse signal then comprises an average or DC value of the LO signal and its harmonics. For the considered square-pulse signal, $T_{\mathrm{LO}} = 2\tau_{\mathrm{LO}}$, and we can

obtain from (13.18) to (13.20)

$$a_o = \frac{V_{\mathrm{LO}}}{2} \tag{13.21}$$

$$a_n = 0 \tag{13.22}$$

$$b_n = \begin{cases} 0, & n \text{ even} \\ \dfrac{2V_{\mathrm{LO}}}{n\pi}, & n \text{ odd} \end{cases} \tag{13.23}$$

and hence the (periodic) square-pulse signal can be obtained from (13.17) as

$$V_{\mathrm{LO}}(t) = \frac{V_{\mathrm{LO}}}{2} + \frac{2V_{\mathrm{LO}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n} \sin 2\pi n f_{\mathrm{LO}} t \tag{13.24}$$

When the switch is turned on, signals, including DC and RF signals, would appear at the output port and currents (both DC and RF) would flow through the switch and the load R_L, and when the switch is off, there is no signal or current at the output. Effectively, the RF input signal is transformed into an output signal as the switch is being modulated by the LO signal. Mathematically, the switch performs a multiplication function and, accordingly, the current flowing through the load R_L can be described as

$$I_{\mathrm{IF}}(t) = [I_{\mathrm{DC}} + I_{\mathrm{RF}}(t)]V_{\mathrm{LO}}(t) = [I_{\mathrm{DC}} + I_{\mathrm{RF}}(t)]\left(\frac{V_{\mathrm{LO}}}{2} + \frac{2V_{\mathrm{LO}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n} \sin 2\pi n f_{\mathrm{LO}} t\right) \tag{13.25}$$

where $I_{\mathrm{RF}}(t)$ and I_{DC} are the RF current (from the RF signal) and the DC current (from the DC voltages applied to M1) flow through the switch M2 under its on-state. This current is indeed formed by "mixing." It is noted that the mixer operates based on three inter-related actions (switching, multiplication, and mixing): switching is a physical action, which is mathematically equivalent to multiplication, resulting in mixing which is a physical action. The output voltage can hence be obtained as

$$\begin{aligned} V_{\mathrm{IF}}(t) &= V_{\mathrm{RF}}(t)V_{\mathrm{LO}}(t) = I_{\mathrm{IF}}(t)R_L \\ &= R_L[I_{\mathrm{DC}} + I_{\mathrm{RF}}(t)]\left(\frac{V_{\mathrm{LO}}}{2} + \frac{2V_{\mathrm{LO}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n} \sin 2\pi n f_{\mathrm{LO}} t\right) \end{aligned} \tag{13.26}$$

Equations (13.25) and (13.26) set the basis for analyzing the output current, voltage and their spectrums for different mixer topologies.

13.4.2 Single-Ended Mixer

Figure 13.24 shows a single-ended mixer core used for the analysis. This mixer topology is already described in Section 13.2.1. In this mixer core, the RF input signal is fed to the gate of the lower transistor (M1) and the LO signal, assumed to be a train of perfect square pulses, drives the upper transistor (M2). The mixer can be considered a switching mixer discussed earlier in which transistor M2 acts as a switch, turning on and off in each half cycle of the LO signal, and transistor M1 functions as an amplifier amplifying the RF input signal, which subsequently enters M2 as the input signal. M1 and M2 are operated in the saturation and linear regions, respectively, which are typical operating conditions for amplifiers and switches. To facilitate the modulation of the LO signal, M2 is biased with a gate–source voltage near the transistor's threshold voltage. When the total voltage (biased voltage and LO voltage) at the gate of M2 is sufficiently small and large, the transistor is turned off and on, respectively, which indeed happen alternately during each half cycle

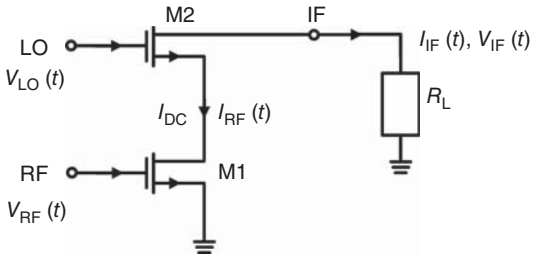


Figure 13.24. Single-ended MOSFET mixer.

of the LO signal. It is noted that for mixers used in a receiver front-end, M1 should act for both low-noise and amplification to achieve an overall low noise figure for mixers. For up-converters, only amplification is needed.

The amplified RF signal feeding the switch M2 is primarily controlled by the transconductance of M1. Its RF output current is approximately given by

$$I_{\mathrm{RF}}(t) = g_m(t)V_{\mathrm{RF}}(t) \tag{13.27}$$

where $I_{\mathrm{RF}}(t)$ is actually the drain–source current of M1, $g_m(t)$ is the transconductance of M1, and $V_{\mathrm{RF}}(t)$ represents the RF input voltage which is the voltage at the gate of M1. For simplicity without loss of generality, we assume that $g_m(t)$ is a constant (g_m). That means that the mixer is operated under small signals; under large signals, $g_m(t)$ is a nonlinear function. Let us also assume that $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t)$; we can then determine the output voltage from (13.26), making use of (13.27) and excluding the supplied DC voltage V_{dd} which will nonetheless be canceled through a blocking capacitor, as

$$V_{\mathrm{IF}}(t) = R_L[I_{\mathrm{DC}} + g_m V_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t)]\left(\frac{V_{\mathrm{LO}}}{2} + \frac{2V_{\mathrm{LO}}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{\mathrm{LO}}t\right) \tag{13.28}$$

where I_{DC} is the DC current, caused by the bias voltages to M1, that goes through M2 when M2 is on. Expanding (13.28) gives

$$\begin{aligned}
V_{\mathrm{IF}}(t) &= \frac{1}{2}V_{\mathrm{LO}}I_{\mathrm{DC}}R_L + \frac{1}{2}g_m R_L V_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + \frac{2R_L V_{\mathrm{LO}}I_{\mathrm{DC}}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{\mathrm{LO}}t \\
&\quad + \frac{2g_m R_L V_{\mathrm{LO}}V_{\mathrm{RF}}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\cos(2\pi f_{\mathrm{RF}}t)\sin 2\pi n f_{\mathrm{LO}}t \\
&= \frac{1}{2}V_{\mathrm{LO}}I_{\mathrm{DC}}R_L + \frac{1}{2}g_m R_L V_{\mathrm{RF}}\cos(2\pi f_{\mathrm{RF}}t) + \frac{2R_L V_{\mathrm{LO}}I_{\mathrm{DC}}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{\mathrm{LO}}t \\
&\quad + \frac{g_m R_L V_{\mathrm{LO}}V_{\mathrm{RF}}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\{\sin[2\pi(nf_{\mathrm{LO}} + f_{\mathrm{RF}})t] + \sin[2\pi(nf_{\mathrm{LO}} - f_{\mathrm{RF}})t]\}
\end{aligned} \tag{13.29}$$

which shows that the output contains signals at DC, f_{RF}, f_{LO} and its odd harmonics, and $(nf_{\mathrm{LO}} \pm f_{\mathrm{RF}})$, where $n = 1, 3, 5, \ldots$ Among those, the signals at $f_{\mathrm{LO}} \pm f_{\mathrm{RF}}$ are normally the interested signals, which $f_{\mathrm{LO}} - f_{\mathrm{RF}}$ and $f_{\mathrm{LO}} + f_{\mathrm{RF}}$ represent the frequencies of the down- and up-converted signals, respectively. It is noted that the foregoing analysis assumes the LO signal is a (simple) perfect square wave. Similar analysis can be performed for other kinds of LO signals such as a sinusoidal LO signal.

As can be recalled, although the analysis is performed through a combined analytical function dictated by (13.26) and hence (13.28), it actually consists of two distinct parts: low-noise amplification through transistor M1 and switching through transistor M2. The low-noise amplification for M1 is needed for mixers used in a receiver front-end. As such, any LNA topologies can be used in lieu of M1. Furthermore, the RF bandwidth is also dictated by the M1 stage and hence of any LNA used in that stage. In the analysis, the switching of M2 is assumed to be perfect. This, however, does not happen in practice. Under on-state, the switch has a nonzero resistance and, under off-state, it has a large but finite resistance. The nonzero on-resistance causes reduction in the conversion gain and increase in the noise figure, while the finite off-resistance reduces the isolation between the RF and IF ports. Transistor M2 thus needs to be configured and/or specially biased to reduce the on-resistance and increase the off-resistance. The RF bandwidth is also affected by the switching performance of M2 over frequencies. Other switching topologies may be used in lieu of M2 to produce a better switching function.

As can be extracted from (13.29), the maximum IF output voltage at $f_{\mathrm{LO}} - f_{\mathrm{RF}}$ is $g_m R_L V_{\mathrm{LO}} V_{\mathrm{RF}}/\pi$. Hence, the voltage conversion gain of the single-ended mixer can be derived as

$$G_{\mathrm{se}} = \frac{g_m R_L V_{\mathrm{LO}} V_{\mathrm{RF}}/\pi}{V_{\mathrm{RF}}} = \frac{V_{\mathrm{LO}}}{\pi} g_m R_L \tag{13.30}$$

Equation (13.30) is of course very approximate and not quantitatively correct, in which it assumes an ideal-behaved amplifier with its gain dictated completely by the transconductance and an ideal switch with perfect on and off states. Practical mixers do not behave this way and hence degrading the performance – for instance, a nonzero on-resistance of the switch (M2) reduces the conversion gain of the mixer. Nevertheless, it can be used to gain insight of the mixer's conversion qualitatively. Equation (13.30) shows that the conversion gain of the mixer is proportional to the transconductance of transistor M1. This is indeed expected as M1 acts as an amplifier in the mixer as described early. The transconductance of MOSFET in saturation, as given in (9.9), is proportional to the ratio of the width and length (W/L) of the transistor and, hence, can be increased by increasing W/L. Increasing W/L, however, also increases the transistor's parasitic capacitances, thereby degrading its performance. The conversion gain, as can be seen from (13.30), can also be increased by using a large load resistance and large LO signal. Using a large but finite load resistance, however, is possible only when a mixer is used in an integrated-circuit environment with the circuit following it has a large input impedance. In this case, a buffer is needed for measuring the mixer. When used as a standalone component, the typical load impedance is 50 Ω. It is noted that the switching transistor M2 is operated in the linear region which is confined with small LO voltage, and so (13.30) is valid only under small LO signal, implying that the conversion gain of a mixer increases only linearly with the LO amplitude under small LO levels; beyond a certain LO level, the conversion gain does not increase with the LO amplitude; instead, it increases to a certain level and then reduces as the LO amplitude is increased.

As any mixer, regardless of topologies, can be considered a switching mixer functioning based on two important constituent parts: amplification stage (amplifier) and switching stage (switch), we can make a conclusion that, in order to achieve low noise figure and high conversion gain across a interested frequency range for a mixer, proper amplifying and switching stages need to be carefully derived and designed. These may include a single transistor, like M1 and M2 in Figure 13.24, or multiple transistors interconnected in a particular fashion. Moreover, as discussed in Section 13.1.1 and shown again in this section for a simple square-wave LO, spurious signals at the output of mixers and the isolation between ports are important design consideration. Although filters can be included in a mixer circuit to suppress unwanted signals, such as the RF signal, the LO signal and its odd harmonics, and others at ($nf_{\mathrm{LO}} \pm f_{\mathrm{RF}}$) except $f_{\mathrm{LO}} - f_{\mathrm{RF}}$ or $f_{\mathrm{LO}} + f_{\mathrm{RF}}$, where $n = 1, 3, 5, \ldots$, at the IF port of the single-ended mixer discussed earlier, the use of filters increases the loss and size of the mixer. This increase can be significant due to the inductors needed for filters. In view of these (noise figure, gain, spurs, isolation), proper topologies for mixers should be and, in fact, can be developed to achieve a particular performance without possible use (or with a minimum number) of filters. The idea is achieving improved performance without increasing much more complexity and size to circuits. The single-balanced

and double-balanced mixers to be described later can be viewed as revolving around implementing better amplification and switching stages and hence better mixers.

13.4.3 Single-Balanced Mixer

As stated earlier in Section 13.2.2, a single-balanced mixer can be considered a combination of two single-ended mixers. Figure 13.25 shows such a single-balance mixer core integrating two single-ended mixers shown in Figure 13.24 sharing a common transistor (M1). This mixer is also shown in Figure 13.9(b). To simplify the analysis and discussion without loss of generality, we assume that transistors M2+ and M2− are the same. As for the single-ended mixer, the common transistor M1 acts as an amplifier for the RF signal to provide amplified RF signals to the upper transistors M2+ and M2− which function as switches modulated by the LO signals. The LO signals, namely LO+ and LO−, drive transistors M2+ and M2− with equal amplitude but 180° out-of-phase – that is, providing a differential signal to these transistors. The differential LO signals may be obtained directly from a differential LO source or obtained from a single-ended LO through an LO balun. It is noted that the total LO power provided to both M2+ and M2− is twice of that reaching M2+ or M2− or twice of that used in the single-ended mixer. In operation, the RF signal arrives at the switching or mixing transistors M2+ and M2− in phase while the LO signal arrives out-of-phase, hence producing two output signals at ports IF+ and IF− with equal amplitude and 180° out-of-phase. These two output signals can be combined via a balun to provide a single-ended output signal or used directly as a differential output signal for a subsequent component. Since the LO signals LO+ and LO− have opposite phases, they switch the respective transistors M2+ and M2− on and off alternately in each LO period. When M2+ is on in one half LO cycle, M2− is off and when M2+ is off in the next half LO cycle, M2− is on. As such, in any half LO cycle (i.e., at any time), mixing always occurs, which is different from the single-ended mixer which only conducts mixing in every other half cycle. Therefore, it is expected that the output signal would be twice as much as that for the single-ended mixer if the RF signal arriving at the source of each mixing transistor (M2+ or M2−) is the same as that feeding the source of the mixing transistor M2 of the single-ended mixer in Figure 13.24. The design of the single-balanced mixer can be based directly on the design of its single-ended counterpart with possible addition of LO and IF baluns.

We assume that transistors M2+ and M2− are operated in the linear region, transistor M1 is operated in the saturation region, the LO signal is a perfect square wave given in (13.24), and the RF signal entering the RF port is a small signal described as $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}} \cos(2\pi f_{\mathrm{RF}} t)$. Assuming the RF signal amplified by transistor M1 is divided and enters equally to transistors M2+ and M2−, the input signal to M2+ or M2− can be written as

$$V_{\mathrm{RF,M1}}(t) = \frac{g_m V_{\mathrm{RF}}}{2} \cos(2\pi f_{\mathrm{RF}} t) \tag{13.31}$$

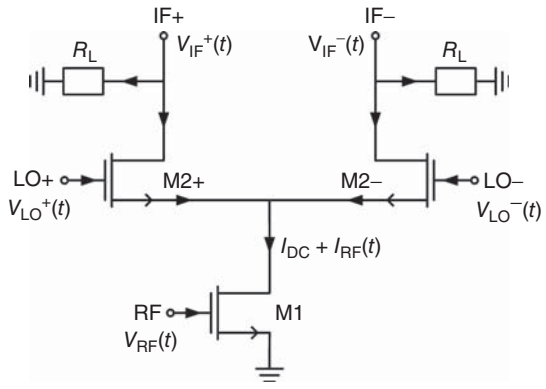


Figure 13.25. Single-balance MOSFET mixer.

where g_m is the tranconductance of M2+ and M2−. We can write the output voltages at port IF+ and IF−, making use of (13.29), (13.31) and excluding the supplied DC voltage V_{dd}, as

$$V_{IF}^{+}(t) = \frac{1}{2}V_{LO}I_{DC}R_L + \frac{1}{4}g_m R_L V_{RF}\cos(2\pi f_{RF}t) + \frac{2R_L V_{LO}I_{DC}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{LO}t$$
$$+ \frac{g_m R_L V_{LO}V_{RF}}{2\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\{\sin[2\pi(nf_{LO}+f_{RF})t] + \sin[2\pi(nf_{LO}-f_{RF})t]\} \tag{13.32}$$

and

$$V_{IF}^{-}(t) = \frac{1}{2}V_{LO}I_{DC}R_L + \frac{1}{4}g_m R_L V_{RF}\cos(2\pi f_{RF}t) - \frac{2R_L V_{LO}I_{DC}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{LO}t$$
$$- \frac{g_m R_L V_{LO}V_{RF}}{2\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\{\sin[2\pi(nf_{LO}+f_{RF})t] + \sin[2\pi(nf_{LO}-f_{RF})t]\} \tag{13.33}$$

recognizing that $V_{LO}^{-}(t) = V_{LO}^{+}(t)e^{j\pi} = -V_{LO}^{+}(t)$ and $V_{LO}^{+} = V_{LO}^{+} \equiv V_{LO}$. The total output voltage, or single-ended voltage, can then be obtained by subtracting $V_{IF}^{-}(t)$ from $V_{IF}^{+}(t)$ as

$$V_{IF}(t) = V_{IF}^{+}(t) - V_{IF}^{-}(t)$$
$$= \frac{4R_L V_{LO}I_{DC}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\sin 2\pi n f_{LO}t$$
$$+ \frac{g_m R_L V_{LO}V_{RF}}{\pi}\sum_{n=1,3,5,\ldots}^{\infty}\frac{1}{n}\{\sin[2\pi(nf_{LO}+f_{RF})t] + \sin[2\pi(nf_{LO}-f_{RF})t]\} \tag{13.34}$$

It is noted again that V_{LO} in (13.34) is the same as V_{LO} for a single-ended mixer; that means that the total LO voltage driving the single-balanced mixer is 1.414 times of the LO voltage required for the single-ended mixer, or the LO power is twice as much. The output thus contains signals at f_{LO} and its odd harmonics, and $(nf_{LO} \pm f_{RF})$, where $n = 1, 3, 5, \ldots$ The desired signal is either at $f_{LO} - f_{RF}$ (for down-conversion) or at $f_{LO} + f_{RF}$ (for up-conversion), while all others are undesired. It is noted that, contrary to the single-ended mixer, the output of a single-balanced mixer has no DC and RF components, which are expected considering the facts that the RF and DC signals arrive at the mixing transistors M2+ and M2− in-phase and, hence, are canceled out at the subtracting point (single-ended output port). This advantage results through exploitation of the unique property of differential ports, which reject any common signals (herein RF and DC) to them, which is discussed in the differential amplifiers in Chapter 11. It is noted that the analysis requires a perfect symmetry of the two switching stages (M2+ and M2−) with respect to the amplifying stage (M1), and so not only electrical design but also physical layout need to be carefully done to ensure symmetry. Otherwise, degradation in performance will result. Moreover, switch M2− should be completely off when M2+ is on and vice versa for optimum performance.

The output voltage amplitude at $f_{LO} - f_{RF}$ can be obtained from (13.34) as $g_m R_L V_{LO} V_{RF}/\pi$. Hence, the voltage conversion gain of the single-balanced mixer can be determined to be

$$G_{sb} = \frac{g_m R_L V_{LO}V_{RF}/\pi}{V_{RF}} = \frac{V_{LO}}{\pi}g_m R_L = G_{se} \tag{13.35}$$

which is equal to the conversion gain of the single-ended mixer. This is expected since, although the IF output voltages are combined, the same RF signal as for the single-ended mixer is shared between the two

single-ended mixers making up the single-balance mixer. Nevertheless, as for the single-ended mixer's conversion gain given in (13.30), (13.35) should be used only for analyzing and estimating the conversion gain qualitatively.

As discussed in the preceding section, better mixer performance can be achieved by employing better topologies for the mixing and amplifying stages. This concept for improving mixer performance has been demonstrated through a single-balanced mixer topology that uses a balanced switching stage, which dictates the need of a balanced (differential) controlled LO signal, instead of an unbalanced switching stage as for the single-ended mixer.

13.4.4 Double-Balanced Mixer

Figure 13.26 shows a double-balanced mixer core, described in Figure 13.11, which consists of two single-balanced mixers shown in Figure 13.25 or four single-ended mixers shown in Figure 13.24. The drains of M2a+ and M2b+ are connected to become the IF+ port, and the drains of M2a− and M2b− are tied together to make the IF− port. The single-balanced mixer 1 consists of M1+, M2a+, and M2a− and the single-balanced mixer 2 consists of M1−, M2b−, and M2b+. For simple illustration, we assume transistors M2a+, M2a−, M2b+, and M2b− are identical. The operation of the double-balanced mixer is essentially based on that of the single-balanced mixer. Transistors M1+ and M1− are used as amplifiers (or specifically LNA for down-conversion in a receiver front-end), while M2a+, M2a−, M2b+, and M2b− work as switches to provide the required mixing action. The RF signal is fed into M1+ and M1− with equal amplitude but 180° out-of-phase. The LO signal drives M2a + /M2b+ and M2a − /M2b− in phase and 180° out-of-phase, respectively, and consequently switching M2a + /M2b+ on (off) and M2a − /M2b− off (on) alternately in every half-period. As a result, the output signals coming from M2a+ and M2b− have the same amplitude and phase and hence are combined constructively at the IF+ port. Similarly, the output signals coming from M2a− and M2b+ have the same amplitude and phase and are added at the IF− port. The IF+ and IF− signals have equal amplitude with 180° out-of-phase. At any instant, it is recognized that there is always conduction in both the IF+ and IF− paths, which is different from the single-balanced mixer in which conduction happens only in the IF+ or IF− path. Accordingly, we can expect that the output signal of the double-balanced mixer would be twice of that for the single-balanced mixer provided that the output RF power of M1+ or M1− is equal to that produced by M1 of the single-balanced mixer in Figure 13.25.

We assume that transistors M2a+, M2a−, M2b+, and M2b− are operated in the linear region, transistors M1+ and M1− are operated in the saturation region, the LO signal is a perfect square wave given in (13.24), and the RF signal entering the single-ended RF port is a small signal described as $V_{RF}(t) = V_{RF}\cos(2\pi f_{RF}t)$.

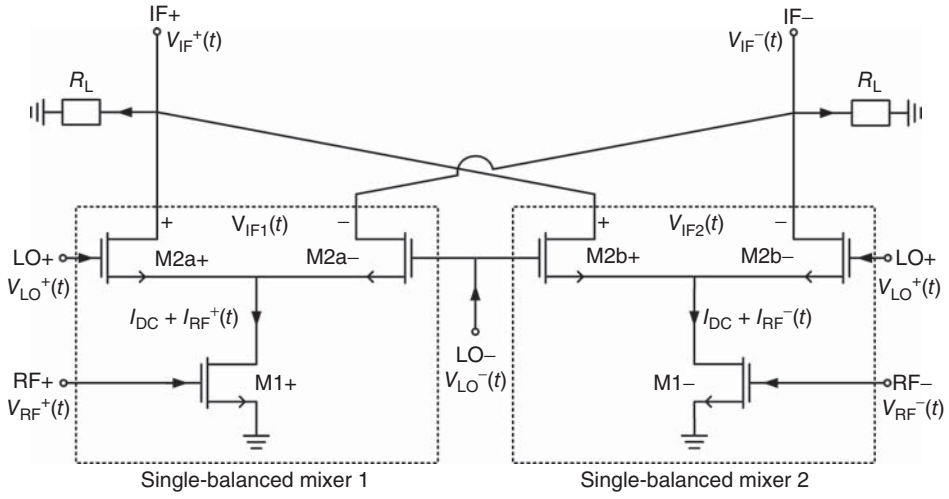


Figure 13.26. Double-balance MOSFET mixer.

We also assume that the (single-ended) RF signal splits equally with opposite phase [$V^{+}_{\mathrm{RF}}(t)$ and $V^{+}_{\mathrm{RF}}(t)$] to M1+ and M1−, respectively. These signals, after amplified by M1+ and M1−, are divided equally to M2a+, M2a− and M2b+, M2b−. The input signal to each of M2a+, M2b+, M2a− and M2b− can thus be written as

$$V^{+}_{\mathrm{RF,M1+}}(t) = -V^{-}_{\mathrm{RF,M1-}}(t) = \frac{g_m V_{\mathrm{RF}}}{4}\cos(2\pi f_{\mathrm{RF}} t) \tag{13.36}$$

where g_m is the tranconductance of M2a+, M2a−, M2b+, and M2b−. Note that $I^{-}_{\mathrm{RF}}(t) = -I^{+}_{\mathrm{RF}}(t)$. The output voltage produced by the single-balance mixer 1 is exactly the same as that of the single-balanced mixer in Figure 13.25 with half of the RF voltage amplitude according to (13.36) and, hence, can be obtained from (13.34) as

$$\begin{aligned} V_{\mathrm{IF1}}(t) &= \frac{4R_L V_{\mathrm{LO}} I_{\mathrm{DC}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n}\sin 2\pi n f_{\mathrm{LO}} t \\ &\quad + \frac{g_m R_L V_{\mathrm{LO}} V_{\mathrm{RF}}}{2\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n}\{\sin[2\pi(nf_{\mathrm{LO}} + f_{\mathrm{RF}})t] + \sin[2\pi(nf_{\mathrm{LO}} - f_{\mathrm{RF}})t]\} \end{aligned} \tag{13.37}$$

It is noted that V_{LO} in (13.37) is the same as V_{LO} for a single-ended or single-balanced mixer; that means that the total LO voltage driving the double-balanced mixer is twice or 1.414 of the LO voltage required for the single-ended or single-balanced mixer, respectively. Now comparing between the single-balanced mixers 1 and 2, the LO signal going to M2b − (M2b+) has opposite phase to that of M2a + (M2a−), and the RF signals arriving at M1+ and M1− are also 180° out-of-phase. In consequence, the output voltage produced the single-balanced mixer 2 can be derived from (13.34) but with the LO and RF signals off-set by 180° (i.e., $V_{\mathrm{LO}} \rightarrow -V_{\mathrm{LO}}$ and $V_{\mathrm{RF}} \rightarrow -V_{\mathrm{RF}}$) as

$$\begin{aligned} V_{\mathrm{IF2}}(t) &= -\frac{4R_L V_{\mathrm{LO}} I_{\mathrm{DC}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n}\sin 2\pi n f_{\mathrm{LO}} t \\ &\quad + \frac{g_m R_L V_{\mathrm{LO}} V_{\mathrm{RF}}}{2\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n}\{\sin[2\pi(nf_{\mathrm{LO}} + f_{\mathrm{RF}})t] + \sin[2\pi(nf_{\mathrm{LO}} - f_{\mathrm{RF}})t]\} \end{aligned} \tag{13.38}$$

The total output voltage of the double-balanced mixer is obtained as the summation of those produced by the constituent single-balanced mixers as

$$\begin{aligned} V_{\mathrm{IF}}(t) &= V_{\mathrm{IF1}}(t) + V_{\mathrm{IF2}}(t) \\ &= \frac{g_m R_L V_{\mathrm{LO}} V_{\mathrm{RF}}}{\pi} \sum_{n=1,3,5,\ldots}^{\infty} \frac{1}{n}\{\sin[2\pi(nf_{\mathrm{LO}} + f_{\mathrm{RF}})t] + \sin[2\pi(nf_{\mathrm{LO}} - f_{\mathrm{RF}})t]\} \end{aligned} \tag{13.39}$$

We can now see that the output contains only signals at $(nf_{\mathrm{LO}} \pm f_{\mathrm{RF}})$, where $n = 1, 3, 5, \ldots$, whose desired component is either at $f_{\mathrm{LO}} - f_{\mathrm{RF}}$ (for down-conversion) or at $f_{\mathrm{LO}} + f_{\mathrm{RF}}$ (for up-conversion). All other signals including DC, RF, LO, and LO harmonics are completely suppressed. The LO and LO odd-harmonics existing at each of the differential IF ports are in-phase and hence canceled completely when subtracting each other at the IF port. The RF (and DC) signals emerging at the output of M2a + (M2a−) and M2b − (M2b+) are 180° out-of-phase, thereby canceling each other at the combining point IF + (IF−) and subsequently at the subtracting point between IF+ and IF−. Therefore, the LO − IF and RF − IF isolation is inherent. It is noted that the double-balanced mixer employs differential ports for RF, LO, and IF, hence also producing good isolation among other ports. Again, all these signal suppressions and isolation between ports are achieved through proper use of differential signals or decomposing single-ended signals into differential signals. Similar to the single-balanced mixer, a perfect symmetry of the switching stages with respect to the amplifying stages

and the LO port, and so not only electrical design but also physical layout need to be carefully done to ensure symmetry. Otherwise, degradation in performance will result. Moreover, switching pair M2a − /M2b+ should be completely off when M2a + /M2b− is on and vice versa for optimum performance.

The output voltage amplitude at $f_{\text{LO}} - f_{\text{RF}}$ can be obtained from (13.39) as $g_m R_L V_{\text{LO}} V_{\text{RF}}/\pi$. Hence the voltage conversion gain of the double-balanced mixer can be derived as

$$G_{\text{db}} = \frac{g_m R_L V_{\text{LO}} V_{\text{RF}}/\pi}{V_{\text{RF}}} = \frac{V_{\text{LO}}}{\pi} g_m R_L = G_{\text{sb}} = G_{\text{se}} \tag{13.40}$$

which shows that the conversion gains for the double-balanced, single-balanced, and single-ended mixers are equal. This is expected since, although the IF output voltages are combined, the same RF signal as for the single-ended mixer is shared between the four single-ended mixers making up the double-balance mixer. It is noted that (13.30) is derived assuming ideal conditions and hence should be used only for determining the conversion gain qualitatively.

Again, by implementing a better-configured mixer topology with a balanced amplifying stage and double-balanced switching stage, better mixer performance can be achieved. This, however, complicates the mixer design such as requiring balanced (differential) LO and RF signals. The concept can be extended to include a triple-balanced switching stage to further improve the mixer performance. Additionally, a better switching topology and/or amplifying topology, instead of a single MOSFET used in the foregoing discussed single-ended, single-balanced, and double-balanced mixers, can be employed to further improve the noise figure, gain, and mixing performance.

Among the mixer topologies, the double-balanced mixer as shown in Figure 13.26 is perhaps most widely used. Its performance is superior as compared to the single-ended and single-balanced mixer while much less complex than other high-level mixers such as the doubly double-balanced mixer.

Appendix 2 described a detailed design of a CMOS RFIC double-balanced mixer. This design also serves as an example to illustrate a design procedure for other CMOS RFICs as well.

13.4.5 Source Degeneration in Mixer Design

All the foregoing discussed mixers assume the transconductance of the MOSFETs used as the amplifying (or amplifying plus low-noise) stage is constant and the transistors are connected to ground. A via-hole is typically used to connect the source to the ground, which inadvertently introduces extra inductance and resistance, hence degrading the overall gain. Under regular operations, the transistor's transconductance is not constant – in fact, it is a nonlinear function of time – due to possible large voltage swings of applied signals, such as that from the LO signal, causing changes in the transistor's biasing and hence its operation and behavior – for instance, nonlinearity and variation in the transconductance. A nonconstant transconductance causes undesired fluctuation of the output signal and the mixer's conversion gain and the nonlinearity generates more spurs and degrades the mixer's linearity. Source degeneration, implemented for source-degenerated amplifiers in Section 11.6.4, can be employed in the amplifying stage for mixers to minimize the change of the transconductance and, with proper selection of the source-degeneration element, can remove the dependence of the mixer on the transistor's transconductance. Using the source-degeneration technique, an impedance is inserted between the amplifying MOSFETs (M1 in Figures 13.24 and 13.25 and M1+, M1− in Figure 13.26) and ground as shown in Figure 13.27. Typically, either an inductor (inductive source degeneration) or a resistor (resistive source degeneration) is used for the degenerative impedance as shown in Figure 13.27(b) and (c), respectively. The degenerative impedance can be viewed as part of the transistor and, as a result, can alter the performance such as gain of the transistor. With proper selection of the degenerative inductance, it can, together with the extra inductance caused by a via-hole, can improve the gain of the transistor as compared to that without a degenerative inductor.

The RF output current with a source-degeneration impedance Z_S can be approximately derived as

$$I_{\text{RF}}(t) = \frac{g_m(t)}{1 + Z_S[g_m(t) + j\omega C_{\text{gs}}]} V_{\text{RF}}(t) \tag{13.41}$$

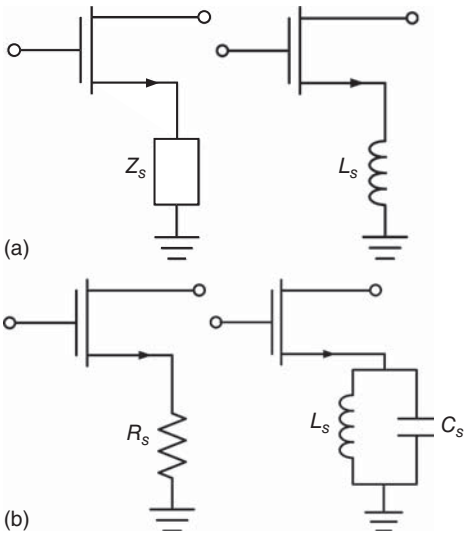


Figure 13.27. MOSFET with source degeneration: impedance source degeneration (a), inductive source degeneration (b), and resistive source degeneration (c).

where $g_m(t)$ and C_{gs} are the transconductance and gate–source capacitance of the transistor and $V_{\text{RF}}(t)$ is the RF input voltage at the gate of the transistor. Comparing (13.41) to (13.27), we can then define an "effective transconductance" or "overall transconductance" for the transistor as

$$g_{m,\text{eff}}(t) = \frac{g_m(t)}{1 + Z_S[g_m(t) + j\omega C_{\text{gs}}]} \tag{13.42}$$

It should be noted that the effective transconductance is simply a mathematically defined parameter and happens only in theory – it is not real. It serves to illustrate some particular points and helps explain certain results. The effective transconductance combines several parameters together in order to simplify the formulation, just like we use the "effective dielectric constant" in Chapter 2 to lump the dielectric constant and conductivity of materials together for characterizing materials. The effective transconductance enables us to use all the equations that exist or are derived involving the transconductance by simply replacing $g_m(t)$ with $g_{m,\text{eff}}(t)$. The appearance of $g_m(t)$ on both the numerator and denominator of the effective transconductance in the effective transconductance signifies the less dependence on the variation of $g_m(t)$ for the source-degenerated mixer. For the inductive source degeneration, the effective transconductance is

$$g_{m,\text{eff}}(t) = \frac{g_m(t)}{1 - \omega^2(L_S + L_v)C_{\text{gs}} + j\omega g_m(t)(L_S + L_v)} \tag{13.43}$$

where L_v is the via-hole's inductance (the resistance of the via-hole is typically small in submicron CMOS processes and is assumed to be zero). The input impedance looking into the transistor's, which is the input impedance at the RF port, can be obtained from Eq. (11.376) as

$$Z_{\text{RF,in}} = \frac{g_m(t)(L_S + L_v)}{C_{\text{gs}}} + j\left[\omega\left(L_S + L_v\right) - \frac{1}{\omega C_{\text{gs}}}\right] \tag{13.44}$$

which shows that this impedance can be modeled as a series RLC network where $R = g_m(L_S + L_v)/C_{\text{gs}}$, $L = L_S + L_v$, and $C = C_{\text{gs}}$. At the resonant frequency where $j\omega(L_S + L_v) + 1/j\omega C_{\text{gs}} = 0$, the effective transconductance and RF input impedance in (13.43) and (13.44) become

$$g_{m,\text{eff}} = -\frac{j}{\omega(L_S + L_v)} = \frac{1}{\omega(L_S + L_v)}\angle 270° \tag{13.45}$$

and

$$Z_{\mathrm{RF},in} = \frac{g_m(t)(L_S + L_v)}{C_{\mathrm{gs}}} \tag{13.46}$$

respectively. We can now see that the overall transconductance of the transistor is theoretically independent with the actual transconductance of the transistor and is constant at a particular frequency. The overall transconductance of the amplifying transistor thus decreases the frequency is increased, hence reducing the mixer's conversion gain. However, at the same time, the RF input impedance becomes purely resistive and can be easily designed to be equal to the RF source impedance over an interested frequency range. It is particularly noted that the inductive source degeneration effectively removes the conversion-gain dependence on the transconductance of the transistors and allows the inductor to play an important role in the conversion. This, however, is possible only in theory under ideal conditions. In reality, the conversion depends on both the inductor and transconductance.

For the resistive source degeneration, the effective transconductance can be determined as

$$g_{m,\mathrm{eff}}(t) = \frac{g_m(t)}{1 + g_m(t)R_S + j\omega R_S C_{\mathrm{gs}}} \tag{13.47}$$

If the degeneration resistance and transistor were selected so that $1/g_m \ll R_S \ll 1/\omega C_{\mathrm{gs}}$ at the operating frequency, then the effective transconductance becomes

$$g_{m,\mathrm{eff}} = \frac{1}{R_S} \tag{13.48}$$

which is indeed a constant. This effective transconductance can therefore be increased by using small values for R_S (while still keeping $R_S \gg 1/g_m$), hence resulting in increased gain. It is noted that using resistive source degeneration would reduce the conversion gain due to the voltage drop across the resistor. The resistive degeneration also increases noise due to an additional noise source.

For an LC source degeneration as shown in Figure 13.27(d), the effective transconductance can be derived as

$$g_{m,\mathrm{eff}}(t) = \frac{g_m(t)(1 - \omega^2 L_s C_s)}{1 - \omega^2 L_s(C_s + C_{\mathrm{gs}}) + j\omega L_s g_m(t)} \tag{13.49}$$

which, again, shows the less dependence on the variation of $g_m(t)$. If the inductance L_s was chosen so that it resonates the capacitance $(C_s - C_{\mathrm{gs}})$ at the design frequency (i.e., $L_s = 1/\omega^2(C_s - C_{\mathrm{gs}})$), then the effective transconductance becomes

$$g_{m,\mathrm{eff}}(t) = \frac{\omega C_{\mathrm{gs}} g_m(t)}{2\omega C_{\mathrm{gs}} - j g_m(t)} \tag{13.50}$$

which, although is not a constant, reduces the effect of the transistor's varying transconductance.

13.5 SAMPLING MIXER

Mixers that we have discussed so far are generally based on the signal-mixing/switching concept. There is another class of mixers that functions based on signal-sampling. These mixers, commonly known as "sampling mixers" or simply "samplers," also find applications in wireless communications and sensing. They are particularly useful for recovering periodic wideband sinusoidal and nonsinusoidal signals, such as pulses, making them a popular receiver front-end structure in pulse-based radars and sensors. Sampling mixers can down-convert a wideband input signal and retain the down-converted signal waveform in the same form as the RF input signal. With the integration of impulse generators, sampling mixers are fully integrated, making them low-cost and small-size. Fully-integrated sampling mixers can conduct sampling over wide bandwidths and can compete with or even outperform diode-based sampling counterparts.

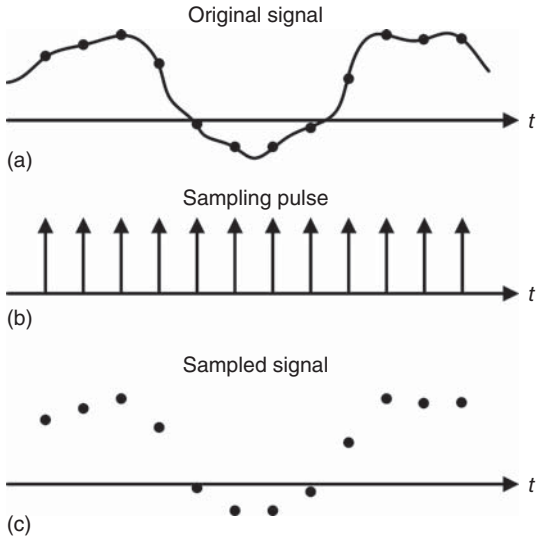


Figure 13.28. General sampling process: (a) original signal, (b) sampling impulse/gate, and (c) sampled signal.

13.5.1 Fundamentals of Sampling

Systems can extract information contained in signals that they receive by sampling these signals. The fundamental principle of sampling is the repeated capturing of subsequent points of a time-varying waveform by a sampling gate. The gate is open and closed by a narrow pulse, which is triggered repeatedly. Sampling techniques have long been used to capture periodic waveforms over wide bandwidths in oscilloscopes. Figure 13.28 shows a general sampling process based on sampling gates controlled by impulses, which recovers an original signal.

Sampling is basically a modulation scheme in which the received signal's information (such as amplitude and hence amplitude modulation) is collected at different instants. There are two kinds of sampling as illustrated in Figure 13.29. One is sampling the (received) RF signal directly to reconstruct the RF signal

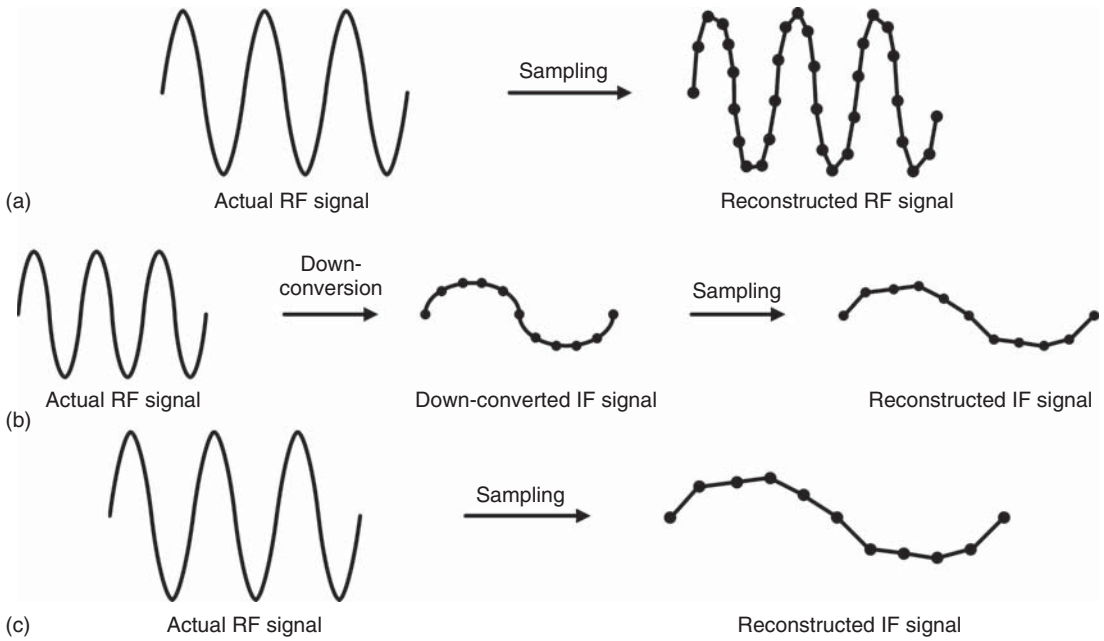


Figure 13.29. Sampling types: (a) direct sampling of RF signal, (b) sampling of down-converted IF signal, and (c) sampling of RF signal to produce IF signal directly.

(Figure 13.29(a)). This approach is convenient but very difficult, if not impossible, to implement at very high frequencies. Another one, which is typically employed, is to sample the IF signal down-converted from the RF signal to produce a reconstructed down-converted signal (Figure 13.29(b)). The down-converted IF signal can be obtained by a conventional mixer such as those described previously. Another approach is employing a "sampling mixer" to sample the actual received RF signal at different instants using a "sampling pulse" or "sampling strobe." The sampling-mixer process converts an actual RF signal into an IF signal at low frequency (typically baseband) by repetitive sampling (Figure 13.29(c)). This method is particularly useful for pulse-based RF systems such as impulse radar.

Sampling can be done using various techniques including real-time sampling, (repetitive) random sampling, and (repetitive) sequential sampling. The real-time sampling technique is the only choice for single-shot waveform and must take all the samples in one occurrence of the waveform. This technique is applicable only to direct sampling of actual received RF signals and direct sampling of down-converted IF signals. The random-sampling approach is used for repetitive signals only and it randomly samples different points on different occurrence of waveform (i.e., on different cycles.) The sequential-sampling method is also used only for repetitive signals and it samples different points on different occurrence of waveform at predetermined times.

13.5.2 Sampling Theory

A signal can be represented by sample values taken at times spaced appropriately. The relation between a sampled signal $f_s(t)$ (i.e., the output IF signal in mixers), an actual signal $f(t)$ (i.e., the input RF signal in mixers), and a sampling signal $s(t)$ (i.e., the LO signal doing the sampling – e.g., a sampling pulse) can be described as

$$f_s(t) = f(t)s(t) \tag{13.51}$$

An ideal sampled signal can be represented by

$$f_s(t) = \sum_{n=-\infty}^{\infty} f(nT_s)\delta(t - nT_s) \tag{13.52}$$

where T_s is the sampling interval defined as the (time) spacing between samples and $\delta(t)$ is a unit impulse function defined as

$$\delta(t - t_o) = \begin{cases} 1 & t = t_o \\ 0 & t \neq t_o \end{cases} \tag{13.53}$$

According to the Nyquist sampling theorem, a signal is completely described by sample values uniformly spaced in sampling time interval $T_s < 1/2f_m$, where f_m is the highest frequency of the signal being sampled, from which the sampling rate or sampling frequency $f_s > 2f_m$. The frequency $f_N = 2f_m$ is the Nyquist frequency and correspondingly $T_N = 1/2f_m$ is the Nyquist sampling interval or period. A signal can therefore be fully reconstructed by sampling the signal at a frequency more than twice of the highest frequency of the signal being sampled. For instance, a signal that has a highest frequency of 2 GHz must be sampled at a rate of at least 4×10^9 times per second in order to recover all the information.

Equation (13.52) can be rewritten as

$$f_s(t) = f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \tag{13.54}$$

recognizing

$$\delta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases}$$

Taking the Fourier transform of (13.54) gives

$$F_s(f) = F(f) \times \left[f_s \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] \tag{13.55}$$

which, after making use of

$$F(f) \times \delta(f - nf_s) = \int_{-\infty}^{\infty} F(x)\delta(f - x - nf_s) = F(f - nf_s) \tag{13.56}$$

becomes

$$F_s(f) = f_s \sum_{n=-\infty}^{\infty} F(f - nf_s) \tag{13.57}$$

Assuming $f(t)$ is band-limited with bandwidth B and the sampling rate $f_s > 2B$, $F_s(f)$ can be deduced from $F(f)$. Figure 13.30 shows an example of constructing $F_s(f)$ from $F(f)$.

Sampling produces a periodic repetition of the signal $f(t)$ being sampled in the frequency domain with the repetition period $1/f_s$. If the sampling frequency $f_s < 2B$, there will be an "overlap" between components of the reproduced $F_s(f)$, and hence $f(t)$ cannot be recovered from $f_s(t)$ without distortion. If $f_s > 2B$, the lowest component of $F_s(f)$ (i.e., $n = 0$) can be completely extracted using an ideal LPF. The signal can be completely reconstructed from the sampled waveform by passing it through an ideal LPF whose bandwidth B satisfies $f_m < B < f_s - f_m$ and $B < f_s/2$. If $f(t)$ is not band-limited or $f_s < 2f_m$, then there is distortion in the reconstructed waveform. This distortion is called "aliasing." For continuous wave (CW) signals, f_m can readily be obtained while, for pulsed signal, f_m is not very clear and must be determined from the signal spectrum. Typically, f_m is determined by a desired bandwidth. Figure 13.31 illustrates how f_m may be determined for a monocycle pulse signal. In this example, the optimum f_m is at the first-zero cross-over (f_{m1}). Under this choice, if $f_s > 2f_{m1}$ then the signal can be recovered theoretically almost without distortion. It is noted that the spectrum still contains a very small portion of energy beyond f_{m1}. Now if $f_m = f_{m2}$ is chosen based on the 4-dB bandwidth as shown, then the signal can also be recovered but with distortion. Note that $f_o \simeq 1/\tau$ and $B_{4\,\mathrm{dB}} \simeq 1/\tau$. For instance, for a 0.33-ns monocycle pulse, its (almost complete) spectrum would go from 1.5 to 4.5 GHz and hence f_m can be selected as 4.5 GHz, from which the sampling frequency must be $f_s > 9$ GHz. It is important, for design purposes, to note that the sampling theory applies to sampling mixers and hence sampling receivers.

13.5.3 Sampling Process

Direct sampling of RF signals using a sampling receiver requires a sampling frequency $f_s > 2f_m$ according to Nyquist theorem, which implies high sampling rate. For high RF, this is very difficult if not feasible. One common solution is to use an "heterodyne sampling receiver" as shown in Figure 13.32 to down-convert an RF signal into a low-frequency IF signal, which is then converted into a digital signal by an analog-to-digital

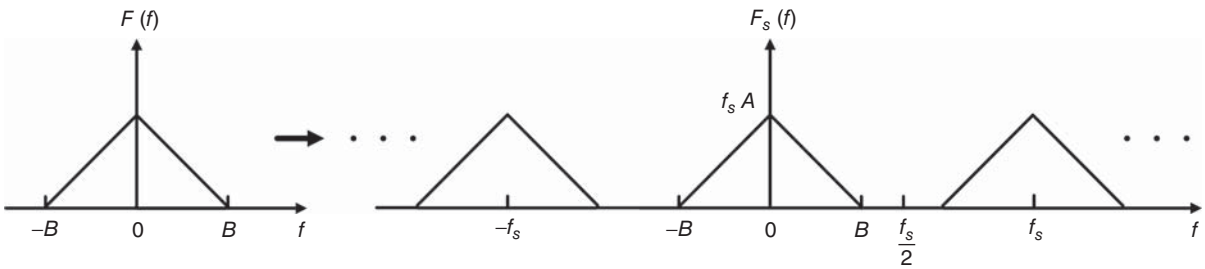


Figure 13.30. Illustration of obtaining $F_s(f)$ from $F(f)$.

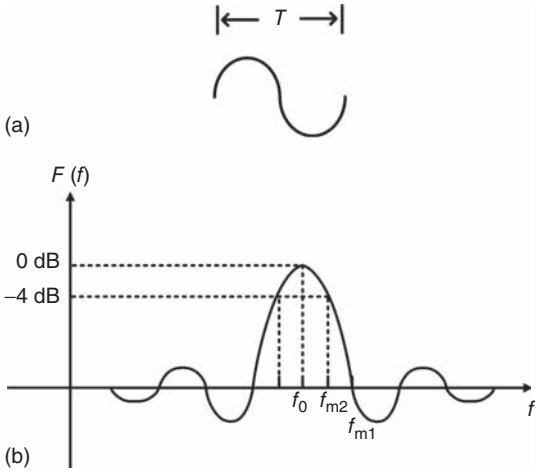


Figure 13.31. Determination of f_m for a monocycle pulse (a) based on its spectrum (b).

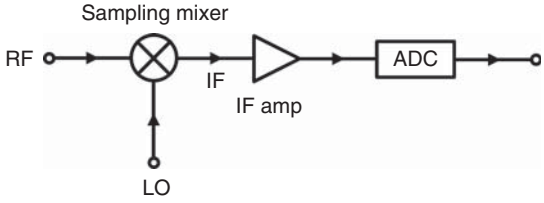


Figure 13.32. Heterodyne sampling receiver.

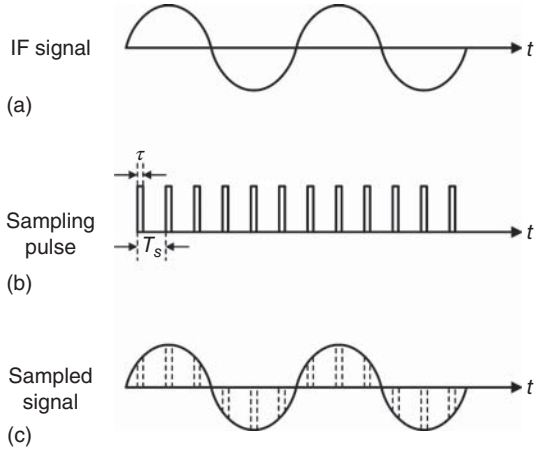


Figure 13.33. Sampling of down-converted IF signal: (a) IF signal, (b) sampling pulse, and (c) sampled signal.

converter (ADC) as shown in Figure 13.33. Sequential sampling is normally used with the ADC. Here, the sampling theory and sampling process apply directly with $f_s > 2f_m$ at very low frequencies. Another approach that is less common but provides a lower cost solution is to convert an RF signal directly to an IF signal, typically in baseband, using a sampling mixer. In this case, the sampling theory and sampling process still apply but is less readily seen. Figure 13.34 illustrates this sampling down-conversion process. In Figure 13.34, $V_{\mathrm{RF}}(t)$ is the waveform of the received input RF signal, $V_{\mathrm{LO}}(t)$ is the LO sampling pulse or strobe signal that triggers the sampling, and $V_{\mathrm{IF}}(t)$ is the output down-converted signal obtained through the sample-and-hold operation of the sampling mixer.

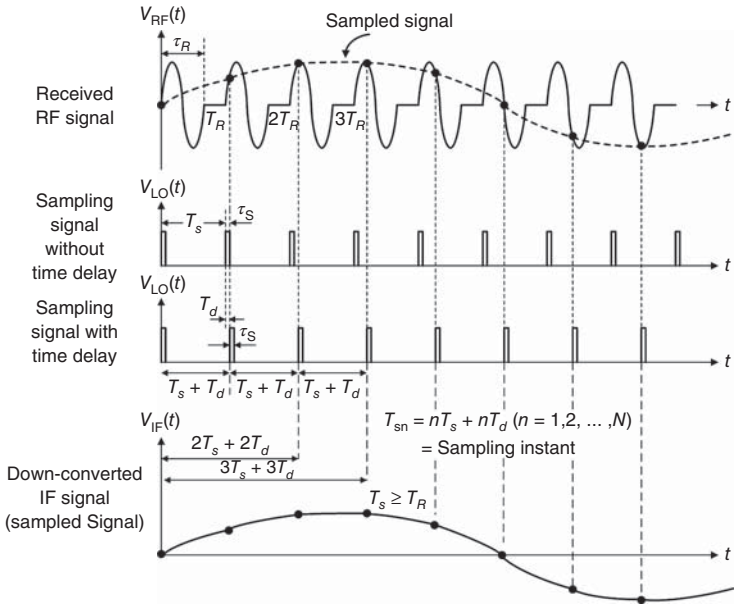


Figure 13.34. Illustration of the sampling down-conversion process. The sampling signal with time delay is the actual sampling signal. We assume $T_d = \tau_s$, but T_d can be different from τ_s. Dots on V_{RF} waveform represent sampled points which have constant time delay T_d from T_s.

As shown in Figure 13.34, we assume the sampling action starts immediately upon receiving the RF signal without any time delay. In practice, however, there is always a time delay between the arrival time of the received RF signal and the first sampling time. In each repetition interval T_R of the received RF signal, the RF signal is sampled once over the pulse length or pulse duration τ_s of the sampling pulse. This sample contains an average of all individual samples taken during τ_s. The sampling time is sequentially increased from pulse to pulse as

$$T_{\text{sn}} = nT_s + nT_d \quad (n = 1, 2, \ldots, N) \tag{13.58}$$

as illustrated in Figures 13.34 and 13.35, where T_s is the period of the sampling pulse, T_d is the time delay for each sampling pulse, and N is the total number of samples. The time delay is needed to facilitate sampling at different points of the RF signal. It is noted that the stability of the delay time T_d is very important; that is, the time delay in the sequential sampling times should be as close as possible. In other words, the sampling instants should follow Eq. (13.58) as faithful as possible. Instability of this time delay produces distortion in the sampled signal, causing it becomes a noisy signal. The pulse length τ_s of the sampling pulse must be smaller than the pulse length τ_R of the received RF signal to allow a portion of the signal to be sampled. The sampling frequency f_s can be much smaller than the highest frequency f_m of the RF signal being sampled and down-converted – for example, $f_s = 10\,\text{MHz}$ and $f_m = 6\,\text{GHz}$. In the sampling process, the received RF signal is converted into another signal over a long time or, in other words, the received RF signal appearing in a

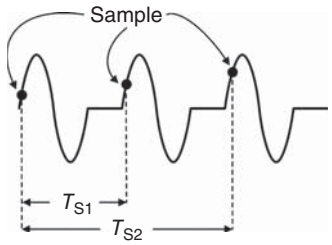


Figure 13.35. Illustration of sampling points.

short time (i.e., at high frequency) is transformed into another signal over a long-time interval (i.e., at low frequency).

Typical sampling mixers have high noise figure. To overcome this problem, a high-gain LNA should be used in front of a sampling mixer in order to improve the overall noise figure of a sampling receiver. In order to increase the sampling efficiency, it is desired to increase the amplitudes of the spectral components of the down-converted IF signal in the interested IF range. The operating bandwidth of a sampling mixer can be optimized by proper circuit design and choice of the sampling pulse.

13.5.4 Sample and Hold

Sampling mixers implement "sample-and-hold" techniques to produce sampling signals. Sample-and-hold techniques thus play the most crucial role in the operation of sampling mixers. We will address the principle of sample-and-hold techniques and then discuss sample-and-hold circuits, which implement these techniques.

13.5.4.1 Principle of Sample-and-Hold. Figure 13.36(a) illustrates the principle of sample and hold considering a received sinusoidal RF signal and a sampling pulse that is a periodic train of rectangular pulses. A sampling pulse consisting of periodic impulses is also shown in Figure 13.36(b) for ideal sampling consideration. The process of sample and hold, as shown in Figure 13.36(a), is as follows. The received RF signal is sampled at sampling instants over the sampling duration τ_s which is the pulse-width of the sampling pulse. The sampled value in each instant is held until the next sampling time – that is, the sampled value is held over the entire pulse-repetition period T_s of the sampling pulse. The sampled signal can be written from (13.54) as

$$f_s(t) = \left[f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \right] \prod \left(\frac{t}{T_s} - \frac{1}{2} \right) \tag{13.59}$$

where

$$\prod \left(\frac{t}{T_s} - \frac{1}{2} \right) = \begin{cases} 1 & 0 \le t \le T_s \\ 0 & \text{otherwise} \end{cases} \tag{13.60}$$

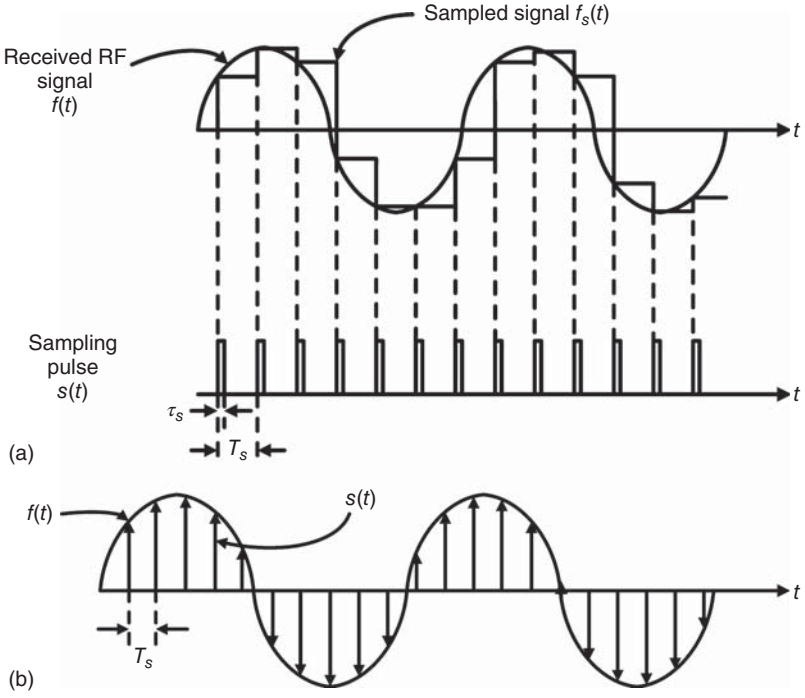


Figure 13.36. Sample and hold using sampling rectangular pulses (a) and impulses (b).

Taking the Fourier transform of $f_s(t)$ in (13.59) gives

$$F_s(f) = \frac{\sin(\pi f T_s)}{\pi f T_s} e^{-j\pi f T_s} \sum_{n=-\infty}^{\infty} F\left(f - \frac{n}{T_s}\right) \tag{13.61}$$

We recall that for an idea sampling, which uses a periodic train of impulses $\delta(t - nT_s)$, the sampled signal and its Fourier transform are given by (13.54) and (13.57), respectively. This ideal sampling is described in Figure 13.36(b). The sampled signal in the frequency domain $F_s(f)$, as described in (13.61), is basically the (frequency-domain) signal being sampled $F(f)$ shifted by multiples of the sampling frequency $f_s(= 1/T_s)$. Therefore, the sampled signal can be recovered exactly from the original signal; that is, the sampled signal is an undistorted version of the original signal. On the other hand, the sampled signal based on sampling rectangular pulses is not an undistorted version of the signal being sampled since the sampled signal's spectrum, although is similar to that of an ideal one, has a "sinc" envelope as can be inferred from (13.61).

13.5.4.2 Sample-and-Hold Circuit. Figure 13.37(a) shows a basic sample-and-hold or sampling circuit, which is essentially a general sampling mixer, consisting of a switch, controlled by a sampling pulse $g(t)$, and a (holding) capacitor C_H. The signal being sampled, represented by the voltage $V_{\text{in}}(t)$, is sent to the circuit. During the sampling time, the switch is on; otherwise, it is off. The switch closes instantaneously at the sampling instant for a short period of time controlled by $g(t)$, allowing the holding capacitor C_H to charge to a portion of the input voltage $V_{\text{in}}(t)$ or to the full value of V_{in} under the optimum condition to produce an output (sampled) voltage $V_{\text{out}}(t)$. The switch is then opened as controlled by $g(t)$ and the voltage stored in the capacitor C_H remains constant until the next sampling instant. The sampling efficiency of a sampling circuit is defined as

$$\eta = \frac{V_{\text{out}}}{V_{\text{in}}} \tag{13.62}$$

We assume that the voltage across C_H resets to zero before each new sample. Practical circuits employ "discharge loop" to discharge the capacitor C_H (e.g., using a resistor across C_H). Ideally, C_H discharges completely

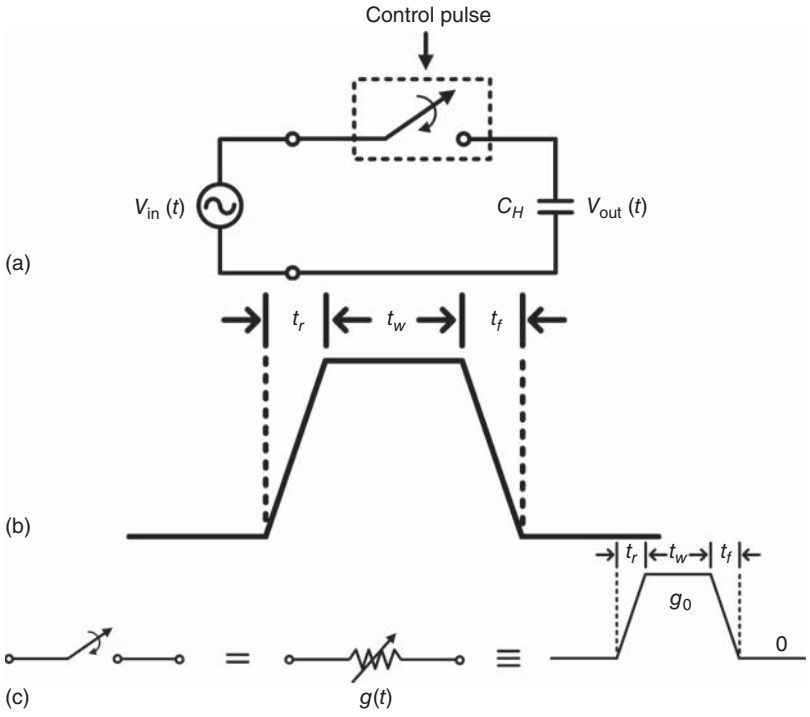


Figure 13.37. Basic sample-and-hold circuit (a), sampling pulse (b), and sampling switch's model (c). It is assumed $t_f = t_r$.

before the next sampling pulse arrives (i.e., before a new sample). The bandwidth of a sampling circuit is normally defined as 3-dB; that is, from DC to f_H, where f_H is the frequency where the output voltage V_{out} is 0.707 of its value at DC or the lowest operating frequency. In practice, a buffer amplifier is normally used before and after a sample-and-hold circuit to isolate the input from the sampling circuit and the sampling circuit from other circuits after it, respectively.

The basic sampling circuit, as shown in Figure 13.37(a), is modeled as a switch loaded by a capacitor C_H, which holds the sampled voltage after sampling. For simplicity, we assume that the sampling pulse used to control the switch is symmetrical and has equal linear rising and falling edges as shown in Figure 13.37(b). An ideal switch would have two distinguished states, on and off with infinite and zero conductance, respectively, with no time delay between them. Practical switches, however, have a finite conductance in either on or off state and time delay between them. The time delay is caused by the response of the solid-state device and the rising and falling times of the controlled pulse. Therefore, neglecting the time response of the solid-state device, a pulse-controlled sampling switch can be represented as a time-varying resistor having time-variant conductance $g(t)$ as described in Figure 13.37(c), where the time delay between the on and off states is the rising/falling time t_r of the controlled pulse over which the conductance is assumed to vary linearly between 0 (minimum) and g_o (maximum). This time-varying conductance $g(t)$ can be written as

$$g(t) = \begin{cases} \dfrac{g_0}{t_r} t, & 0 < t < t_r \\ g_0, & t_r < t < t_w - t_r \\ \dfrac{g_0}{t_r} (t_w - t), & t_w - t_r < t < t_w \end{cases} \tag{13.63}$$

where t_w is the duration of the pulse and g_o is the on-conductance of the (on-state) sampling switch. Note that we assume the switch has zero conductance under the off-state.

A first-order linear differential equation can be written for the sampling circuit in Figure 13.37(a) as

$$C_H \frac{dV_{\text{out}}(t)}{dt} = [V_{\text{in}}(t) - V_{\text{out}}(t)]g(t) \tag{13.64}$$

The (input-to-output) transfer function $h(t)$ of the sampling circuit can be analytically solved, making use of (13.63) and (13.64), as [6]

$$h(t) = \begin{cases} \dfrac{g_0 t}{C_H t_r} e^{\frac{\left(t^2 + 2t_r^2 - 2t_w t_r\right) g_0}{2C_H t_r}}, & 0 < t < t_r \\ \dfrac{g_0}{C_H} e^{\frac{(2t + t_r - 2t_w) g_0}{2C_H}}, & t_r < t < t_w - t_r \\ \dfrac{g_0 (t_w - t)}{C_H t_r} e^{\frac{(t - t_w)^2 g_0}{2C_H t_r}}, & t_w - t_r < t < t_w \end{cases} \tag{13.65}$$

Taking a Fourier transformation of $h(t)$ gives the frequency response of the sampling circuit. The transfer function in both time and frequency domain depends on the controlled pulse. Calculations of the sampling circuit's frequency responses for various parameters of the controlled pulse indicate that the bandwidth of the sampling is relatively independent of the pulse duration t_w while directly affected by the rising/falling time t_r. The expressions in (13.65) and hence their Fourier transforms are, unfortunately, not simple to give an insightful design guideline which is useful (in fact desirable) for design purposes. To overcome this problem, the calculated bandwidth of a sampling circuit with holding capacitance $C_H = 1\,\text{pF}$ and switch's on-conductance

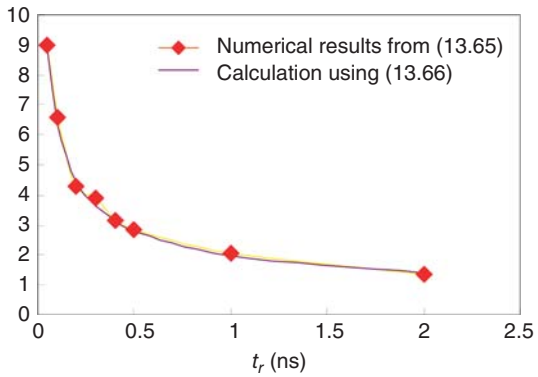


Figure 13.38. Calculated sampling bandwidth based on Eqs. (13.65) and (13.66). The pulse width t_w is set to 0 when calculating the bandwidth using (13.65) and the results are similar for other values of t_w.

$g_o = 1\,\text{S}$ is curve-fitted to derive an empirical equation for estimating the 3-dB bandwidth of the sampling circuit as

$$B \simeq \frac{6.3}{\sqrt{t_r}} \tag{13.66}$$

where t_r is in nanoseconds and bandwidth B is in gigahertz. Equation (13.66) shows that the 3-dB bandwidth is approximately inversely proportional to the square root of t_r. Figure 13.38 compares the sampling bandwidth calculated using (13.66) and that based on the numerical results obtained from (13.65). As can be seen, a good agreement is obtained between the two results, making the simple equation (13.66) a good equation for design purposes. Considering that the bandwidth of a sampling circuit is also affected by its RC constant as $B = 1/\sqrt{C_H R_{on}}$, where $R_{on} = 1/g_o$ is the on-resistance of the switch, (13.66) can be modified to take into account the switch's on-resistance and holding capacitance as

$$B \simeq \frac{6.3}{\sqrt{t_r C_H / g_o}} \tag{13.67}$$

where t_r is in nanoseconds, B is in gigahertz, and C_H and g_o are in picofarad and siemens, respectively.

In addition to the parameters in (13.67), the jitter of a sampling oscillator or clock, which sets the PRF of the sampling pulse, can also cause significant sampling bandwidth reduction [7]. Figure 13.39 depicts how the jitter of a sampling clock increases the noise on the sampled voltage. The dithered clock causes the sampling position to vary randomly, which appears as a sampling error. The rms of the resultant sampled voltage error

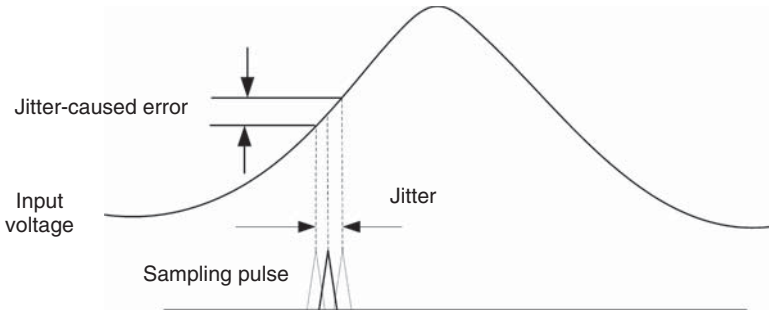


Figure 13.39. Illustration of jitter-caused noise on sampled voltage.

$\widetilde{N}_s$ due to a certain rms jitter Δt of the sampling clock can be derived from the plot in Figure 13.39 as

$$\widetilde{N}_s = \left.\frac{dV(t)}{dt}\right|_{\text{rms}} \times \Delta t = \sqrt{\frac{1}{T}\int_0^T \left(\frac{dV(t)}{dt}\right)^2 dt} \times \Delta t \tag{13.68}$$

where T is the period of the sampled signal. Assume the input RF voltage is $V(t) = A\sin(2\pi ft)$, where f and A are the frequency and voltage amplitude of the RF signal, respectively, then

$$\widetilde{N}_s = A\sqrt{2}\pi \mathrm{f}\Delta \mathrm{t} \tag{13.69}$$

It can be deduced from (13.69) that, as the frequency of the input signal increases, the jitter-caused noise rises accordingly. This explains the reduction of the sampling bandwidth with the existence of a sampling clock jitter.

The conversion gain/loss of a sampling circuit, even for a very basic one as shown in Figure 13.37(a), however, cannot be accurately modeled by (13.65) since such a model is linear and there is no frequency transformation involved. The nonlinear model in [8] indicates that the smaller duty cycle the sampling pulse has, the higher conversion gain the sampling circuit has. Figure 13.40 demonstrates the relationship between conversion gain and duty cycle based on the model in [8]. This is easy to understand because a wider sampling pulse with larger duty cycle has more "switch-on time," during which the holding capacitor has RF-signal passing. Accordingly, the sampled output voltage will contain more RF power but less IF power, resulting a lower conversion gain. Figure 13.41 shows the simulated conversion gain with 300 ps- and 10 ns-width sampling pulse. The period of the sampling clock is 25 ns. The simulated 4 dB gain difference between the two sampling pulses is quite close to the results given in Figure 13.40.

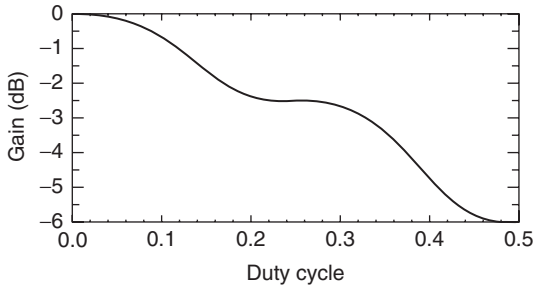


Figure 13.40. Relationship between conversion gain and duty cycle for an ideal sampling circuit.

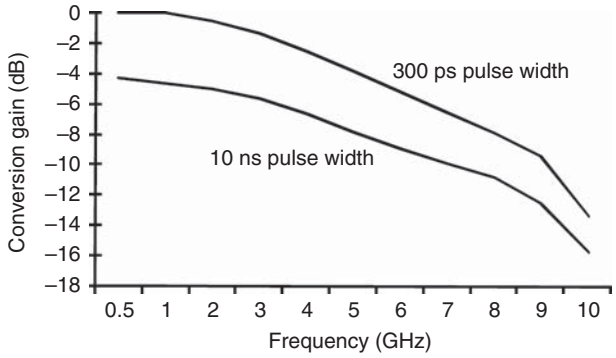


Figure 13.41. Conversion-gain comparison for different sampling pulse durations.

13.5.5 Sampling Switch

The function of sampling switches, like any switch, is simply switching on and off. A sampling switch can hence be implemented using any switch topology such as those described in Chapter 14. Here we consider either an NMOS switch or an NMOS/PMOS transmission gate consisting of NMOS and PMOS connected in parallel as shown in Figure 13.42. The NMOS switch has a simpler structure and requires only one control signal. The NMOS/PMOS transmission gate requires a pair of control signals but is more linear. To investigate their effects on the noise figure of sampling mixers employing them, a frequency-domain simulation is conducted for a simple sampling mixer consisting of a sampling switch and capacitor. The sampling clock frequency is set to 40 MHz. We consider three different switch topologies. The first topology uses a single NMOS as the sampling switch with 150 ps rising/falling time (t_r), the second topology implements an NMOS/PMOS pair switch with 150 ps rising/falling time, and third one uses an NMOS/PMOS pair switch with 100 ps rising/falling time. As can be seen in Figure 13.42, the effect of the $1/f$ noise can be observed in the frequency range below 100 kHz. The NMOS/PMOS transmission gate has dramatically less $1/f$ noise due to better noise performance of PMOS. For the NMOS switch, increasing the transistor size increases the $1/f$ noise according to the simulation. Shorter rising/falling time can bring down the $1/f$ noise floor. At 40, 80, and 120 MHz, noise peaks appear. This is due to the harmonics of the sampling clock at the output of the sampling capacitor. From the comparison, we can conclude that an NMOS/PMOS transmission gate is preferred for a sampling switch in order to achieve better noise performance for sampling mixers.

13.5.6 Integrated Sampling Mixer

LNA is typically placed before an RF sampling mixer to provide gain and lower the overall noise figure. In integrated circuit environments such as a single CMOS RF receiver chip, the LNA should be integrated with the sampling mixer as a single unit to improve the overall performance and reduce the circuit complexity, size, and cost. This section describes such a fully integrated sampling mixer incorporating an LNA and sampling switch together (with or without a buffer) [9].

13.5.6.1 Integrated Sampling Mixer Architecture and Analysis. Figure 13.44 shows the block diagram of the integrated sampling mixer which consists of an LNA integrated with a pulse-controlled sampling switch, a holding capacitor (C_H) and an (optional) buffer. When a sampling strobe (pulse) arrives at the switch, the switch will be turned on to allow the sample-and-hold capacitor keep track of the LNA output signal. Besides providing the essential input impedance matching (which is not necessarily 50 Ω), the LNA also produces a voltage gain that contributes to the conversion gain and reduces the overall noise figure of the integrated sampling mixer. Figure 13.44 shows an equivalent-circuit model for noise analysis based on

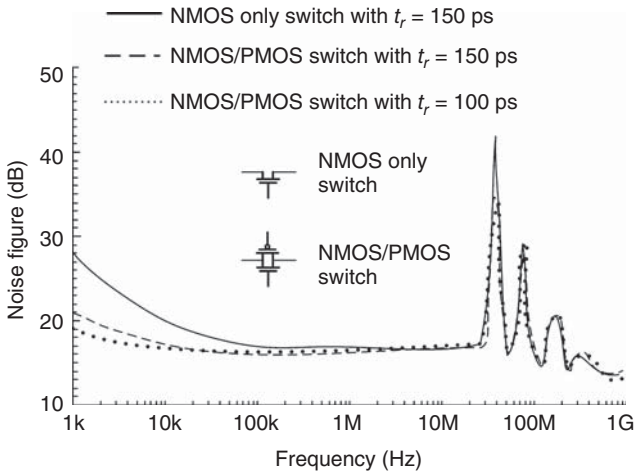


Figure 13.42. Noise figure of a simple sampling mixer with different switch structures.

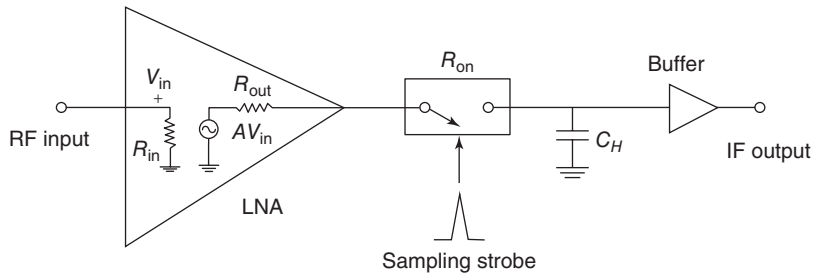


Figure 13.43. Integrated sampling mixer. The schematic inside the LNA represents its simplified small-signal model. The buffer is optional. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

Figure 13.43. Two noise sources are included for the analysis: one is n_s representing the output noise of the LNA and the other is n_{Ron} caused by the thermal noise of the sampling switch's on-resistance R_{on}. It can be easily shown that

$$n_s^2 = n_i^2 FG \tag{13.70}$$

where n_i^2 is the input noise density due to the input source impedance, and F and G are the noise figure and gain of the LNA, respectively. We can also derive

$$n_{\text{Ron}}^2 = 4KTR_{\text{on}} \tag{13.71}$$

where K is the Boltzmann constant and T is the temperature in kelvin (K). In the $R-C$ based equivalent circuit of Figure 13.44, the equivalent noise bandwidth can be derived as

$$\Delta f = \int_0^\infty |H(j\omega|^2\, df = \int_0^\infty \left| \frac{1/j\omega C_H}{1/j\omega C_H + R_{\text{on}} + R_{\text{out}}} \right|^2 df = \frac{1}{4(R_{\text{on}} + R_{\text{out}})C_H} \tag{13.72}$$

If the sampling clock jitter-caused noise in (13.69) is included, the total output noise of the integrated sampling mixer (including LNA) can be approximately derived as

$$\begin{aligned} \overline{N_o}^2 &= (n_s^2 + n_{\text{Ron}}^2)\Delta f + 2(\pi f A \Delta t)^2 \\ &= \frac{n_i{}^2 FG + 4KTR_{\text{on}}}{4C_H(R_{\text{on}} + R_{\text{out}})} + 2(\pi f A \Delta t)^2 \end{aligned} \tag{13.73}$$

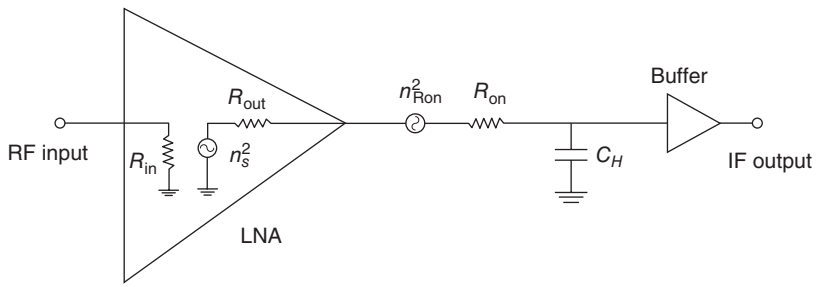


Figure 13.44. Equivalent circuit of the integrated sampling mixer for calculating output noise. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

It is noted that $2(\pi fA\Delta t)^2$ represents the noise (power) error caused by the sampling clock jitter. The total noise figure of the integrated sampling mixer can hence be derived as

$$F_{\text{mixer}} = \frac{n_o^2}{Gn_i^2} = \frac{\overline{N}_o^{\,2}}{Gn_i^2}\frac{2}{f_s}$$
$$= \left[\frac{F}{4C_H\left(R_{\text{on}} + R_{\text{out}}\right)} + \frac{4KTR_{\text{on}}}{4C_H(R_{\text{on}} + R_{\text{out}})Gn_i^2} + \frac{2(\pi fA\Delta t)^2}{Gn_i^2}\right]\frac{2}{f_s} \tag{13.74}$$

where f_s is the sampling frequency. This formula is useful for evaluating the noise figure qualitatively. In order to minimize the noise figure of the integrated sampling mixer, the constituent LNA needs to have a high gain and maintain a low noise figure itself, hence rendering difficulty in achieving low power consumption. For subsystems such as multi-band receivers or those used in wideband measurement instruments, where narrow-band prefiltering typically implemented in narrow-band systems to remove noise is not feasible, the noise reduction for the integrated sampling mixer mainly depends on the gain of the LNA. Under this circumstance, the LNA needs to maintain a certain gain over a wide frequency range and, as a result, may draw a large amount of DC current that may jeopardize the power consumption of the whole subsystem. It is noted that in the integrated sampling mixer shown in Figure 13.43, while samplings occur only at discrete times, the LNA runs continuously, resulting in not only RF-power inefficiency but also unnecessarily increased power consumption.

As indicated in Section 13.5.4.2 (e.g., see Figure 13.41), the narrower the sampling pulse, the higher the conversion efficiency. Additionally, the sampling bandwidth depends on the sampling aperture time which, in turn, depends on the sampling pulse width. A narrow sampling pulse is thus typically preferred. The duration of CMOS-based generated pulses is usually hundreds of picoseconds while the period of a sampling clock is typically several nanoseconds at least. The resultant low duty cycle of sampling pulses means that, during most of the time, the sampling switch is off and the circuit is in "hold" status within which no RF signal needs to be transmitted or amplified. A large portion of RF power will then be wasted if the LNA is turned on all the time. An integrated sampling mixer incorporating a switching LNA using a power-efficient two-stage switching technique can solve this problem. Figure 13.45 shows the block diagram of such an integrated sampling mixer [9]. The LNA is gated through its internal switch by a pulse produced by pulse generator 1, which should be wide enough to allow the LNA to start and reach stabilization. Pulse generator 2 generates a narrower subnanosecond pulse to conduct fast sampling after the LNA is turned on and stabilized. The output time domain waveforms of different blocks in the integrated sampling mixer are illustrated in Figure 13.46. It is noted that the sampling pulse from pulse generator 2 needs to fall behind the rising edge of the LNA's switching signal from pulse generator 1 to accommodate possible slow switching time of the LNA. The synchronization of the two pulse generators is realized by using a common sampling clock. Assuming a 100-MHz sampling frequency is used and pulse generator 1 has a pulse duration of 2 ns, then the LNA is turned on only during 20% of the time. In other words, 80% of the power can be saved by switching the LNA. If lower sampling frequencies were used then even more power can be saved.

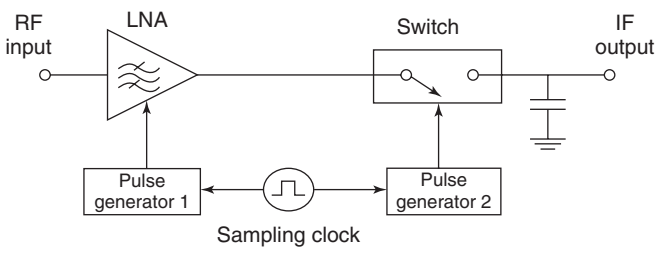


Figure 13.45. Block diagram of an integrated sampling mixer integrating a double-stage switching sampling mixer with a switching LNA and pulse generators. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

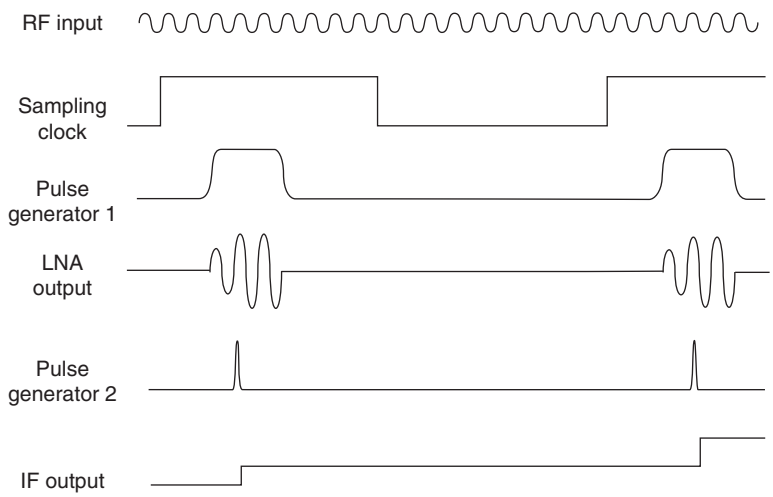


Figure 13.46. Waveforms of building blocks in the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

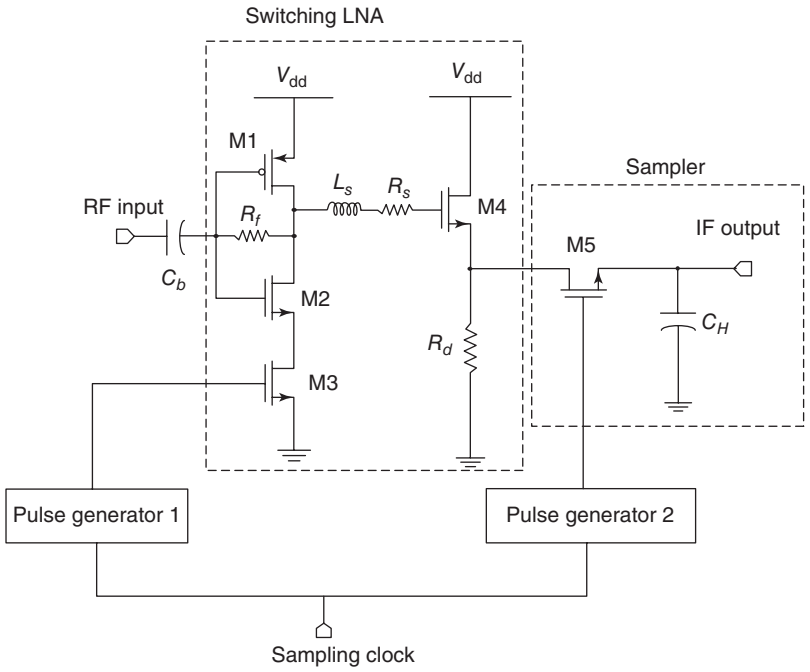


Figure 13.47. Schematic of the integrated sampling mixer. The pulse generators' schematics are shown in Figure 13.51. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

13.5.6.2 Switching LNA Design. As shown in Figure 13.45, a switching LNA is needed for the integrated sampling mixer. Herein, we choose a resistive feedback push–pull structure for the switching LNA. Figure 13.47 shows the schematic of the CMOS switching LNA as part of the integrated sampling mixer. Transistor pair M1 and M2 forms a push–pull configuration to provide a large transconductance g_m. Serial feedback resistor R_f is utilized to achieve broadband gain. Transistor M3, controlled by the pulse generated by pulse generator 1, functions as a switch to turn on and off the DC current of the first LNA stage. The DC-switch M3 is placed as close to the ground as possible to ensure its source terminal is at a low DC biasing point. Therefore, with a fixed turn-on gate voltage, the voltage (V_{gs}) that determines the switching-on resistance should have enough swing as close as possible to, yet always smaller than, V_{dd} to turn on the DC path.

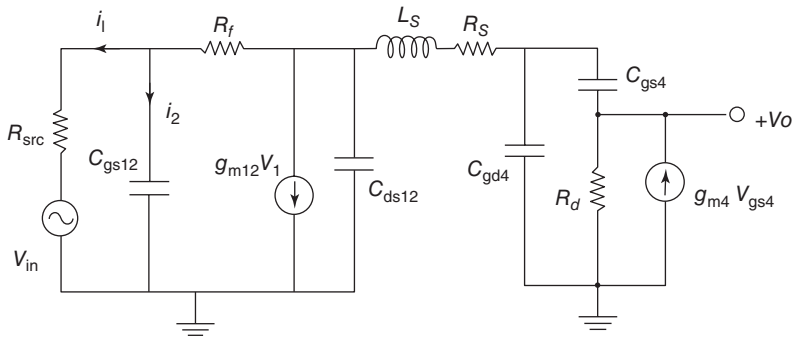


Figure 13.48. Small-signal equivalent circuit of the switching LNA with M3 turned on. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

This also makes the single-ended LNA structure implemented in the integrated sampling mixer preferred than a differential LNA for better switching performance. A differential LNA might require the amplitude of the switching signals to exceed V_{dd}. Transistor M5 is the sampling switch and controlled by the sampling strobe from pulse generator 2.

The small-signal equivalent circuit of the switching LNA with M3 turned on is given in Figure 13.48, in which C_{gs12}, C_{gd12}, g_{m12}, r_{o12}, and C_{d12} are the combined gate–source capacitance, gate–drain capacitance, transconductance, output resistance, and drain–ground capacitance of M1 and M2, respectively. C_{gs4}, C_{gd4}, and g_{m4} stand for the gate–source capacitance, gate–drain capacitance, and transconductance of transistor M4. V_{in} is the input RF voltage and V_{gs4} is the gate–source voltage across transistor M4. The input impedance of the switching LNA can be estimated as

$$Z_{\text{in}} \simeq \frac{1}{g_{m12} + j\omega C_{\text{gs12}}} \tag{13.75}$$

where $g_{\text{m12}} = g_{\text{m1}} + g_{\text{m1}}$ and $C_{\text{gs12}} = C_{\text{gs1}} + C_{\text{gs2}}$. This estimation is valid only at low frequencies where R_f is relatively small compared with the impedance of the parasitic capacitance of the transistors.

A serial peaking inductor L_s is inserted between the first stage and source follower of the LNA to extend the bandwidth of the gain. Resistor R_s is connected in series with L_s to ensure stability for the LNA and help achieve a flat gain over the interested frequency range. The use of R_s herein allows the (internal) parasitic resistance of the inductor to be absorbed into R_s so that the inductor can be electrically represented by its inductance only instead of with its parasitic resistance, effectively helping relax the quality-factor requirement for inductor L_s. Based on the small-signal equivalent circuit in Figure 13.48, the voltage-gain expression of the LNA can be written as

$$\begin{aligned} G(j\omega) &= G_1(j\omega) \cdot H_{12}(j\omega) \cdot G_2(j\omega) \\ &\simeq G_1(j\omega) \cdot \frac{1}{1 - \omega^2 L_s C_{g4} + j\omega R_s C_{g4}} \cdot G_2(j\omega) \end{aligned} \tag{13.76}$$

where $G_1(j\omega)$ and $G_2(j\omega)$ are the voltage gains of the first and second stage, respectively, $H_{12}(j\omega)$ is the frequency response of the LC component between the first and second stage, and $C_{g4} = C_{\text{gd4}} + C_{\text{gs4}}/(1 + R_d g_{\text{m4}})$. Equation (13.76) can be rearranged as

$$G(j\omega) = G_1(j\omega) \cdot \frac{H_0}{{\omega_n}^2 - \omega^2 + j\omega\omega_n\xi} \cdot G_2(j\omega) \tag{13.77}$$

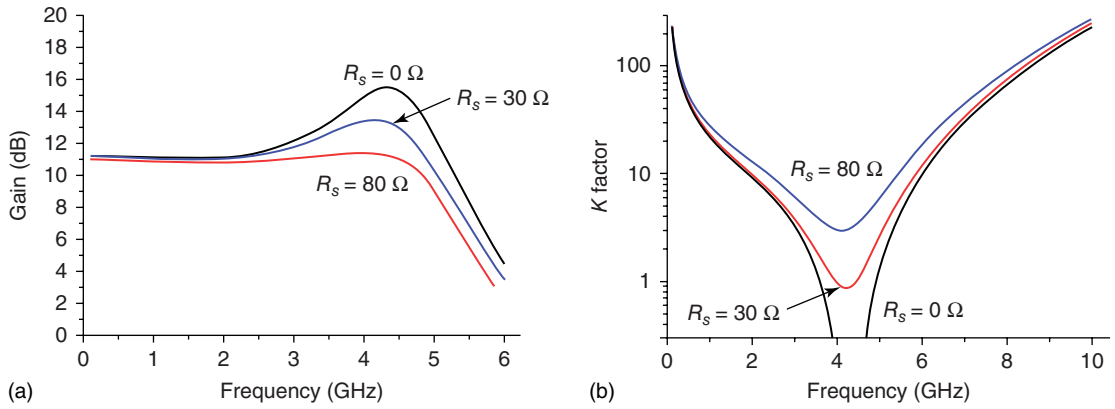


Figure 13.49. Effects of serial resistor R_s on LNA gain (a) and stability factor K (b). (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

where

$$H_0 = \frac{1}{L_s C_{g4}}$$

$$\omega_n = \sqrt{\frac{1}{L_s C_{g4}}}$$

$$\xi = R_s \sqrt{\frac{C_{g4}}{L_s}} \tag{13.78}$$

Without the peaking inductor L_s, the gain of the LNA will drop with increasing frequency due to the drain–source capacitance C_{ds12} and gate–drain capacitance C_{gd4} of the transistors. Adding such an inductor (L_s) between the first and second stage creates a peak to the total gain at the natural frequency ω_n. If L_s is carefully chosen so that the peak can be used to compensate for the gain drop, then the bandwidth of the LNA can be expanded. However, the amplitude of the peak also needs to be controlled so that a flat frequency response can be achieved. Besides, a large peak may lead to instability for the LNA which necessitates the use of the serial resistor R_s to suppress a potential negative resistance. Figure 13.49 illustrates the effects of the serial resistor R_s upon the LNA gain's frequency response and the stability factor K. It can be seen that at the natural frequency ω_n, a peak and a trough appear for the gain and K factor, respectively. As the value of

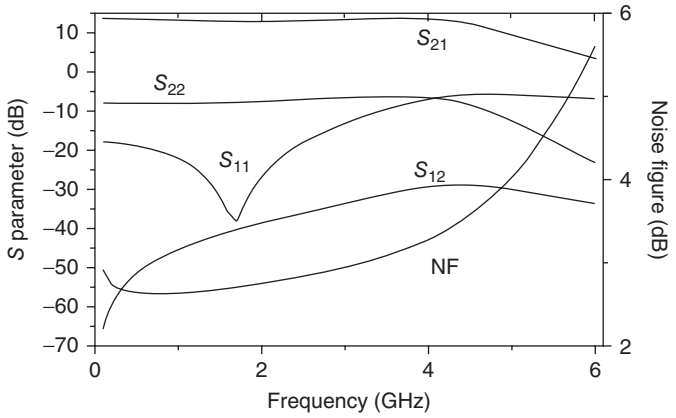


Figure 13.50. Simulated S-parameters and noise figure of the switching LNA. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

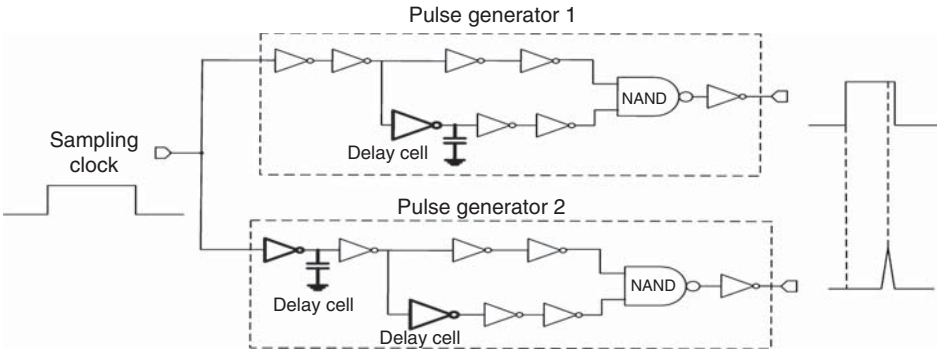


Figure 13.51. Pulse generators 1 and 2 with relatively large and narrow output pulses, respectively. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

R_s grows, the gain becomes flatter and the K factor becomes larger at the trough. The optimal value for R_s is finally chosen as 80 Ω. The simulated noise figure and S-parameters of the LNA are presented in Figure 13.50. From DC to 4 GHz, the calculated gain is 13 ± 1 dB and the input return loss is larger than 10 dB. The noise figure is below 4 dB within this frequency band.

13.5.6.3 Pulse Generator Design. Figure 13.51 shows the schematics of pulse generators 1 and 2 of the integrated sampling mixer shown in Figures 13.45 and 13.47. These pulse generators use digital NAND gates to generate pulses. They can provide large voltage swings required for turning the switches on and off. They contain inverter chains for delay cells and edge-sharpening purpose. Sharpening of the rising/falling edge of the sampling clock is needed for subsequent generation of narrow impulse-like signals. The sharpened clock signal is split into two branches. In one of the branch, the clock is delayed and inverted with respect to the other. The NAND gate then combines the rising and falling edges of these clocks to form an impulse. The duration of the generated impulse is determined by the delay between the rising and falling edges. The use of inverter chains, however, inadvertently introduces more jitter that can contaminate the input sampling clock. The thermal current noise in each inverter stage causes random position change on the crossing point of half-V_{dd} as illustrated in Figure 13.52. The thermal current noise spectral density can be modeled as

$$\overline{i}_n = \sqrt{4kT\gamma g_m} \tag{13.79}$$

where g_m is the transconductance of the NMOS or PMOS transistor in the inverter, and γ is a coefficient approximately equal to 2/3 for long-channel transistors in saturation region and typically two to three times larger for short-channel devices. The load of the inverter is a parallel combination of the inverter's load

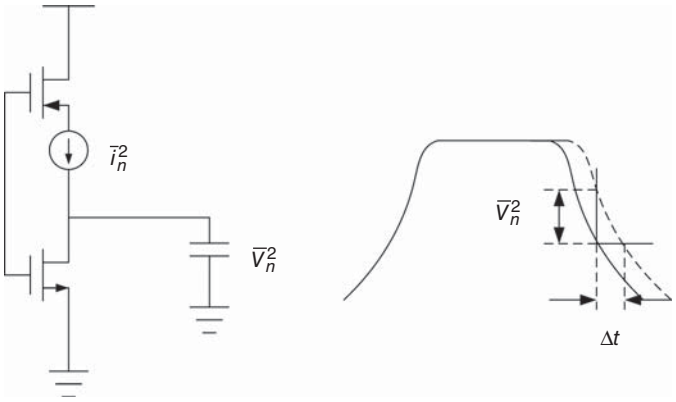


Figure 13.52. Transformation from thermal current noise to jitter. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

capacitance and output resistance. Therefore,

$$\overline{V}_n^{\,2} = \int_0^{\infty} \bar{i}_n \left(r_o + \frac{1}{j\omega C_L} \right) df = \sqrt{\frac{kT\gamma g_m r_o}{C_L}} \tag{13.80}$$

where r_o is the output resistance of the NMOS or PMOS transistor in the inverter and C_L is the load capacitance of the inverter. If the rising and falling edge is linear (i.e., the slope is constant), the jitter in each inverter stage can be estimated as

$$\Delta t_i = \frac{\overline{V}_n}{\mathrm{SR}} = \frac{\overline{V}_n}{I_D/C_L} = \frac{\sqrt{kT\gamma g_m r_o C_L}}{I_D} \tag{13.81}$$

where I_D is the drain current of the inverter and SR is the slew rate of the charging procedure.

Assuming the jitter of each inverter stage is independent from each other, the total jitter for an N-stage inverter chain is $(\Delta t)^2 = N(\Delta t_i)^2$. The calculated jitter of each inverter stage is below 1 ps. In (13.81), PMOS is used for calculating the rising-edge jitter while NMOS is used for the falling-edge jitter. The capacitor-loaded delay chain brings in a jitter of no more than a few picoseconds. The total introduced jitter of the pulse generator is estimated to be less than 20 ps based on its circuit topology and component parameters.

13.5.6.4 Simulation of the Integrated Sampling Mixer. Switch M3 in the integrated sampling mixer, as shown in Figure 13.47, is turned on within the sampling time-window across the width of the pulse produced by pulse generator 1. Under the on-operation, this switch can be considered a small resistor. The resistance of this resistor is affected by the gate–source voltage, which may cause trembling gain at the first stage of the LNA as illustrated in Figure 13.53(a). Therefore, it is necessary to investigate the effect of the ripple of the sampling clock to the output of the LNA's first stage. To that end, a transient simulation is conducted with zero RF input. The simulated output ripple amplitude is shown in Figure 13.53(b). From the results, we can see that the effect of the sampling-clock tremble can be omitted as long as its amplitude is kept below 50 mV. Further ripple compression can be achieved by increasing the size of M3, making the switch behaves closer to a short circuit when turned on.

In order to investigate the effects of clock jitter on the sampling bandwidth, time domain simulation is conducted for a basic sampler, consisting of an NMOS switch and a holding capacitor, with a jittered sampling clock, as shown in Figure 13.54(a). The rising/falling edge of the sampling strobe is set to 100 ps, which is roughly the same as that in the designed pulse generator. Figure 13.54(b) shows the simulated conversion

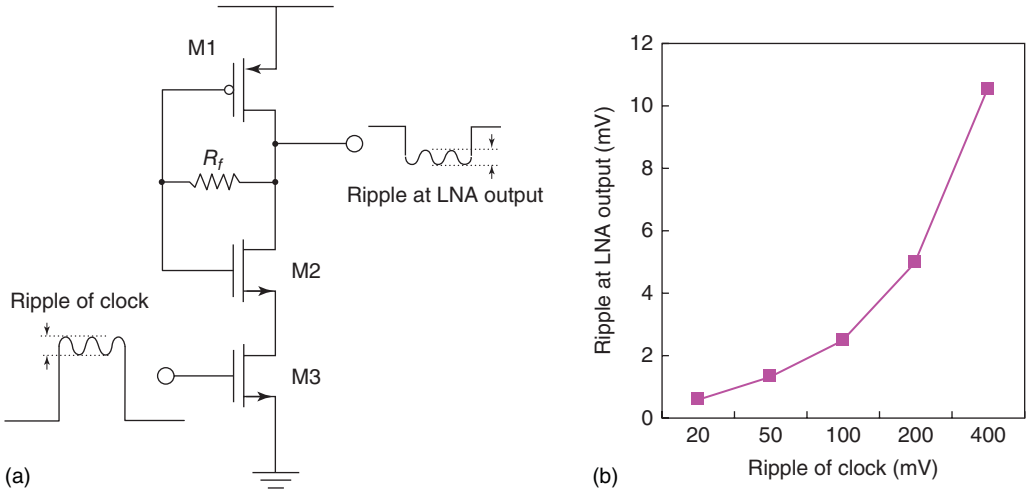


Figure 13.53. (a) Effect of sampling clock ripple and (b) simulated ripple at LNA output. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

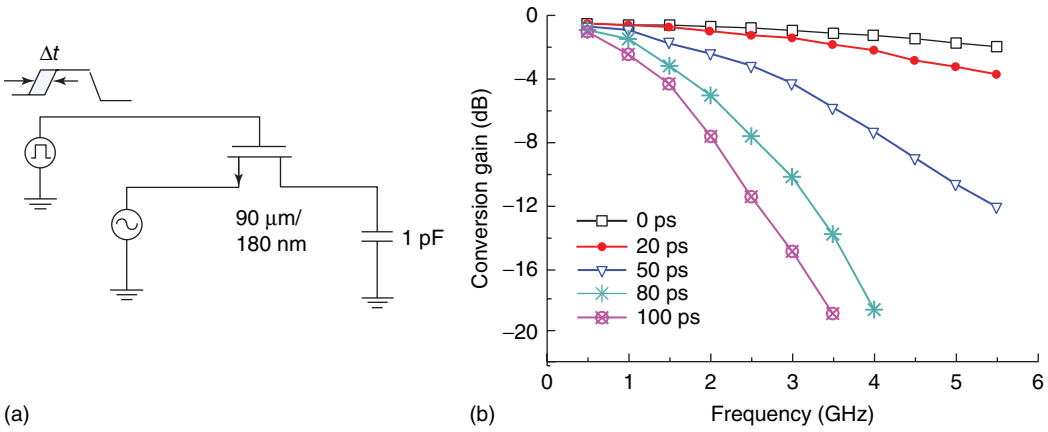


Figure 13.54. Basic sampler with jittered clock (a) and the clock jitter's effects on the sampler's conversion gain (b). The NMOS has a width of 90 μm and a length of 0.18 μm. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

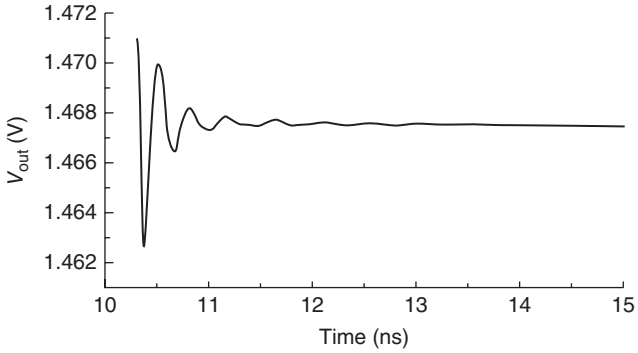


Figure 13.55. Transient response of the output of the switching LNA's first stage. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

gain of the sampler with different sampling clock jitters. As can be seen, a 3-dB sampling bandwidth of over 5.5 GHz can be obtained with a clock jitter smaller than 20 ps, which is large enough to cover the entire bandwidth of the designed switching LNA.

Figure 13.55 shows the output transient response of the first stage of the LNA, indicating that the LNA needs roughly 2 ns to reach stabilization after switch M3 is turned on. Therefore, the sampling strobe from pulse generator 2 should appear at least 2 ns after the rising edge of the pulse coming from pulse generator 1. Figure 13.56 shows the simulated time domain waveforms of different building blocks in the integrated sampling mixer. During the hold status, the drain current of switch M3 is kept at 5.6 μA and the LNA does not produce any RF output. Before the sampling occurs, the gating signal produced by pulse generator 1 arrives and turns on the LNA. As a result, the drain current of switch M3 jumps to 8.3 mA. After the LNA starts up and stabilizes, the sampling strobe coming from pulse generator 2 will reach M5 and conduct the sampling. This sampling strobe has rising/falling edge of 80 ps and is 2 ns behind the gating signal coming from pulse generator 1. The duration of the pulse produced by pulse generator 1 is 3.5 ns, which limits the sampling frequency to below 200 MHz. For sampling frequency higher than 200 MHz, the duration of this pulse and the delay between it and that from pulse generator 2 is different from the designed values of 3.5 and 2 ns, respectively.

Figures 13.57 and 13.58 show the output spectrum and time-domain waveform of the integrated sampling mixer with 505-MHz RF signal and 100-MHz sampling clock. The peak-to-peak swing of the input RF signal is about 20 mV. The observed IF swing is more than 100 mV from the waveform seen in Figure 13.58. As can be seen in the spectrum, besides the down-converted signal at 5 MHz, spurs appear at every harmonic of 5 MHz. These spurs, however, are well suppressed with the ratio between the down-converted signal to the largest harmonic signal greater than 20 dBc.

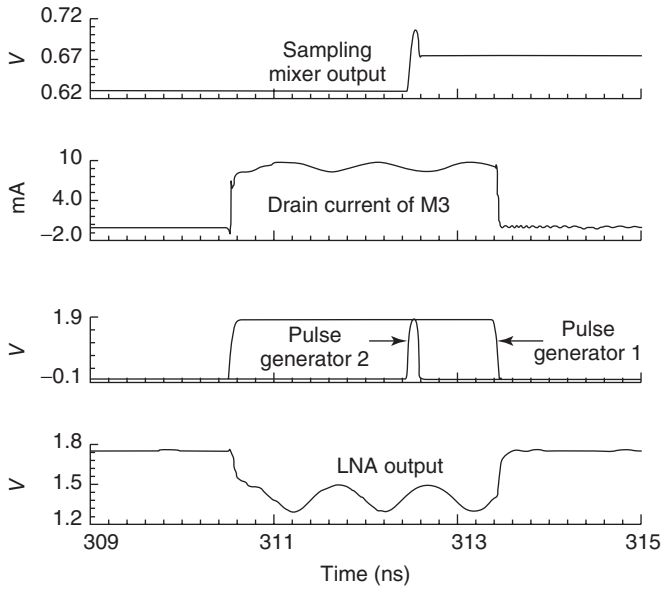


Figure 13.56. Time domain waveforms of different building blocks in the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

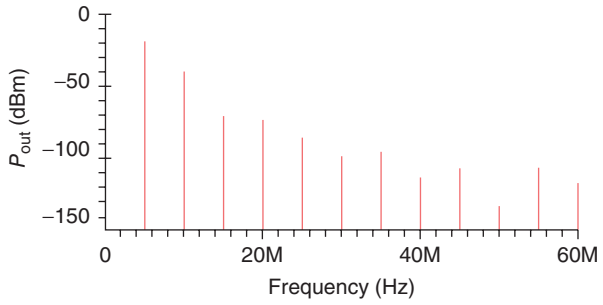


Figure 13.57. Output spectrum of the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

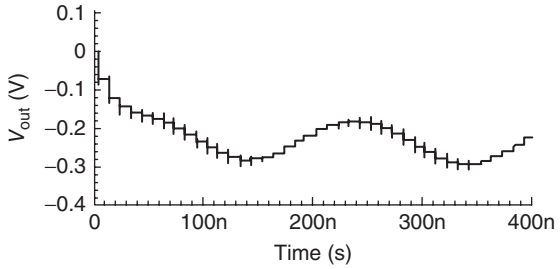


Figure 13.58. Output waveform of the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

13.5.6.5 Integrated Sampling Mixer Performance. The integrated sampling mixer was implemented on the Jazz 0.18-μm enhanced RF CMOS process. The entire sampling mixer chip, including the switching LNA, pulse generators, sampler, RF pads and output buffer, occupies a die area of 750 μm × 730 μm and is shown in Figure 13.59. Its performance was measured on-chip.

Figure 13.60 shows the measured input return loss and the measured conversion gain with 100-MHz and 10-MHz sampling frequency. The input return loss is below −10 dB up to 3.8 GHz. The voltage conversion gain ranges from 9 to 12 dB across DC-3.5 GHz for 100-MHz sampling frequency. When the sampling frequency is reduced to 10 MHz, the conversion-gain bandwidth reduces significantly. This is due to the larger

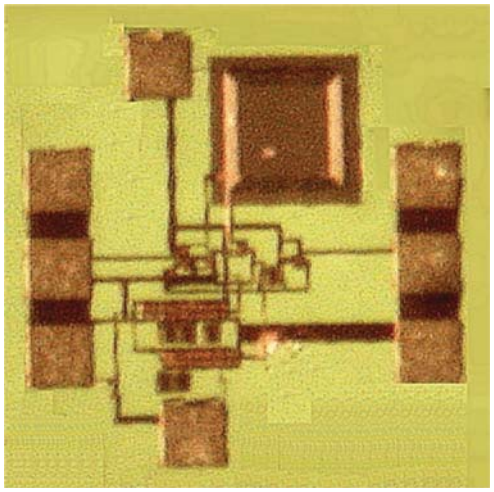


Figure 13.59. Microphotograph of the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

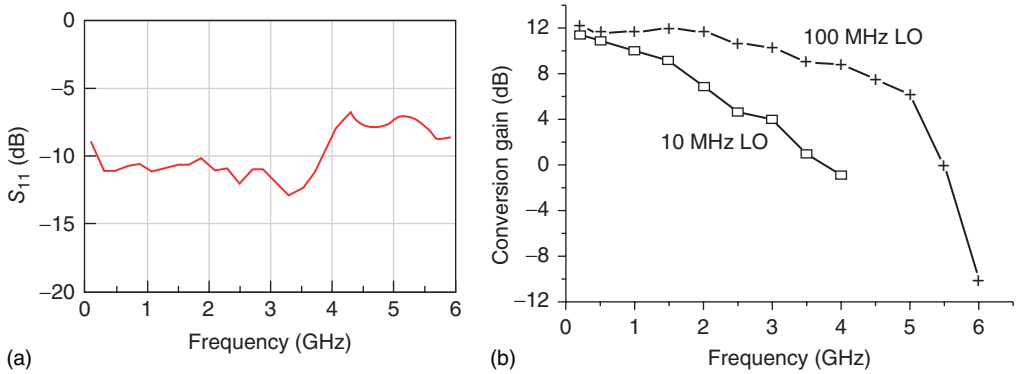


Figure 13.60. Measured input return loss (a) and conversion gain (b) of the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

jitter at 10 MHz. It is noted that the output buffer has a calculated loss of 5 dB. This buffer is used to drive a 50-Ω load for measurement purposes only and is not needed in complete receiver integration. Therefore, 5 dB was added to the measured power conversion gain to obtain the actual voltage conversion gain of the integrated sampling mixer itself (without buffer) as shown in Figure 13.60(b). It is worth noticing that the measured conversion gain drops faster than the power gain shown in Figure 13.50 as frequency is increased. This is mainly due to the jitter from the external generator employed as the sampling clock. Figure 13.61 shows the measured input 1-dB power compression point, around −11 dBm, with the sampling and RF frequencies of 100 and 500 MHz, respectively. As the gain of the switching LNA drops with increasing frequency, the 1-dB compression point will increase accordingly.

Figure 13.62 shows the measured and calculated noise figure with 100-MHz sampling frequency. During the measurement, the output noise floor was recorded with the existence of −20-dBm RF input signal so that the sampling error could be included. As the RF frequency is increased, the elevation of the output noise floor can be observed. This agrees with the fact that the sampling error rises with frequency due to the clock jitter [7]. Using the 100-MHz clock, the noise figure of the integrated sampling mixer lies between 15 and 25 dB from DC to 3.5 GHz. It is about 4–5 dB larger than the simulated value due to the external sampling clock jittering introduced during the measurement. Since the noise figure of the LNA is less than 4 dB, the main noise contribution comes from the sampling stage. Although the sampling switch's loss is only less than 1 dB from DC to 3 GHz, the down-converted noise as well as the sampling clock jitter raise the output noise-floor significantly.

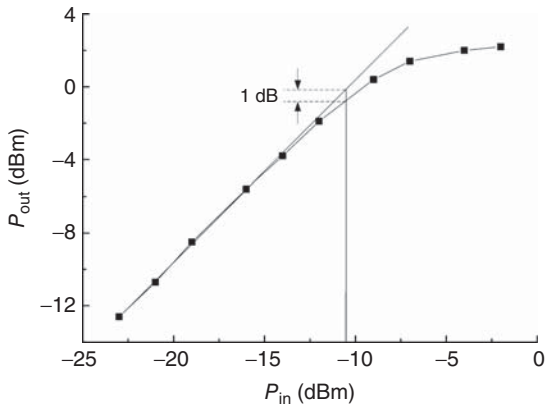


Figure 13.61. Measured 1-dB compression point of the integrated sampling mixer with 100-MHz sampling clock and 500-MHz RF frequency. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

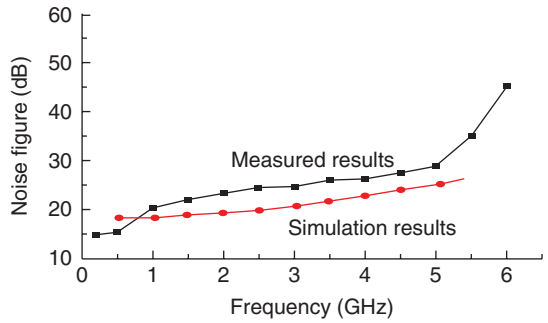


Figure 13.62. Measured and simulated noise figure of the integrated sampling mixer. (After Rui and Nguyen [9]. Reprinted with permission of IEEE.)

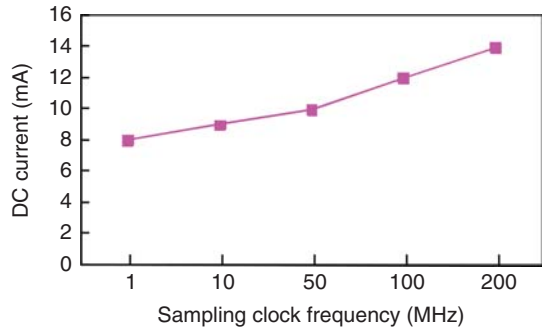


Figure 13.63. Current consumption of the integrated sampling mixer.

The measured power consumption indicates that the integrated sampling mixer including the switching LNA, two pulse generators, sampler, and output buffer consumes 12-mA DC current under a 1.8-V supply voltage. The output buffer draws around 5.5-mA DC current. As indicated earlier, this buffer is used solely for measurement purpose. If this sampling mixer was integrated into a receiver, where a buffer is not needed, then the 5.5-mA output buffer current can be saved, thus leading to even lower power consumption. It is noted that if only one switching stage was used (i.e., using only pulse generator 2), the integrated sampling mixer would draw an additional 8-mA DC current due to the LNA being on all the time. The current consumption of the integrated sampling mixer increases as the (LO) sampling clock frequency is increased due to more "turn-on" time of the LNA. The measured DC current as a function of the sampling clock frequency is displayed in Figure 13.63.

REFERENCES

1. B. Gilbert, "A Precise Four-Quadrant Multiplier with Subnanosecond Response," *IEEE J. Solid-State Circuits*, Vol. 3, pp. 365–373, Dec. 1968.
2. S. A. Mass, "A GaAs MESFET Mixer with Very Low Intermodulation," *IEEE Trans. Microwave Tech.*, Vol. 35, pp. 425–429, Apr. 1987.
3. S. A. Maas, *Microwave Mixer*, Artech House, Inc., Norwood, MA, 1993, pp. 341–343.
4. X. Guan and C. Nguyen, "A Novel CMOS Distributed Receiver Front-End for Wireless Ultra-Wideband Receivers," *Microwave Opt. Technol. Lett.*, Vol. 52, pp. 1790–1792, Aug. 2010.
5. CA18HR 0.18-μm Enhance RF CMOS Process, Jazz Semiconductor, Newport Beach, CA, USA, 2007.
6. P. Eriksson and H. Tenhunen, "A Model for Predicting Sampler RF Bandwidth and Conversion Loss," *Proc. IEEE Int. Conf. Circuits Syst.*, Vol. 6, pp. 18–21, May 1999.
7. M. Shinagawa, Y. Akazawa, and T. Wakimoto, "Jitter Analysis of High-Speed Sampling System," *IEEE J. Solid-State Circuits*, Vol. 25, pp. 220–224, Feb. 1990.
8. A. Parssinen, R. Magoon, S. Long, and V. Porra, "A 2-GHz Subharmonic Sampler for Signal Down Conversion," *IEEE Trans. Microwave Theory Tech.*, Vol. 12, pp. 2344–2351, Dec. 1997.
9. R. Xu and C. Nguyen, "An Ultra-Wideband Low Power-Consumption Low Noise-Figure, High-Gain, RF-Power Efficient DC-3.5-GHz CMOS Integrated Sampling Mixer," *IEEE Trans. Microwave Theory Tech.*, Vol. 56, pp. 1069–1075, May 2008.

PROBLEMS

13.1 Derive Eq. (13.5).

13.2 Consider the two-tone $(m_1 + m_2)^{\text{th}}$ order IM and assume that the two tones and the resultant products have equal amplitudes, respectively. Prove that a 1-dBm increase in the input RF power causes a $(m_1 + m_2)$-dB increase in the IM level. What is the slope of the $(m_1 + m_2)$th order IM curve?

13.3 Sketch the spectrum for the two-tone fifth-order IM signal with respect to other pertinent signals.

13.4 Consider the two-tone fifth-order IM. Sketch the IF and IM power as a function of the input RF power and show the fifth-order intercept point.

13.5 Is there a relation between the 1-dB power compression point ($P_{1\,\text{dB}}$) and third-order intercept point (IP3)? Provide your rationale. If there is a relationship, determine approximately the difference (in decibel) between the input $P_{1\,\text{dB}}$ and IIP3.

13.6 The IF signals and IM3 signals corresponding to two closely spaced RF tones are usually assumed to be equal, respectively. In practice, however, this does not happen precisely. In the IIP3 calculation based on Eq. (13.11), IM(dBc) is defined as the IM signal's power relative to the IF power. What do you think the IM(dBc) should be defined when the two IF signals are not equal and the two IM3 signals are different to best describe the actual IP3 performance of mixers? Provide your rationale.

13.7 Derive Eq. (13.12).

13.8 Draw a double-balanced mixer block diagram similar to that shown in Figure 13.10 but with the RF signal being fed in-phase to the first and third single-ended mixer and 180° out-of-phase to the second and fourth single-ended mixer, and the LO signal arriving to these constituent mixers as in Figure 13.10. Describe the mixer operation and discuss its characteristics.

13.9 Compare the single-ended and single-balanced mixers as shown in Figures 13.6(c) and 13.9(b), respectively. Assume all elements including filters, baluns, and interconnects are ideal. Estimate the improvement in the 1-dB power compression point and third-order intercept point, and the additional LO power needed for the single-balanced mixer as compared to the single-ended mixer.

Compare the conversion gain between these two mixer types. Provide your rationale in arriving at the estimates.

13.10 Repeat Problem 13.9 for the single- and double-balanced mixers as shown in Figures 13.9(b) and 13.12, respectively.

13.11 Repeat Problem 13.9 for the double-balanced and doubly double-balanced mixers as shown in Figures 13.12 and 13.13, respectively.

13.12 Consider a doubly double-balanced CMOS RFIC mixer as shown in Figure 13.13. Is it worthwhile to consider this mixer for systems as compared to a doubled-balanced CMOS RFIC mixer? If not in general, then is there any specific case under which it should be used? Qualitatively and (quantitatively if possible) point out the characteristics to back-up your comments.

13.13 Design a passive CMOS double-balanced quad-FET ring mixer as shown in Figure 13.14 using any available 0.18-μm CMOS process. The RF frequency is 10–11 GHz and the LO frequency is 9.5 GHz. There are no other required specifications for this mixer. However, you are required to try your best to achieve best possible performance. Both circuit and EM simulators need to be used in the design. Draw the schematic with all element values. Prepare the layout of the mixer ready for tape-out for fabrication. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance. Discuss the design of the RF, LO, and IF baluns (or transformers) and appropriate filters. Simulate the conversion loss as a function of the transistor's gate width and, from which, discuss and choose the gate width for optimum conversion loss. Plot and discuss the mixer performance including conversion loss, noise figure, IIP3, $P_{1\,\mathrm{dB}}$, isolation between RF, LO, and IF ports, return loss, LO power, and output spectrum at the IF port.

13.14 The image rejection (IR) in image-reject mixers can be approximated as

$$\mathrm{IR} = -10\log\left[\frac{1 - 2\sqrt{\Delta A}\cos(\Delta\phi) + \Delta A}{1 + 2\sqrt{\Delta A}\cos(\Delta\phi) + \Delta A}\right]$$

where ΔA and $\Delta\phi$ represent the amplitude and phase imbalance, respectively.

a) Compute and plot the image rejection (in decibel) as a function of the amplitude imbalance from 0 to 10 dB for the phase imbalance of 1°, 2°, 5°, and 10°. Discuss the results.

b) Compute and plot the image rejection (in decibel) as a function of the phase imbalance from 0° to 20° for the amplitude imbalance of 1, 2, 5, and 10 dB. Discuss the results.

13.15 The existence of the image signal increases the noise figure of mixers. The noise-figure increase ΔF can be estimated as

$$\Delta F = 10\log(1 + \mathrm{IR})$$

where IR is the image-rejection level. Compute and plot the noise-figure increase (in decibel) in mixers as a function of the image rejection (in decibel) from 0 to 30 dB. Discuss the results.

13.16 Design a CMOS distributed mixer based on the schematic as shown in Figure 13.17 using any available 0.18-μm CMOS process. The number of sections (mixer cores) used in the synthetic transmission lines can be different from 3 as used in Figure 13.17. The RF frequency is 5–20 GHz, the LO frequency is 4.5–19.5 GHz, and the IF frequency is 500 MHz. There are no other required specifications for this mixer. However, you are required to try your best to achieve best possible performance. Both circuit and EM simulators need to be used in the design. Draw the schematic with all element values. Prepare the layout of the mixer ready for tape-out for fabrication. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance. Discuss the design in details. Simulate the conversion gain as a function of the transistor's gate width and, from which, discuss and choose the gate width for optimum conversion gain. Plot and discuss

the mixer performance including conversion gain, noise figure, IIP3, $P_{1\,\mathrm{dB}}$, isolation between RF, LO, and IF ports, return loss, LO power, DC power consumption, and output spectrum at the IF port.

13.17 Repeat Problem 13.16 for a distributed mixer based on the schematic as shown in Figure 13.16 with and without LO and IF baluns.

13.18 As can be recognized in this chapter, the gain cell of distributed mixers plays a crucial role in achieving high gain and low noise figure for the mixers. Derive a new gain cell different from what included in Figures 13.16 and 13.17 and study its gain and noise figure as compared to the gain cell used in Figure 13.16 (i.e., the lower transistor connecting to the RF line) and in Figure 13.17 (i.e., the cascade gain cell shown in Figure 13.18c). Repeat Problem 13.17 with the newly derived gain cell.

13.19 Design a CMOS double-balanced Gilbert mixer using any available 0.18-μm CMOS process with and without RF, LO, and IF baluns. The RF frequency is 9–10 GHz, the LO frequency is 8.9–9.9 GHz, and the IF frequency is 100 MHz. There are no other required specifications for this mixer. However, you are required to try your best to achieve best possible performance. Both circuit and EM simulators need to be used in the design. Draw the schematic with all element values. Prepare the layout of the mixer ready for tape-out for fabrication. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance. Discuss the design in details. Simulate the conversion gain as a function of the transistor's gate width and, from which, discuss and choose the gate width for optimum conversion gain. Plot and discuss the mixer performance including conversion gain, noise figure, IIP3, $P_{1\,\mathrm{dB}}$, isolation between RF, LO, and IF ports, return loss, LO power, DC power consumption, and output spectrum at the IF port.

13.20 Repeat Problem 13.19 for a CMOS singly-balanced mixer.

13.21 Repeat Problem 13.19 for a CMOS single-ended mixer.

13.22 Consider the single-ended mixer as shown in Figure 13.24. Assume the RF input voltage is $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}} \sin(2\pi f_{\mathrm{RF}} t)$ and the LO signal is a square-pulse signal described in Eq. (13.24). Derive the output voltage $V_{\mathrm{IF}}(t)$ and provide a discussion on the signals appearing at the mixer's output port.

13.23 Consider the single-ended mixer as shown in Figure 13.24. Assume the RF input voltage is $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}} \sin(2\pi f_{\mathrm{RF}} t)$ and the LO signal is a CW signal having voltage $V_{\mathrm{LO}}(t) = V_{\mathrm{LO}} \sin(2\pi f_{\mathrm{LO}} t)$. Derive the output voltage $V_{\mathrm{IF}}(t)$ and provide a discussion on the signals appearing at the mixer's output port.

13.24 Consider the single-balanced mixer as shown in Figure 13.25. Assume the RF input voltage is $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}} \sin(2\pi f_{\mathrm{RF}} t)$ and the LO signal is a square-pulse signal described in Eq. (13.24). Derive the output voltage $V_{\mathrm{IF}}(t)$ and provide a discussion on the signals appearing at the mixer's output port.

13.25 Consider the single-balanced mixer as shown in Figure 13.25. Assume the RF input voltage is $V_{\mathrm{RF}}(t) = V_{\mathrm{RF}} \sin(2\pi f_{\mathrm{RF}} t)$ and the LO signal is a CW signal having voltage $V_{\mathrm{LO}}(t) = V_{\mathrm{LO}} \sin(2\pi f_{\mathrm{LO}} t)$. Derive the output voltage $V_{\mathrm{IF}}(t)$ and provide a discussion on the signals appearing at the mixer's output port.

13.26 Consider the double-balanced mixer as shown in Figure 13.26. Derive Eq. (13.39) using a different approach by determining the individual output voltages emerged at the drains of M2a+, M2a−, M2b+, and M2b− and, from which, determining the voltages at ports IF+ and IF− and hence that between IF+ and IF−.

13.27 Consider a doubly double-balanced mixer core, as can be extracted from Figure 13.13, with differential RF and LO input signals. Assume the RF and LO are sinusoidal and square waves, respectively. Derive expressions for the output voltage and conversion gain. Discuss the results and compare to those of the double-balanced mixer.

13.28 Rederive Eqs. (13.43), (13.45), (13.47), and (13.48) for the effective transconductance and Eqs. (13.44) and (13.46) for the input impedance when the resistance (R_v) of a via-hole at the source of the transistor is not negligible.

13.29 Derive Eq. (13.42) based on a small-signal equivalent circuit model for MOSFET.

13.30 Consider a sampling mixer operating with a sampling frequency (f_s) of 1 MHz. Assume the time delay (T_d) is 200 ps, period of the received RF pulse signal (T_R) is 1 μs, number of samples or steps (N) is 256, and each sample is averaged from 16 samples over τ_s.

a) Determine the duration of the sample time-window (i.e., the total time it takes to sample the received signal in order to produce one down-converted signal).

b) The sampling action is done every 1 μs. Assume that the maximum timing stability is 10% of the sampling time delay, calculate the timing stability in nanoseconds. Does this timing stability impose a very strict design for a timing circuit?

c) What is the reconstruction time for reconstructing the down-converted signal in one cycle?

13.31 Consider a sampling mixer using a sampling pulse having a sampling rate of 20 MHz. The time delay used in the sampling process is $T_d = 100$ ps. The number of steps taken is $N = 512$, and each sample is averaged from values taken in 32 steps over the sampling pulse's duration. The period of the received RF signal is 0.1 μs. Assume the maximum stability is 5% of the sampling delay. Calculate:

a) Timing stability.

b) Sample time-window's width.

c) Reconstruction time.

13.32 According to the theory of sampling receivers discussed in this chapter, the time delay T_d has a maximum value of $1/2B$, where B is the receiver's bandwidth. Compute the maximum time delay allowed for B of $1, 2, \ldots, 9, 10$ GHz.

13.33 Consider a sampling mixer having bandwidth (B) of 1, 5, or 10 GHz and pulse repetition frequency ($f_R = 1/T_R$) of 1, 3, or 5 MHz. For each of the mixer's bandwidth, calculate the corresponding sampling frequency (f_s) of the sampling pulse for $n = 1$, where n is defined in Eq. (13.58), and the ratio f_s/f_R. List the results in a table. From these results, is it better to use a high or low f_R? What about the bandwidth? Provide your rationale.

13.34 Derive Eq. (13.65).

13.35 Design a basic CMOS sampling mixer including a controlled pulse generator using any available 0.18-μm CMOS process. The sampling frequency is 100 MHz and the frequency of the input signal being sampled is from DC-2 GHz. There are no other required specifications for this mixer. However, you are required to try your best to achieve best possible performance. Draw the schematic with all element values. Prepare the layout of the mixer ready for tape-out for fabrication. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance. Discuss the design in details. Plot and discuss the mixer performance including conversion gain, noise figure, $P_{1\,\mathrm{dB}}$, input return loss, DC power consumption, output spectrum, and output voltage waveform.

13.36 Repeat Problem 13.35 for a CMOS integrated sampling mixer based on the circuit diagram presented in Figure 13.45.

CHAPTER 14

SWITCHES

Switches are important components found in many radio frequency (RF) systems from single-pole single-throw (SPST) to multi-pole multi-throw switches. They can be used as a stand-alone component or integrated within subsystems or systems – for instance, transmit/receive (T/R) switches used in systems to accommodate the operation of transmitters and receivers sharing a common antenna. Switches are also used as an integral part of a component to enable the component to perform a particular function such as an SPST switch in an RF pulse-former component to modulate a continuous-wave (CW) signal to form an RF pulse signal. More complex multi-pole multi-throw switches can be employed to achieve various system functions such as transmission and reception of dual-polarized signals using a single dual-polarized antenna. Besides the fundamental function of switching seen in commonly used switches, other functions such as filtering and attenuation can also be incorporated into a switch to form a multi-function switch such as a filter switch or a switching attenuator. This chapter discusses the fundamentals and analyses of switches, and the design of SPST and T/R switches for RFICs. It also addresses ultra-wideband distributed switches, ultra-high-isolation switches, and switches implementing filtering functions. It is noted that the design of CMOS RF switches for low insertion loss, high isolation, high linearity, large power handling, wide bandwidth, and/or fast speed is more challenging than their counterparts employing III–V semiconductors such as GaAs transistors due to the low mobility, high substrate conductivity, low break-down voltage, and various parasitic parameters of Si-based processes.

14.1 FUNDAMENTALS OF SWITCHES

14.1.1 Switch Operation

Switches function based on the on- and off-sate of the employed semiconductor devices (diodes or transistors). For discussion purposes without loss of generality, we will use MOSFET unless otherwise noted. At low frequencies, MOSFET exhibits a very small resistance between the drain and source terminals when the DC-biased voltage applied to the gate–source is higher than the threshold voltage (on state) and a very large resistance when the gate–source voltage is lower than the threshold voltage (off state). It is this difference in impedances between the on- and off-state that switch operation relies on. The function and performance

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

of switches are based on the difference in the reflection of signals instead of the dissipation of signals. As MOSFETs and hence switches are typically operated as a passive device, there is no DC power dissipated in switches. The switch operation is therefore dependent upon the operation or, specifically, passive models of MOSFET under its on- and off-state. These passive models are covered in Section 9.2.4.

The simplest switch type is SPST which is operated as on and off states between the switch's two ports. In more complex switches, the operation is typically based on the on-state between the input port and one of the output ports while the paths between the input and the remaining output ports are turned off. For instance, a T/R switch, or equivalently a single-pole double-throw (SPDT) switch, operates under on-state between the antenna port and the transmitter port and off-state between the antenna and receiver ports in the transmission mode while, in the reception mode, its antenna-receiver port is on and the antenna-transmitter port is off. However, in more complex switches, several paths between ports can be turned on simultaneously while the remaining paths between other ports are turned off. Figure 14.1 shows an example of such switches which connects ports 1 and 2 to an antenna's ports (e.g., vertical and horizontal ports of a dual-polarized antenna), port 4 to a transmitter (TX), and ports 3 and 5 to receivers 1 and 2 (RX1 and RX2). This switch can be operated under various modes depending on the polarization of the radiating signal. For instance, under the horizontally polarized transmission, the transmitter transmits a signal to port 4, which is connected to port 1 under the on-state between 1 and 4, which is then radiated by the antenna's horizontal polarization port, while the paths between other ports (4–2, 1–3, and 2–5) are open. The horizontally polarized transmitting signal hits a target and produces two return signals: one having horizontal polarization and another having vertical polarization. Under the reception mode corresponding to the horizontally polarized transmitting signal, the paths between 1–3 and 2–5 are switched on, while 4–1 and 4–2 are off, which allow the two return signals received at ports 1 and 2 to be directed to the respective receivers.

14.1.2 Important Parameters

14.1.2.1 Insertion Loss. Typical switches are operated as passive devices and, hence, suffer losses. This loss, typically known as insertion loss, just like other passive components such as filters, is important for the use of switches in systems. This importance is more pronounced when a switch is used in front of a low-noise amplifier (LNA) such as a T/R switch, as the receiver noise figure is increased by an amount approximately equal to the insertion loss. This is especially important in the millimeter-wave regime, where the IL of a millimeter-wave switch is relatively high while high gain and low noise figure of a millimeter-wave LNA are difficult to obtain. It is noted that designing a switch to operate in an active mode or including a gain stage (acting as an amplifier) in a switch to overcome the loss, while may improve the performance in certain use such as for transmitters, would degrade the performance for receivers due to increased noise figure.

14.1.2.2 Isolation. Isolation in switches measures how well the two ports in a switch are isolated from each other when the switch between these ports is off. In the design of switches, normally, it is relatively difficult to achieve high isolation than low IL. Moreover, in some applications, isolation is more important than IL. For instance, in a T/R switch, the isolation between the transmitter port and antenna port (when the switch

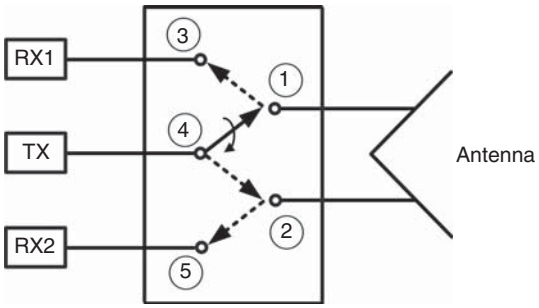


Figure 14.1. Multi-pole multi-throw switch.

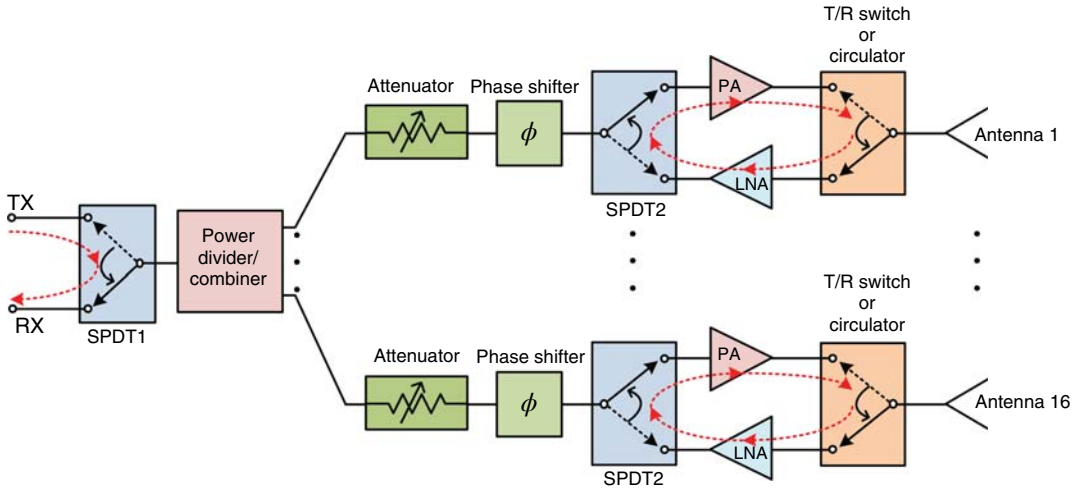


Figure 14.2. System employing various SPDT and T/R switches. Dotted loop/arrow shows RF leakage problems under reception mode.

between the transmitter and antenna is turned off) and the isolation between the receiver port and antenna port (when the switch between the receiver and antenna is turned off) are normally more important than the IL, particularly the isolation between the transmitter and antenna since, while the loss may be overcome by increasing the gain or output power of the power amplifier (PA), poor isolation causes large signal leakage from the transmitter to receiver, resulting in undesired effects. To illustrate the importance of isolation, we consider a system as shown in Figure 14.2. This system represents a phased array, which employs SPDT and T/R switches to share common components. This system suffers serious RF leakage problems under the reception mode due to the limited isolation of the employed SPDT and T/R switches. The first RF leakage occurs at SPDT1 where the transmitted signal from the transmitter (TX) port leaks to the receiver (RX) port and adds to the received signal, causing detrimental effects such as distorting the received signal and reduced system's dynamic range and linearity. The second RF leakage occurs in the loop along the T/R switch, LNA, SPDT2, and PA, where the received signal from the antenna, after being amplified by the LNA, leaks to the transmitting path through SPDT2, amplified by the PA, and then goes back to the antenna. This leaking signal may be larger than the received signal, hence distorting the received signal at the antenna.

14.1.2.3 Linearity. Linearity is an important figure of merit for switches, particularly when it is used in place where high RF power exists such as in the path connecting an antenna and a transmitter in a T/R switch. Due to the high power typically used in transmitters, it is crucial that the T/R switch has high linearity. It is important to note that switches are typically designed to operate in their passive state, which is essentially linear state, to maintain the purity of transmitting or receiving signals, except a reduction in signal amplitude. Nonlinear behavior, however, occurs with signals of sufficiently large power which causes undesirable intermodulation products at different frequencies to occur, and needs to be avoided. The switch performance needs to be as linear as possible in order to produce high-quality signal. The linearity measures the quality of switches in handling RF signals or, specifically, it measures the ability of switches to operate without distorting RF signals at high input power levels. The important aspect in signal handling is the maximum tolerable level of the input RF signal which is described by a power compression magnitude, typically 1 dB. As the RF signal is increased beyond the linear range, reaching a large-signal level, the IL is no longer constant, and the output power begins to saturate and then reduces as the input power is increased, causing switches to compress. The 1-dB power compression point ($P_{1\,\mathrm{dB}}$) measures the departure of switches from their linear operation in which the output signal power increases linearly with the input RF signal power. The 1-dB input power compression is the input RF signal level at which the actual output RF power is 1 dB less than the linearly increased output RF level. The 1-dB input power compression point is typically used as the maximum input power that a switch can handle.

14.1.2.4 Switching Time. Switching time is important for switches requiring high-speed switching. The switching speed of a MOSFET switch depends not only on the MOSFET itself but also on the circuit environment in which the MOSFET is embedded. For instance, the gate resistor typically used for DC biasing at the gate affects the switching time and should be optimized to reduce the (RC) time constant, and hence resulting in a faster speed, while keeping a minimum effect on the IL. The switching time can be determined by simulating or measuring the RF pulse of a CW signal produced by turning the switch on and off. The rising and falling times between 10% and 90% of the maximum voltage are typically used for the switching time. These times are normally different for practical switches.

One important remark that needs to be made at this point is that switches should be designed, simulated, and measured under the same conditions at which they are intended to be operated. For instance, for switches intended for high-power applications such as T/R switches, the performance of a switch such as IL and isolation needs to be simulated and measured under both small- and large-signal conditions.

14.2 ANALYSIS OF SWITCHING MOSFET

Switches are formed by transistors connected in series only, shunt only or alternating series, and shunt in single or multiple paths. Therefore, a series transistor, shunt transistor, or combined series and shunt transistors serve as the basic building element for various switches. Among the various types of switches, the SPST switch having a single path between the input and output ports represents a basic element found in other switches with multi-path like SPDT switches having two paths between the input and two output ports. As such, the analysis and performance of a series transistor, shunt transistor, or combination of series and shunt transistors and SPST switches, as well the design of SPST switches, form the foundation for the analysis and design of switches.

14.2.1 Analysis of Shunt Transistor

Figure 14.3 shows a MOSFET connected in shunt which functions as a very basic SPST switch. The MOSFET is represented by an admittance having two different values: $Y_{\text{on}} = G_{\text{on}} + jB_{\text{on}}$ and $Y_{\text{off}} = G_{\text{off}} + jB_{\text{off}}$ under the MOSFET's on- and off-state, respectively. The switch passes and rejects a signal under the MOSFET's off-state (corresponding to large impedance) and on-state (corresponding to small impedance), respectively. The scattering parameter S_{21} of the switch can be obtained from (7.59) as

$$S_{21} = \frac{2Y_o}{Y + 2Y_o} = \frac{1}{1 + \overline{Y}/2} \tag{14.1}$$

where Y represents Y_{on} or Y_{off}, Y_o is the terminating source/load admittance, and $\overline{Y}$ is the normalized value of Y with respect to Y_o. The IL or isolation of the switch is given as $|S_{21}|^2$.

The insertion loss (IL) is obtained from (14.1) as

$$\text{IL} = \frac{4Y_o^2}{(G_{\text{off}} + 2Y_o)^2 + B_{\text{off}}^2} \tag{14.2}$$

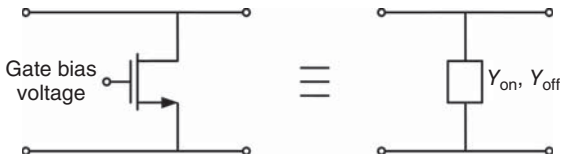


Figure 14.3. Shunt MOSFET switch.

which, upon simplification, gives

$$\mathrm{IL} = \left(1 + \overline{G}_{\mathrm{off}} + \frac{1}{4}\overline{G}_{\mathrm{off}}^2 + \frac{1}{4}\overline{B}_{\mathrm{off}}^2\right)^{-1} \tag{14.3}$$

Similarly, the isolation (ISO) can also be obtained from (14.1) as

$$\mathrm{ISO} = \left(1 + \overline{G}_{\mathrm{on}} + \frac{1}{4}\overline{G}_{\mathrm{on}}^2 + \frac{1}{4}\overline{B}_{\mathrm{on}}^2\right)^{-1} \tag{14.4}$$

where $\overline{Y}_{\mathrm{on(off)}} = \overline{G}_{\mathrm{on(off)}} + j\overline{B}_{\mathrm{on(off)}}$ are normalized with respect to Y_o. We can see that under the perfect condition of the MOSFET's off-state (infinite impedance) and on-state (zero impedance), the IL is zero and the isolation approaches infinity, as expected theoretically.

14.2.2 Analysis of Series Transistor

Figure 14.4 shows a MOSFET connected in series which also represents a very basic SPST switch. The MOSFET is represented by an impedance having two different values: $Z_{\mathrm{on}} = R_{\mathrm{on}} + jX_{\mathrm{on}}$ and $Z_{\mathrm{off}} = R_{\mathrm{off}} + jX_{\mathrm{off}}$ under the MOSFET's on- and off-state, respectively. The switch passes and rejects a signal under the MOSFET's on-state (corresponding to small impedance) and off-state (corresponding to large impedance), respectively. The scattering parameter S_{21} of the switch can be derived as

$$S_{21} = \frac{2Z_o}{Z + 2Z_o} = \frac{1}{1 + \overline{Z}/2} \tag{14.5}$$

where Z represents Z_{on} or Z_{off}, Z_o is the terminating source/load impedance, and $\overline{Z}$ is the normalized value of Z with respect to Z_o. The insertion loss (IL) and isolation (ISO) can be derived from (14.5) as

$$\mathrm{IL} = \left(1 + \overline{R}_{\mathrm{on}} + \frac{1}{4}\overline{R}_{\mathrm{on}}^2 + \frac{1}{4}\overline{X}_{\mathrm{on}}^2\right)^{-1} \tag{14.6}$$

and

$$\mathrm{ISO} = \left(1 + \overline{R}_{\mathrm{off}} + \frac{1}{4}\overline{R}_{\mathrm{off}}^2 + \frac{1}{4}\overline{X}_{\mathrm{off}}^2\right)^{-1} \tag{14.7}$$

where $\overline{Z}_{\mathrm{on(off)}} = \overline{R}_{\mathrm{on(off)}} + j\overline{X}_{\mathrm{on(off)}}$ are normalized with respect to Z_o. Under the perfect condition of the MOSFET's on-state (zero impedance) and off-state (infinite impedance), the IL is zero and the isolation approaches infinity as expected.

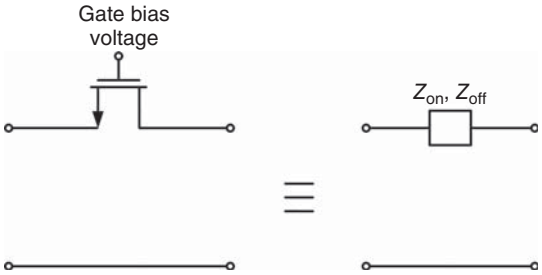


Figure 14.4. Series MOSFET switch.

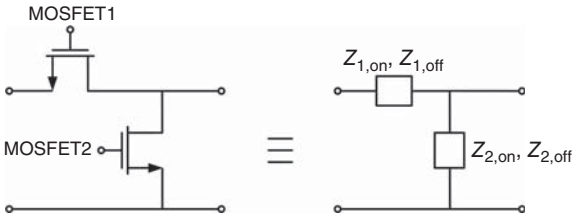


Figure 14.5. MOSFET switch employing both series and shunt MOSFETs.

14.2.3 Analysis of Combined Series and Shunt Transistors

Figure 14.5 shows a switch consisting of two MOSFETs connected in series and shunt. The MOSFETs are represented by impedances $Z_{i,\text{on}} = R_{i,\text{on}} + jX_{i,\text{on}}$ and $Z_{i,\text{off}} = R_{i,\text{off}} + jX_{i,\text{off}}$, with i being 1 and 2 corresponding to MOSFET 1 and 2, respectively, under the MOSFETs' on- and off-state, respectively. The switch passes a signal under the respective on-state of MOSFET 1 and off-state of MOSFET 2, and rejects a signal under the respective off-state of MOSFET 1 and on-state of MOSFET 2. The scattering parameter S_{21} of the switch can be derived as

$$S_{21} = \frac{2}{2 + Z_1/Z_2 + Z_1/Z_o + Z_o/Z_2} = \frac{2}{2 + \overline{Z}_1/\overline{Z}_2 + \overline{Z}_1 + 1/\overline{Z}_2} \tag{14.8}$$

where Z_i represents $Z_{i,\text{on}}$ or $Z_{i,\text{off}}$, Z_o is the terminating source/load impedance, and $\overline{Z}_i$ is the normalized impedance with respect to Z_o. The insertion loss (IL) and isolation (ISO) can be derived from (14.8) as

$$\text{IL} = \frac{4(\overline{R}_{2,\text{off}}^2 + \overline{X}_{2,\text{off}}^2)}{[1 + \overline{R}_{1,\text{on}} + \overline{R}_{2,\text{off}}(2 + \overline{R}_{1,\text{on}}) - \overline{X}_{1,\text{on}}\overline{X}_{2,\text{off}}]^2 + [\overline{X}_{1,\text{on}}(1 + \overline{R}_{2,\text{off}}) + \overline{X}_{2,\text{off}}(2 + \overline{R}_{1,\text{on}})]^2} \tag{14.9}$$

and

$$\text{ISO} = \frac{4(\overline{R}_{2,\text{on}}^2 + \overline{X}_{2,\text{on}}^2)}{[1 + \overline{R}_{1,\text{off}} + \overline{R}_{2,\text{on}}(2 + \overline{R}_{1,\text{off}}) - \overline{X}_{1,\text{off}}\overline{X}_{2,\text{on}}]^2 + [\overline{X}_{1,\text{off}}(1 + \overline{R}_{2,\text{on}}) + \overline{X}_{2,\text{on}}(2 + \overline{R}_{1,\text{off}})]^2} \tag{14.10}$$

where $\overline{Z}_{i,\text{on(off)}} = \overline{R}_{i,\text{on(off)}} + j\overline{X}_{i,\text{on(off)}}$ are normalized with respect to Z_o. Under the perfect condition of the MOSFETs' on-state (zero impedance) and off-state (infinite impedance), the IL is zero and the isolation approaches infinity as expected.

14.2.4 Selection of MOSFET

One of the most important tasks in the design of RF switches is the selection of proper MOSFETs. In general, series transistors dominate the IL, whereas shunt devices primarily dictates the isolation of RF switches. The IL and isolation of a switch depend on the size (i.e., gate width for a given gate length) of the series and shunt transistors. Increasing the size of the series transistor reduces the on-state resistance, hence reducing the IL at low frequencies, but, at the same time, also increases the parasitic capacitances, resulting in increase in IL and reduction of isolation at high frequencies. On the other hand, increasing the size of the shunt transistor enhances the isolation, but also increases the IL. Typically, switches employing only shunt MOSFETs are preferred at high frequencies, whereas, both series and shunt MOSFETs are used for very broadband switches down to DC to achieve good isolation in the low- and high-frequency regions of the switch's operating frequency range, respectively. It is noted that the on-state resistance is scaled down approximately as L/W with L and W being the gate length and width, respectively, in advanced sub-micrometer CMOS processes. In general, optimum IL and isolation do not occur for the same MOSFET. A trade-off between low IL and high isolation needs to be considered in the switch design. This analysis can be easily performed

for any MOSFET using its on- and off-state impedances along with (14.3), (14.4), (14.6), (14.7), and (14.9), (14.10), or a commercially available program along with the MOSFET's equivalent-circuit model. The results can be displayed in graphs to enable convenient selections of devices. As examples to illustrate a procedure for selecting proper sizes for series transistors, shunt transistors, and combined series and shunt transistors for given IL and/or isolation, we use MOSFETs having gate length of 0.18 μm, finger's width of 4 μm, frequency of 35 GHz, and bias voltages of 0 (off-state) and 1.8 V (on-state). The IL and isolation of switches using different sizes of these 0.18 μm transistors, represented by the numbers of fingers, are simulated using the transistor's equivalent-circuit models for three cases: shunt MOSFET, series MOSFET, and combined series and shunt MOSFETs.

14.2.4.1 Shunt MOSFET. Figure 14.6 shows the IL and isolation for 0.18-μm MOSFETs connected in shunt and series versus the numbers of fingers (and hence transistor size). As can be seen from Figure 14.6, a device size for shunt connection can be conveniently chosen to produce certain IL and/or isolation. For example, an IL of around 1 dB and isolation of around 10.5 dB can be achieved using a 0.18-μm shunt MOSFET with 60-μm gate width.

14.2.4.2 Series MOSFET. As can be seen from Figure 14.6, a device size for series connection can be conveniently chosen to produce certain IL and/or isolation. For example, an IL of around 1.5 dB and isolation of 4 dB can be achieved using a 0.18-μm series MOSFET with 60-μm gate width.

Comparison between MOSFET connected in series and shunt in Figure 14.6 shows that a series MOSFET generally has higher IL and lower isolation than a shunt MOSFET. In practice, shunt MOSFETs are typically used to enhance the isolation in switches.

14.2.4.3 Combined Series and Shunt MOSFETs. As Figure 14.6 shows, it is very straightforward to generate the IL and isolation curves for MOSFETs and to select a proper device. On the other hand, the generation of the IL and isolation contours for combined shunt and series MOSFETs and the selection of proper shunt and series devices, while are still simple, are more involved due to the fact that there are two transistors and the need to take into account their interactions with respect to IL and isolation. To facilitate the selection of optimum sizes for the series and shunt MOSFETs for a given IL and/or isolation using charts, contours corresponding to fixed IL and isolation are plotted as the sizes of the series and shunt transistors are simultaneously swept at an interested frequency. Figure 14.7 shows the IL and isolation contours for a pair of series and shunt 0.18-μm MOSFETs at 35 GHz versus the finger numbers of the series and shunt transistors. As

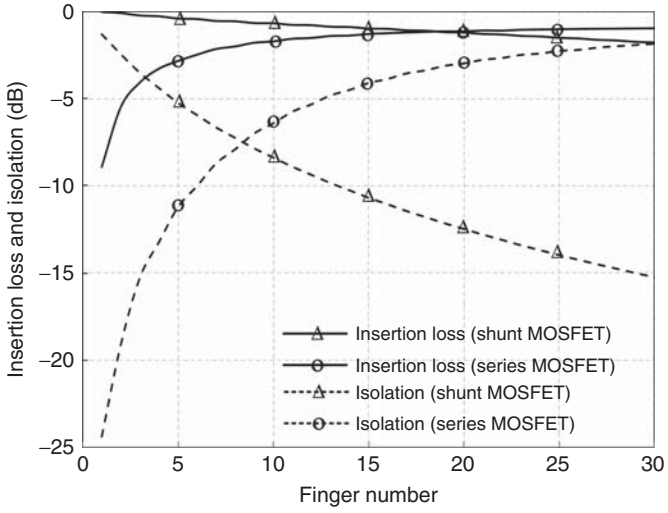


Figure 14.6. Insertion loss and isolation for series and shunt 0.18-μm MOSFETs of different sizes at 35 GHz. The finger width is 4 μm and the gate-bias resistor is 10 kΩ.

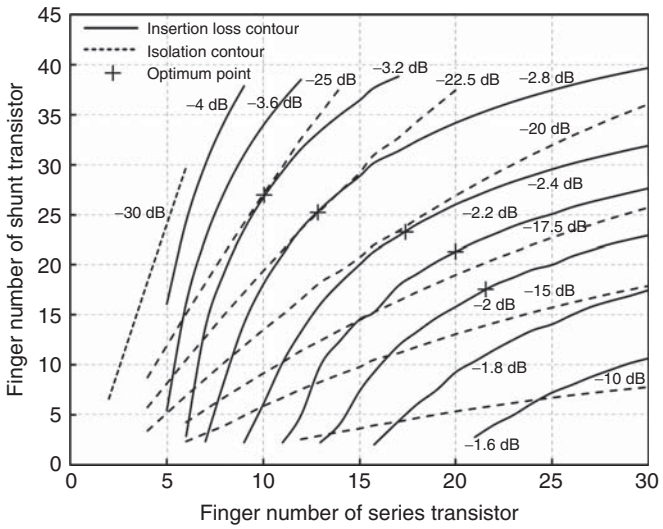


Figure 14.7. Insertion loss and isolation contours for combined series and shunt 0.18-μm MOSFETs of different sizes at 35 GHz. The finger width is 4 μm and the gate-bias resistor is 400 Ω.

can be seen, the IL contours have nearly parabolic form, while those for the isolation behave almost linearly. We then expect mathematically that the IL and isolation contours would cross or be tangential to each other. These phenomena are indeed seen through the simulation results, some of which are presented in Figure 14.7. The series and shunt transistor sizes at the crossing point between an IL curve and an isolation curve would produce the corresponding IL and isolation. This crossing point, however, is not the optimum point corresponding to the lowest possible IL and highest possible isolation. Depending on the values of the IL and isolation at that point, it may correspond to arbitrary loss and isolation, minimum IL and non-maximum isolation, or maximum isolation and non-minimum IL. The tangential point between the contours represents the optimum point for the device sizes at which either the highest isolation is obtained for a given IL or the lowest IL is achieved for a given isolation. It is expected that, for a given insertion-loss (or isolation) contour, there is always an isolation (or insertion-loss) contour that is tangential to it. Therefore, for a given IL or isolation, it is always possible that an optimum point exists, and hence the optimum sizes for the series and shunt transistors can be chosen at that point. The plus (+) on the curves in Figure 14.7 shows some of these optimum points. The sizes of transistors corresponding to these optimum points should be chosen for the switch design. For example, the optimum sizes for the series and shunt MOSFETs, corresponding to 20-dB isolation and 2.4-dB IL for a switch, are 64 and 84 μm, respectively, according to Figure 14.7.

14.2.5 Design Consideration for Improved Insertion Loss and Isolation

For a given MOSFET, the impedances (or admittances) under the on- and off-state are fixed at a particular frequency. Equations (14.3), (14.4), (14.6), (14.7), and (14.9), (14.10) show that the IL and isolation depend on these impedances. Reducing R_{on} and increasing R_{off} using additional MOSFETs in parallel and series, respectively, can improve the isolation for shunt and series switches, respectively. This, however, may be undesirable due to increased IL. On the other hand, the susceptance (for shunt MOSFET) and reactance (for series MOSFET) can be optimized to improve the switch's performance in IL and isolation. For instance, considering a MOSFET connected in shunt and assuming, for illustration purpose, it is represented by a large resistance R_{off} in parallel with a capacitor C_{off} for off-state and a small resistance R_{on} in series with an inductor L_{on} for on-state. To achieve minimum IL for the switch when the MOSFET is off, the effect of C_{off} is removed by connecting an inductor in parallel with the MOSFET to form a parallel resonator with C_{off} at the design frequency. This added inductor does not degrade the isolation when the MOSFET is on since the isolation is primarily determined by the small impedance across the device. On the other hand, to achieve maximum isolation for the switch when the MOSFET is on, the effect of L_{on} is removed by connecting a capacitor in

series with L_{on} to form a series resonator at the design frequency. The added capacitor is usually very large and therefore only slightly affects the impedance, and hence the IL, under off-state and, often, it is needed for DC-blocking purpose.

14.3 SPST SWITCHES

SPST switch is the basic element for all switches and hence plays an important part in the design, analysis, and performance of switches. Practical MOSFETs have finite impedances under their on- and off-state, which limit the amount of isolation that can be obtained. These finite impedances make it difficult to achieve high isolation using a single MOSFET, particularly in the high RF range where MOSFET's parasitics such as inductors have large reactances. In order to achieve high isolation or isolation over a wide bandwidth, multiple MOSFETs, placing next to each other or far apart, are needed. It is noted that, to maintain symmetry in circuit layouts which is crucial for some circuits, especially at millimeter-wave frequencies, use of two parallel transistors at the same location may be needed, and the use of symmetrical transmission lines such as coplanar waveguide (CPW) facilitates this kind of connection. SPST switches employing only a series MOSFET, a shunt MOSFET, and a pair of series/shunt MOSFETs are covered in Section 14.2. The operating principle and analysis of multi-MOSFET SPST switches is similar to that for a two-MOSFET SPST switch. Therefore, without loss of generality, we will specifically focus the analysis on two-MOSFET SPST switches in this section.

14.3.1 SPST Switch Employing Two Parallel MOSFETs

Figure 14.8 shows an SPST switch consisting of two MOSFETs connected in parallel. The total admittance of the two MOSFETs is $2Y_{on} = 2(G_{on} + jB_{on})$ and $2Y_{off} = 2(G_{off} + jB_{off})$ under on- and off-state, respectively, where Y_{on} and Y_{off} are the respective admittances for one MOSFET. The increased admittance under on-state leads to enhanced isolation, yet at a cost of increased IL due to increased off-state admittance. From (14.4), the isolation is obtained as

$$\mathrm{ISO} = (1 + 2\overline{G}_{on} + \overline{G}_{on}^2 + \overline{B}_{on}^2)^{-1} \tag{14.11}$$

Typical MOSFETs used for switches have sufficiently large G_{on} or, correspondingly, reasonable isolation for shunt connection. Comparing (14.11) to (14.4) shows that using two parallel MOSFETs can improve the isolation by 6 dB approximately. This isolation can be further improved by using additional transistors (e.g., around 12 dB using three MOSFETs). The result can be generalized as

$$\mathrm{ISO(dB)} \simeq \mathrm{ISO'(dB)} - (N-1)6\,\mathrm{dB} \tag{14.12}$$

where ISO(dB) and ISO′(dB) are the isolation (in decibel) for N parallel MOSFETs and one MOSFET, respectively. It is noted that (14.12) is valid only for MOSFETs having sufficiently large G_{on}. Such devices are, in fact, always used for switch design; therefore, (14.12) is normally valid in practice.

It is recognized that the IL for two devices in parallel also increases. For good MOSFETs and/or at low frequencies, the increase in loss may be tolerated. However, for less quality devices and/or high operating frequencies, the loss increase may be significant for a switch to operate properly and hence the use of multiple MOSFETs may be infeasible for switch design.

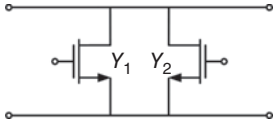


Figure 14.8. SPST switch employing two MOSFETs next to each other in parallel. Y_1 and Y_2 represent the admittances of the MOSFETs.

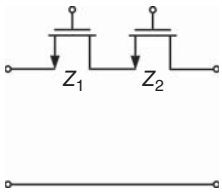


Figure 14.9. SPST switches employing two series MOSFETs. Z_1 and Z_2 represent the impedances of the MOSFETs.

14.3.2 SPST Switch Employing Two Series MOSFETs

Considering two MOSFETs connected in series as shown in Figure 14.9, we also find out that, for good MOSFETs typically used for switches, the isolation is enhanced by about 6 dB with respect to a single series MOSFET. However, the IL for two series MOSFETs is increased more than that for two parallel MOSFETs, making this configuration less attractive than two parallel MOSFETs.

14.3.3 SPST Switch Employing Two Series and Two Shunt MOSFETs

Figure 14.10 shows an SPST switch consisting of two MOSFETs in series and two MOSFETs in shunt. We can expect that the isolation is enhanced by about 6 dB as compared to the case when only one series MOSFET and one shunt MOSFET are connected. However, the IL also increases and may not be tolerable.

14.3.4 SPST Switch Using Impedance or Admittance Inverters

We learn in Section 5.5 that an ideal impedance (or admittance) inverter is characterized by its constant image impedance K (or admittance $J = 1/K$) and $\pm 90°$ transmission phase at all frequencies (or, specifically, over an interested frequency range). We also learn that an ideal impedance (or admittance) inverter would transform an impedance Z_b or admittance $Y_b = 1/Z_b$ connected at one port of the inverter into another impedance Z_a or admittance $Y_a = 1/Z_a$ at the other port of

$$Z_a = \frac{K^2}{Z_b} \tag{14.13}$$

$$Y_a = \frac{J^2}{Y_b} \tag{14.14}$$

respectively. Equations (14.13) and (14.14) show that a short or open would be transformed into an open or short via an ideal inverter over an operating bandwidth, respectively. This result can help facilitate the enhancement of isolation in the switch design. It is recalled from Section 7.5 that impedance and admittance inverters are basically the same.

We recall that when two MOSFETs are placed in parallel or series at the same location, the overall isolation is increased by about 6 dB due to the resulting half or double impedance of the MOSFETs in parallel or series, respectively. Now we consider an SPST switch consisting of two series (or shunt) MOSFETs separated by an ideal impedance or admittance inverter. We can see that the switch would undergo isolations at two different locations: one isolation at the first MOSFET location and another at the second MOSFET location

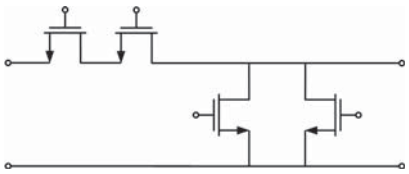


Figure 14.10. SPST switches employing two series and two shunt MOSFETs.

that combines the isolation of the image of the first MOSFET, transformed through the inverter, and the isolation of the second MOSFET. We can then expect that the isolation of the switch would be approximately equal to the isolations of the two individual MOSFETs plus 6 dB, which is indeed the result that we will obtain from an analysis as follows.

14.3.4.1 SPST Switch with Shunt MOSFETs and Admittance Inverters. Figure 14.11 shows an SPST switch formed by two shunt MOSFETs spaced apart by an admittance inverter. The chain matrix of this switch is obtained, utilizing the chain matrix of an ideal inverter in (5.198), as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ Y_1 & 1 \end{bmatrix} \begin{bmatrix} 0 & \pm j/J \\ \pm jJ & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Y_2 & 1 \end{bmatrix} = \begin{bmatrix} \pm jY_2/J & \pm j/J \\ \pm jY_1Y_2/J \pm jJ & \pm jY_1/J_1 \end{bmatrix} \tag{14.15}$$

The IL or isolation can be expressed using the conversion formula from Table 7.1 as

$$|S_{21}|^2 = \frac{4}{|A + BY_o + C/Y_o + D|^2} \tag{14.16}$$

Making use of (14.15) and (14.16), we can express the IL and isolation as

$$\text{IL} = \frac{4}{|(1/J)[Y_o + J^2/Y_o + Y_{1,\text{off}} + Y_{2,\text{off}} + Y_{1,\text{off}}Y_{2,\text{off}}/Y_o]|^2} \tag{14.17}$$

$$\text{ISO} = \frac{4}{|(1/J)[Y_o + J^2/Y + Y_{1,\text{on}} + Y_{2,\text{on}} + Y_{1,\text{on}}Y_{2,\text{on}}/Y_o]|^2} \tag{14.18}$$

Expanding (14.17) and (14.18) gives

$$\text{IL} = \frac{4}{\begin{array}{l}(1/J^2)[Y_o + J^2/Y_o + G_{1,\text{off}} + G_{2,\text{off}} + (G_{1,\text{off}}G_{2,\text{off}} - B_{1,\text{off}}B_{2,\text{off}})/Y_o]^2 \\ +(1/J^2)[B_{1,\text{off}} + B_{2,\text{off}} + (G_{1,\text{off}}B_{2,\text{off}} + G_{2,\text{off}}B_{1,\text{off}})/Y_o]^2\end{array}} \tag{14.19}$$

$$\text{ISO} = \frac{4}{\begin{array}{l}(1/J^2)[Y_o + J^2/Y_o + G_{1,\text{on}} + G_{2,\text{on}} + (G_{1,\text{on}}G_{2,\text{on}} - B_{1,\text{on}}B_{2,\text{on}})/Y_o]^2 \\ +(1/J^2)[B_{1,\text{on}} + B_{2,\text{on}} + (G_{1,\text{on}}B_{2,\text{on}} + G_{2,\text{on}}B_{1,\text{on}})/Y_o]^2\end{array}} \tag{14.20}$$

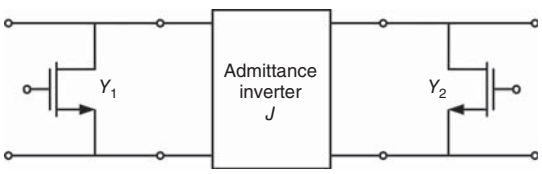


Figure 14.11. SPST switch consisting of two shunt MOSFETs and an admittance inverter. Y_1 and Y_2 represent the MOSFETs' admittances.

By normalizing the admittances to the image admittance J, we can rewrite (14.19) and (14.20) as

$$\mathrm{IL} = \frac{4}{[\overline{Y}_o + 1/\overline{Y}_o + \overline{G}_{1,\mathrm{off}} + \overline{G}_{2,\mathrm{off}} + (\overline{G}_{1,\mathrm{off}}\overline{G}_{2,\mathrm{off}} - \overline{B}_{1,\mathrm{off}}\overline{B}_{2,\mathrm{off}})/\overline{Y}_o]^2 + [\overline{B}_{1,\mathrm{off}} + \overline{B}_{2,\mathrm{off}} + (\overline{G}_{1,\mathrm{off}}\overline{B}_{2,\mathrm{off}} + \overline{G}_{2,\mathrm{off}}\overline{B}_{1,\mathrm{off}})/\overline{Y}_o]^2} \tag{14.21}$$

$$\mathrm{ISO} = \frac{4}{[\overline{Y}_o + 1/\overline{Y}_o + \overline{G}_{1,\mathrm{on}} + \overline{G}_{2,\mathrm{on}} + (\overline{G}_{1,\mathrm{on}}\overline{G}_{2,\mathrm{on}} - \overline{B}_{1,\mathrm{on}}\overline{B}_{2,\mathrm{on}})/\overline{Y}_o]^2 + [\overline{B}_{1,\mathrm{on}} + \overline{B}_{2,\mathrm{on}} + (\overline{G}_{1,\mathrm{on}}\overline{B}_{2,\mathrm{on}} + \overline{G}_{2,\mathrm{on}}\overline{B}_{1,\mathrm{on}})/\overline{Y}_o]^2} \tag{14.22}$$

where $\overline{Y}_{1,\mathrm{on(off)}} = \overline{G}_{1,\mathrm{on(off)}} + j\overline{B}_{1,\mathrm{on(off)}}$ and $\overline{Y}_{2,\mathrm{on(off)}} = \overline{G}_{2,\mathrm{on(off)}} + j\overline{B}_{2,\mathrm{on(off)}}$ are the normalized admittances with respect to J for MOSFET 1 and MOSFET 2 under on(off) state, respectively, and $\overline{Y}_o = 1/\overline{Z}_o$ are the normalized source (or load) admittance with respect to J, with $\overline{Z}_o$ being the normalized source (or load) impedance.

In practice, identical MOSFETs are typically used, and the isolation from (14.22) becomes

$$\mathrm{ISO} = \frac{4}{[\overline{Y}_o + 1/\overline{Y}_o + 2\overline{G}_{\mathrm{on}} + (\overline{G}_{\mathrm{on}}^2 - \overline{B}_{\mathrm{on}}^2)/\overline{Y}_o]^2 + 4(\overline{B}_{\mathrm{on}} + \overline{G}_{\mathrm{on}}\overline{B}_{\mathrm{on}}/\overline{Y}_o)^2} \tag{14.23}$$

where $\overline{Y}_{\mathrm{on(off)}} = \overline{G}_{\mathrm{on(off)}} + j\overline{B}_{\mathrm{on(off)}}$ are the MOSFET's normalized admittances with respect to J. To enable extraction of some interesting results, we simplify the formulation, without loss of generality, by assuming that J is equal to the source (or load) admittance υ and, accordingly, reduce (14.23) to

$$\mathrm{ISO} = \frac{4}{(2 + 2\overline{G}_{\mathrm{on}} + \overline{G}_{\mathrm{on}}^2 - \overline{B}_{\mathrm{on}}^2)^2 + 4[\overline{B}_{\mathrm{on}}(1 + \overline{G}_{\mathrm{on}})]^2} \tag{14.24}$$

Typical MOSFETs used for switches have relatively large conductance and small parasitic susceptance under on-state which correspond to reasonable isolation when a MOSFET is connected in shunt. For instance, if we assume that a MOSFET has $\overline{G}_{\mathrm{on}} = 10$ and $\overline{B}_{\mathrm{on}} = 1$ then, according to (14.4), the isolation for a shunt MOSFET is −15.6 dB which should be attainable. Therefore, for typical MOSFETs used for RF switching, it is reasonable to assume that

$$\overline{G}_{\mathrm{on}}^2 - \overline{B}_{\mathrm{on}}^2 \gg 2 + 2\overline{G}_{\mathrm{on}} \tag{14.25}$$

which is equivalent to

$$\overline{G}_{\mathrm{on}} \gg 1 + \sqrt{1 + (2 + \overline{B}_{\mathrm{on}}^2)} \tag{14.26}$$

Note that the negative root in (14.26) is discarded since it is redundant. We can deduce from (14.26) that

$$\overline{G}_{\mathrm{on}} \gg 1 \tag{14.27}$$

With (14.25) and (14.27), we can then rewrite (14.24) as

$$\mathrm{ISO} \simeq \frac{4}{(\overline{G}_{\mathrm{on}}^2 - \overline{B}_{\mathrm{on}}^2)^2 + 4(\overline{G}_{\mathrm{on}}\overline{B}_{\mathrm{on}})^2} = \frac{4}{|(\overline{G}_{\mathrm{on}} + j\overline{B}_{\mathrm{on}})(\overline{G}_{\mathrm{on}} + j\overline{B}_{\mathrm{on}})|^2} = \frac{4}{|\overline{Y}_{\mathrm{on}}^2|^2} \tag{14.28}$$

We can deduce through (14.28) that the isolation of an SPST switch employing two different shunt MOSFETs separated by an admittance inverter, as shown in Figure 14.13, can be approximately given as

$$\mathrm{ISO} \simeq \frac{4}{|\overline{Y}_{1,\mathrm{on}}\overline{Y}_{2,\mathrm{on}}|^2} \tag{14.29}$$

The isolation for a shunt MOSFET is obtained from (14.1) as

$$\text{ISO} = \frac{1}{|1 + \overline{Y}_{\text{on}}/2|^2} \simeq \frac{4}{|\overline{Y}_{\text{on}}|^2} \tag{14.30}$$

making use of the fact that $\overline{Y}_{\text{on}}$ is relatively large for typical MOSFETs under on-state. The isolation of the SPST switch can now be expressed approximately as

$$\text{ISO} \simeq \frac{1}{4}\text{ISO}_1\text{ISO}_2 \tag{14.31}$$

through comparison of (14.29) and (14.30), where ISO_1 and ISO_2 are the isolations of MOSFET 1 and MOSFT 2, respectively. In terms of decibel (dB), we can write

$$\text{ISO(dB)} \simeq \text{ISO}_1\text{(dB)} + \text{ISO}_2\text{(dB)} - 6\,\text{dB} \tag{14.32}$$

Equation (14.32) shows that the isolation of an SPST switch employing two different MOSFETs connected in shunt and spaced apart by an admittance inverter is equal to the sum of the isolations of individual MOSFETs and 6 dB, demonstrating a significant enhancement in isolation as compared to two MOSFETs connected in parallel at the same location as given in (14.12).

The isolation can be further enhanced by using additional MOSFETs separated by admittance or impedance inverters as shown in Figure 14.12. Following the same analysis, we can derive the approximate isolation for an SPST switch consisting of N shunt MOSFETs separated by $(N-1)$ admittance inverters as

$$\text{ISO} \simeq \frac{1}{4^{(N-1)}}\prod_{n=1}^{N}\text{ISO}_n \tag{14.33}$$

or

$$\text{ISO}(dB) \simeq \sum_{n=1}^{N}\text{ISO}_n\text{(dB)} - (N-1)6\,\text{dB} \tag{14.34}$$

where ISO_n $(n = 1, 2, \ldots, N)$ is the isolation of the nth MOSFET. We can see that a substantial improvement in the isolation can be achieved by using multiple inverters separating MOSFETs. This, however, also increases the IL due to the added losses resulting from additional MOSFETs and inverters. The use of multiple inverters between MOSFETs (and the resulting increased isolation and IL) is conceptually analogous to the use of multiple inverters between resonators in band-pass filters (and the resultant increase in rejection and loss).

14.3.4.2 SPST Switch with Series MOSFETs and Impedance Inverters. Similarly, we can also derive equations for the IL and isolation of SPST switches employing series MOSFETs spaced apart by impedance

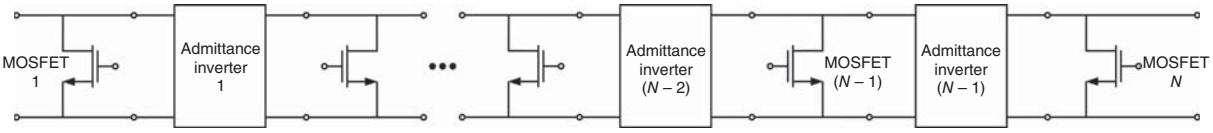


Figure 14.12. SPST switch employing N shunt MOSFETs and $(N-1)$ admittance or impedance inverters.

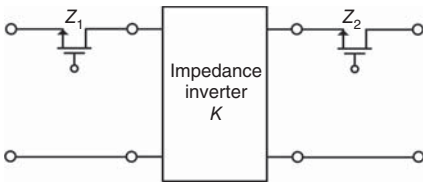


Figure 14.13. SPST switch consisting of two series MOSFETs and an impedance inverter. Z_1 and Z_2 represent the MOSFETs' impedances.

inverters. We consider an SPST switch consisting of two series MOSFETs separated by an impedance inverter as shown in Figure 14.13. The chain matrix of this switch is given as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & Z_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & \pm jK \\ \pm j/K & 0 \end{bmatrix} \begin{bmatrix} 1 & Z_2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \pm jZ_1/K & \pm jK \pm jZ_1Z_2/K \\ \pm j/K & \pm jZ_2/K \end{bmatrix} \tag{14.35}$$

The IL and isolation can be derived using (14.16) and (14.35) as

$$\text{IL} = \frac{4K^2}{[Z_o + R_{1,\text{on}} + R_{2,\text{on}} + (K^2 + R_{1,\text{on}}R_{2,\text{on}} - X_{1,\text{on}}X_{2,\text{on}})/Z_o]^2 + [X_{1,\text{on}} + X_{2,\text{on}} + (R_{1,\text{on}}X_{2,\text{on}} + R_{2,\text{on}}X_{1,\text{on}})/Z_o]^2} \tag{14.36}$$

$$\text{ISO} = \frac{4K^2}{[Z_o + R_{1,\text{off}} + R_{2,\text{off}} + (K^2 + R_{1,\text{off}}R_{2,\text{off}} - X_{1,\text{off}}X_{2,\text{off}})/Z_o]^2 + [X_{1,\text{off}} + X_{2,\text{off}} + (R_{1,\text{off}}X_{2,\text{off}} + R_{2,\text{off}}X_{1,\text{off}})/Z_o]^2} \tag{14.37}$$

By normalizing the impedances to the image impedance K, we can rewrite (14.36) and (14.37) as

$$\text{IL} = \frac{4}{[\overline{Z}_o + 1/\overline{Z}_o + \overline{R}_{1,\text{on}} + \overline{R}_{2,\text{on}} + (\overline{R}_{1,\text{on}}\overline{R}_{2,\text{on}} - \overline{X}_{1,\text{on}}\overline{X}_{2,\text{on}})/\overline{Z}_o]^2 + [\overline{X}_{1,\text{on}} + \overline{X}_{2,\text{on}} + (\overline{R}_{1,\text{on}}\overline{X}_{2,\text{on}} + \overline{R}_{2,\text{on}}\overline{X}_{1,\text{on}})/\overline{Z}_o]^2} \tag{14.38}$$

$$\text{ISO} = \frac{4}{[\overline{Z}_o + 1/\overline{Z}_o + \overline{R}_{1,\text{off}} + \overline{R}_{2,\text{off}} + (\overline{R}_{1,\text{off}}\overline{R}_{2,\text{off}} - \overline{X}_{1,\text{off}}\overline{X}_{2,\text{off}})/\overline{Z}_o]^2 + [\overline{X}_{1,\text{off}} + \overline{X}_{2,\text{off}} + (\overline{R}_{1,\text{off}}\overline{X}_{2,\text{off}} + \overline{R}_{2,\text{off}}\overline{X}_{1,\text{off}})/\overline{Z}_o]^2} \tag{14.39}$$

where $\overline{Z}_{1,\text{on(off)}} = \overline{R}_{1,\text{on(off)}} + j\overline{X}_{1,\text{on(off)}}$ and $\overline{Z}_{2,\text{on(off)}} = \overline{R}_{2,\text{on(off)}} + j\overline{X}_{2,\text{on(off)}}$ are the normalized impedances with respect to K for MOSFET 1 and MOSFET 2 under on (off) state, respectively, and $\overline{Z}_o$ are the normalized source (or load) impedance with respect to K.

For two identical MOSFETs, the isolation from (14.39) is reduced to

$$\text{ISO} = \frac{4}{[\overline{Z}_o + 1/\overline{Z}_o + 2\overline{R}_{\text{off}} + (\overline{R}_{\text{off}}^2 - \overline{X}_{\text{off}}^2)/\overline{Z}_o]^2 + 4[\overline{X}_{\text{off}} + \overline{R}_{\text{off}}\overline{X}_{\text{off}}/\overline{Z}_o]^2} \tag{14.40}$$

where $\overline{Z}_{\text{on(off)}} = \overline{R}_{\text{on(off)}} + j\overline{X}_{\text{on(off)}}$ are the MOSFET's normalized impedances with respect to K. Eq. (14.40) becomes, assuming $K = Z_o$:

$$\text{ISO} = \frac{4}{(2 + 2\overline{R}_{\text{off}} + \overline{R}_{\text{off}}^2 - \overline{X}_{\text{off}}^2)^2 + 4[\overline{X}_{\text{off}}(1 + \overline{R}_{\text{off}})]^2} \tag{14.41}$$

Typical MOSFETs used for switches have relatively large resistance and small parasitic reactance under off-state which correspond to reasonable isolation when a MOSFET is connected in series. This enables us to assume that

$$\overline{R}_{\text{off}}^2 - \overline{X}_{\text{off}}^2 \gg 2 + 2\overline{R}_{\text{off}} \tag{14.42}$$

or

$$\overline{R}_{\text{off}} \gg 1 + \sqrt{1 + (2 + \overline{X}_{\text{off}}^2)} \tag{14.43}$$

Making use (14.42) and (14.43), we can rewrite (14.41) as

$$\begin{aligned} \text{ISO} &\simeq \frac{4}{(\overline{G}_{\text{on}}^2 - \overline{B}_{\text{on}}^2)^2 + 4(\overline{G}_{\text{on}}\overline{B}_{\text{on}})^2} = \frac{4}{|(\overline{G}_{\text{on}} + j\overline{B}_{\text{on}})(\overline{G}_{\text{on}} + j\overline{B}_{\text{on}})|^2} = \frac{4}{|\overline{Y}_{\text{on}}^2|^2} \\ \text{ISO} &\simeq \frac{4}{(\overline{R}_{\text{off}}^2 - \overline{X}_{\text{off}}^2)^2 + 4(\overline{R}_{\text{off}}\overline{X}_{\text{off}})^2} = \frac{4}{|\overline{Z}_{\text{off}}^2|^2} \end{aligned} \tag{14.44}$$

Using the result in (14.44), we can approximate the isolation of an SPST switch employing two different series MOSFETs separated by an impedance inverter as

$$\text{ISO} \simeq \frac{4}{|\overline{Z}_{1,\text{off}}\overline{Z}_{2,\text{off}}|^2} \tag{14.45}$$

Comparing (14.45) to the isolation of a typical series MOSFET having large off-state impedance $\overline{Z}_{\text{off}}$ obtained from (14.5)

$$\text{ISO} = \frac{1}{|1 + \overline{Z}_{\text{off}}/2|^2} \simeq \frac{4}{|\overline{Z}_{\text{off}}|^2} \tag{14.46}$$

we can approximate the isolation of the SPST switch in terms of the isolations of MOSFET 1 (ISO_1) and MOSFT 2 (ISO_2) as in (14.31) and (14.32), which show enhanced isolation with respect to two MOSFETs connected in series at the same location.

The isolation for an SPST switch consisting of N series MOSFETs separated by $(N-1)$ impedance inverters can be derived as in (14.33) and (14.34), which demonstrates a substantial enhancement in isolation by using multiple inverters separating series MOSFETs. This enhanced isolation, however, is also coupled with an increased IL as expected.

14.3.4.3 SPST Switch with Series/Shunt MOSFETs and Impedance Inverters. We consider an SPST switch consisting of two pairs of series/shunt MOSFETs separated by an impedance inverter as shown in Figure 14.14. The chain matrix of this switch is given as

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= \begin{bmatrix} 1 & Z_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Y_2 & 1 \end{bmatrix} \begin{bmatrix} 0 & \pm jK \\ \pm\frac{j}{K} & 0 \end{bmatrix} \begin{bmatrix} 1 & Z_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Y_2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \pm j\left\{KY_2 + Z_1\left[KY_2^2 + \frac{1}{K}(1 + Z_1Y_2)\right]\right\} & \pm j\left[K + Z_1\left(KY_2 + \frac{Z_1}{K}\right)\right] \\ \pm j\left[KY_2^2 + \frac{1}{K}(1 + Z_1Y_2)\right] & \pm j\left(KY_2 + \frac{Z_1}{K}\right) \end{bmatrix} \end{aligned} \tag{14.47}$$

The IL and isolation can be derived using (14.16) and (14.47) as

$$|S_{21}|^2 = \frac{4K^2}{E^2 + F^2} \tag{14.48}$$

where

$$E = K^2[2G_2 + Y_o(1 + R_1G_2) + (G_2^2 - B_2^2)(R_1 + Z_o) - X_1B_2(2G_2 + Y_o)] + Z_o + 2R_1 + (R_1 + Z_o)(R_1G_2 - X_1B_2) - X_1(R_1B_2 + X_1G_2) + Y_o(R_1^2 - X_1^2) \quad (14.49)$$

$$F = K^2\{2B_2[1 + G_2(Z_o + R_1)] + X_1(G_2^2 - B_2^2) + Y_o(X_1G_2 + R_1B_2)\} + 2X_1(1 + R_1Y_o) + X_1(R_1G_2 + X_1B_2) + (Z_o + R_1)(X_1G_2 + R_1B_2) \quad (14.50)$$

with $Z_1 = R_{1,\text{on}} + jX_{1,\text{on}}$ and $Y_2 = G_{2,\text{off}} + jB_{2,\text{off}}$ for IL and $Z_1 = R_{1,\text{off}} + jX_{1,\text{off}}$ and $Y_2 = G_{2,\text{on}} + jB_{2,\text{on}}$ for isolation. By normalizing the impedances and admittances to the image impedance K and admittance J, respectively, we can rewrite (14.48)–(14.50) for the IL and isolation as

$$\text{IL} = \frac{4}{\overline{E}_{\text{IL}}^2 + \overline{F}_{\text{IL}}^2} \quad (14.51)$$

$$\text{ISO} = \frac{4}{\overline{E}_{\text{ISO}}^2 + \overline{F}_{\text{ISO}}^2} \quad (14.52)$$

where

$$\overline{E}_{\text{IL}} = \overline{Z}_o + 2(\overline{R}_{1,\text{on}} + \overline{G}_{2,\text{off}}) + \overline{Y}_o(1 + \overline{R}_{1,\text{on}}\overline{G}_{2,\text{off}} + \overline{R}_{1,\text{on}}^2 - \overline{X}_{1,\text{on}}^2) - \overline{X}_{1,\text{on}}\overline{B}_{2,\text{off}}(2\overline{G}_{2,\text{off}} + \overline{Y}_o) + (\overline{R}_{1,\text{on}} + \overline{Z}_o)(\overline{G}_{2,\text{off}}^2 - \overline{B}_{2,\text{off}}^2 + \overline{R}_{1,\text{on}}\overline{G}_{2,\text{off}} - \overline{X}_{1,\text{on}}\overline{B}_{2,\text{off}}) - \overline{X}_{1,\text{on}}(\overline{R}_{1,\text{on}}\overline{B}_{2,\text{off}} + \overline{X}_{1,\text{on}}\overline{G}_{2,\text{off}}) \quad (14.53)$$

$$\overline{F}_{\text{IL}} = 2\overline{B}_{2,\text{off}} + (\overline{G}_{2,\text{off}} + \overline{X}_{1,\text{on}}\overline{G}_{2,\text{off}} + \overline{R}_{1,\text{on}}\overline{B}_{2,\text{off}})(\overline{Z}_o + \overline{R}_{1,\text{on}}) + \overline{Y}_o(\overline{X}_{1,\text{on}}\overline{G}_{2,\text{off}} + \overline{R}_{1,\text{on}}\overline{B}_{2,\text{off}}) + \overline{Y}_o(\overline{X}_{1,\text{on}}\overline{G}_{2,\text{off}} + \overline{R}_{1,\text{on}}\overline{B}_{2,\text{off}}) + \overline{X}_{1,\text{on}}(2 + 2\overline{Y}_o\overline{R}_{1,\text{on}} + \overline{R}_{1,\text{on}}\overline{G}_{2,\text{off}} + \overline{X}_{1,\text{on}}\overline{B}_{2,\text{off}} + \overline{G}_{2,\text{off}}^2 - \overline{B}_{2,\text{off}}^2) \quad (14.54)$$

$$\overline{E}_{\text{ISO}} = \overline{Z}_o + 2(\overline{R}_{1,\text{off}} + \overline{G}_{2,\text{on}}) + \overline{Y}_o(1 + \overline{R}_{1,\text{off}}\overline{G}_{2,\text{on}} + \overline{R}_{1,\text{off}}^2 - \overline{X}_{1,\text{off}}^2) - \overline{X}_{1,\text{off}}\overline{B}_{2,\text{on}}(2\overline{G}_{2,\text{on}} + \overline{Y}_o) + (\overline{R}_{1,\text{off}} + \overline{Z}_o)(\overline{G}_{2,\text{on}}^2 - \overline{B}_{2,\text{on}}^2 + \overline{R}_{1,\text{off}}\overline{G}_{2,\text{on}} - \overline{X}_{1,\text{off}}\overline{B}_{2,\text{on}}) - \overline{X}_{1,\text{off}}(\overline{R}_{1,\text{off}}\overline{B}_{2,\text{on}} + \overline{X}_{1,\text{off}}\overline{G}_{2,\text{on}}) \quad (14.55)$$

$$\overline{F}_{\text{ISO}} = 2\overline{B}_{2,\text{on}} + (\overline{G}_{2,\text{on}} + \overline{X}_{1,\text{off}}\overline{G}_{2,\text{on}} + \overline{R}_{1,\text{off}}\overline{B}_{2,\text{on}})(\overline{Z}_o + \overline{R}_{1,\text{off}}) + \overline{Y}_o(\overline{X}_{1,\text{off}}\overline{G}_{2,\text{on}} + \overline{R}_{1,\text{off}}\overline{B}_{2,\text{on}}) + \overline{Y}_o(\overline{X}_{1,\text{off}}\overline{G}_{2,\text{on}} + \overline{R}_{1,\text{off}}\overline{B}_{2,\text{on}}) + \overline{X}_{1,\text{off}}(2 + 2\overline{Y}_o\overline{R}_{1,\text{off}} + \overline{R}_{1,\text{off}}\overline{G}_{2,\text{on}} + \overline{X}_{1,\text{off}}\overline{B}_{2,\text{on}} + \overline{G}_{2,\text{on}}^2 - \overline{B}_{2,\text{on}}^2) \quad (14.56)$$

with $\overline{Z}_{1,\text{on(off)}} = \overline{R}_{1,\text{on(off)}} + j\overline{X}_{1,\text{on(off)}}$ and $\overline{Y}_{2,\text{on(off)}} = \overline{G}_{2,\text{on(off)}} + j\overline{B}_{2,\text{on(off)}}$ being the normalized impedance for MOSFET 1 with respect to K and normalized admittance of MOSFET 2 with respect to J under on (off) state, respectively, and $\overline{Z}_o$ and $\overline{Y}_o$ being the normalized source (or load) impedance and admittance with respect to K and J, respectively.

Through a simple but lengthy manipulation, we can approximate the isolation of SPST switches employing typical MOSFETs having large off-state impedance $\overline{Z}_{\text{off}}$ and large on-state admittance $\overline{Y}_{\text{on}}$ as in (14.31) and (14.32), where ISO_1 and ISO_2 are the isolations of the two constituent switches, each consisting of MOSFET 1 in series and MOSFT 2 in shunt. We can also derive that the isolation for an SPST switch consisting of N pairs of series/shunt MOSFETs separated by $(N - 1)$ impedance inverters are approximately given as in (14.33) and (14.34), showing a significant enhancement in isolation which, however, is also coupled with an increase in IL.

14.3.4.4 Impedance and Admittance Inverters. It is apparent that the isolation and IL of SPST switches employing series MOSFETs, shunt MOSFETs or combined series/shunt MOSFETs, and impedance or admittance inverters depend on the on- and off-state impedances of MOSFETs as well as the inverters through their image impedance K, or admittance J, and transmission phase (ideally $\pm 90°$) over the switch's operating frequency range. For given MOSFETs, the design of an SPST switch then depends on K or J and the transmission phase of the impedance or admittance inverters. It is expected that a fixed K or J would not provide maximum isolation and minimum IL at the same time. It is noted that in general K and J vary as a function of frequency - for instance, K and J of the inverters discussed in Section 5.5.1.2. From the circuit design point of view and considering the fact that the isolation is already enhanced through the use of inverters, it may be best that we choose K or J so that an IL as small as possible is achieved. Moreover, an appropriate transmission phase for the inverters needs to be used to provide optimum isolation and IL. For ideal MOSFETs with zero and infinite impedance under on- and off-state, respectively, we can see from the switch topologies shown in Figures 14.13–14.16 that a minimum IL is achieved when $K = Z_o$ or $J = Y_o$ and the transmission phase is $\pm 90°$. Under this condition, the IL of the SPST switch is only slightly larger than the total ILes of individual MOSFETs provided that low-loss inverters are used. For practical MOSFETs having finite on- and off-impedances, however, $J \neq Y_o$ or $K \neq Z_o$. It is noted that the enhanced isolations given in (14.31)–(14.34) are also applicable when $J \neq Y_o$ or $K \neq Z_o$. We also note that isolation in switches occurs due to mismatch in the switches' off-state, which is intentionally imposed in the design. Therefore, better isolation is obtained when additional mismatches are introduced. The IL, however, needs to be checked to make sure that it does not increase beyond a tolerable level due to possible increase in mismatch under on-state even low-loss inverters are used.

As discussed in Section 5.5, there are many kinds of impedance and admittance inverters. The most basic impedance or admittance inverter is a quarter-wavelength transmission line whose characteristic impedance is equal to K and transmission phase is 90°. For MOSFETs behaving closely as a short and an open under on- and off-state, respectively, the characteristic impedance of the quarter-wavelength transmission line is selected to be equal to the source/load impedance Z_o, which results in minimum reflection and hence minimum IL. Practical MOSFETs, however, do not behave as an open or short under its respective off- and on-state, thus necessitating optimization of the characteristic impedance and length of the quarter-wavelength transmission line for desired switch performance. It is noted, however, that a quarter-wavelength transmission line can function only as an impedance or admittance inverter within a narrow frequency range around the quarter-wavelength frequency at which the impedance or admittance inversion is exact. Moreover, a quarter-wavelength transmission-line inverter is relatively long for RFICs, resulting in high chip cost, unless the operating frequency is sufficiently high.

Figures 5.40 and 5.41 show some practical impedance and admittance inverters based on lumped elements or combined lumped elements and transmission lines. These inverters have broader bandwidth than the quarter-wavelength transmission-line inverter. Lumped-element inverters are particularly attractive at frequencies where the lumped elements' characteristics are preserved and/or their size is significantly smaller than a quarter-wavelength transmission-line inverter.

The impedance inverters in Figures 5.40(a) and (b) contain negative inductance and capacitance, respectively, and are suitable for use with series MOSFETs that can absorb the negative elements to form circuits

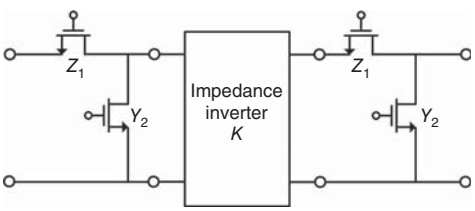


Figure 14.14. SPST switch consisting of two pairs of series/shunt MOSFETs and an impedance inverter. Z_1 and Y_2 represent the impedance and admittance of MOSFET 1 and 2, respectively.

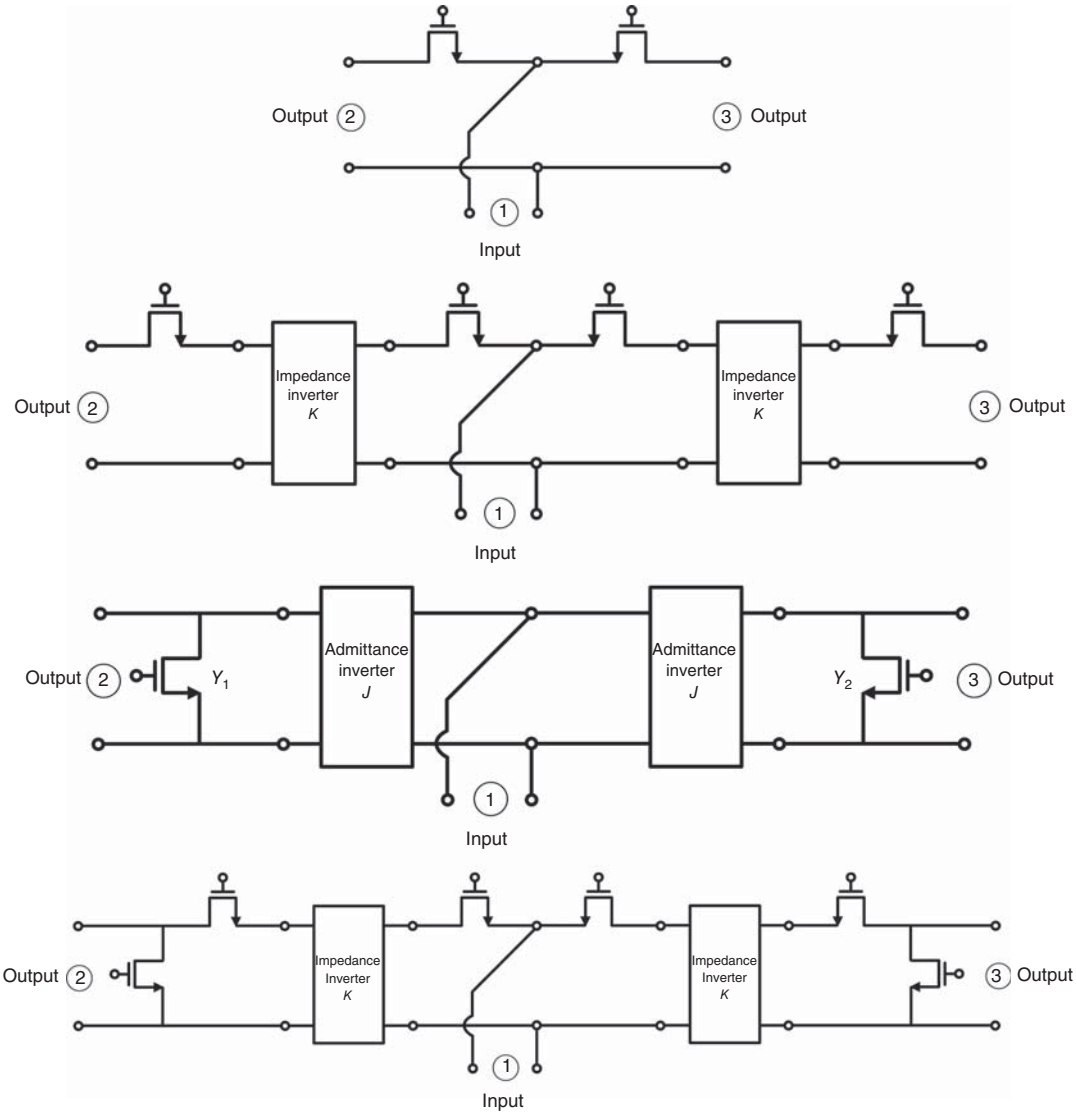


Figure 14.15. SPDT switch topologies.

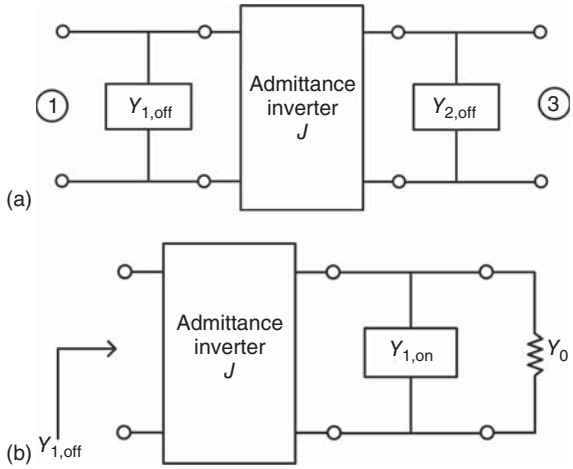


Figure 14.16. (a) On-path equivalent circuit of the SPDT switch in Figure 14.15(c) and (b) equivalent circuit looking into on-state MOSFET 1 through the admittance inverter. The source admittance at input port 1 and load admittances at output ports 2 and 3 are Y_o.

having only positive components. The impedance inverters shown in Figures 5.40(c) and (d) contain transmission lines having negative and positive electrical lengths, respectively, that can be subtracted from or added to adjacent transmission lines having the same characteristic impedance.

The admittance inverters in Figures 5.41(a) and (b) contain negative inductance and capacitance, respectively, and are suitable for use with shunt MOSFETs that can absorb the negative elements to form circuits having only positive components. The admittance inverters shown in Figures 5.41(c) and (d) have transmission lines of positive and negative electrical lengths, respectively, that can be added to or subtracted from adjacent transmission lines having the same characteristic impedance.

The choice of a particular inverter is dictated by the switch configuration (series MOSFETs only, shunt MOSFETs only or combined series/shunt MOSFETs, and lumped elements or transmission lines) and the characteristics of the employed MOSFETs at the operating frequencies. For instance, the inverter in Figure 5.41(b) may be suitable for the shunt switch configuration when the MOSFETs have parasitic capacitances greater than C. Another important criterion for selecting an inverter is the dependence of the image impedance or admittance and transmission phase on frequency. These parameters should not vary substantially over the design frequency range, which puts a constraint on inverters for wide-band design.

14.4 SPDT SWITCHES

SPDT switches are one of the most widely used switches in RF systems, particularly as T/R switches for transceivers connecting to a single antenna. A SPDT switch has one input and two output branches and, in operation, one output branch is turned on while the other output branch is turned off. A SPDT switch hence requires two SPST switches effectively connected in parallel, one in each output branch with at least one MOSFET, to function. As such, the design of SDT switches is based on that for SPST switches and can consist of series, shunt, or combined series and shunt MOSFETs with or without impedance (or admittance) inverters in each output arm.

14.4.1 SPDT Switch Topologies

SPDT switches work based on the principle that one output arm is on while the other is off. Figure 14.15 shows some possible topologies for SPDT switches based on that principle and the SPST switches discussed in the previous sections.

Figure 14.15(a) shows the simplest SPDT switch topology consisting of two series MOSFETs, one in each output branch. One of the MOSFETs is turned on whereas the other is turned off simultaneously to allow an RF signal to pass from the input to the output corresponding to the on-MOSFET. Figure 14.15(b) shows an SPDT switch utilizing series MOSFETs and impedance inverters (e.g., quarter-wavelength inverters). This switch is operated similar to that in Figure 14.15(a) but with higher isolation between the input and the output corresponding to the off-state. Figure 14.15(c) shows an SPDT switch utilizing shunt MOSFETs and admittance inverters. When one MOSFET is turn on and the other is turned off, the RF signal would pass from the input to the output corresponding to the off-MOSFET while being isolated from the output port containing the on-MOSFET. To understand this operation, we assume ideal MOSFETs which behave as an open and short under off- and on-state, respectively, and ideal admittance inverters having K equal to the (input) source admittance. The on-MOSFET, represented by a short, appears as an open at the other end of the admittance inverter which is directly connected to the input port, while the off-MOSFET behaves as an open, thus allowing RF signals to pass from the input port to the output port at the off-MOSFET. It is noted that, while the impedance inverters are optional for the SPDT switch in Figure 14.15(b), the admittance inverters in the SPDT switch in Figure 14.15(c) are required as they enable the input signal to be routed to either output port. These inverters, however, limit the bandwidth of the SPDT switches due to the frequency-dependent image admittance and transmission phase. The SPDT switch in Figure 14.15(d) employs a pair of series/shunt MOSFET, based on the SPST switch in Figure 14.14, and a series MOSFET in each output arm. In one output arm, the series and shunt MOSFETs are turned on and off, respectively, whereas the series and shunt

MOSFETs in the other arm are turned off and on, respectively. These functions are executed simultaneously, enabling an RF signal to be transmitted from the input port to output port corresponding to the on series MOSFET and off shunt MOSFET, while keeping it from appearing at the other output port. It is noted that, only a series MOSFET can be used at the other end of the impedance inverter; a shunt MOSFET cannot be used since it would short-circuit the input port (assume ideal MOSFET), hence preventing the RF signal from going to an output port.

14.4.2 SPDT Switch Analysis

The analysis of SPDT switches such as those in Figure 14.15 is in general very similar and can be adapted from that of SPST switches presented in Section 14.3 considering the fact that, under on- or off-mode, the corresponding transmission or isolation path behaves as an SPST switch under on- and off-mode, respectively. Therefore, without loss of generality, we choose to analyze the SPDT switch as shown in Figure 14.15(c). This switch includes admittance inverters to increase the isolation as analyzed earlier. We begin the analysis by assuming that MOSFET 2 represented by Y_2 is off and MOSFET 1 represented by Y_1 is on, which implies that an RF signal entering the input port 1 would be directed to the output port 3 (transmission or on path) and (ideally) no signal would appear at the output port 2 (isolation or off path).

We consider the IL for the SPDT switch first and draw in Figure 14.16 an equivalent circuit for the on-path and that looking into MOSFET 1 under on-state through the admittance inverter. Figure 14.16(a) is essentially equivalent to Figure 14.11 that represents an SPST switch. Hence, the IL of the SPDT switch from port 1 to port 3 can be obtained using (14.21) derived for the IL of the SPST in Figure 14.11 as

$$\mathrm{IL}_{31} = \frac{4}{\begin{array}{c}[\overline{Y}_o + 1/\overline{Y}_o + \overline{G}_{1,\mathrm{off}} + \overline{G}_{2,\mathrm{off}} + (\overline{G}_{1,\mathrm{off}}\overline{G}_{2,\mathrm{off}} - \overline{B}_{1,\mathrm{off}}\overline{B}_{2,\mathrm{off}})/\overline{Y}_o]^2 \\ +[\overline{B}_{1,\mathrm{off}} + \overline{B}_{2,\mathrm{off}} + (\overline{G}_{1,\mathrm{off}}\overline{B}_{2,\mathrm{off}} + \overline{G}_{2,\mathrm{off}}\overline{B}_{1,\mathrm{off}})/\overline{Y}_o]^2\end{array}} \tag{14.57}$$

where $\overline{Y}_{1,\mathrm{off}} \equiv Y_{1,\mathrm{off}}/J = \overline{G}_{1,\mathrm{off}} + j\overline{B}_{1,\mathrm{off}}$ represents the normalized admittance with respect to J looking into MOSFET 1 under on-state via the admittance inverter, $\overline{Y}_{2,\mathrm{off}} \equiv Y_{2,\mathrm{off}}/J = \overline{G}_{2,\mathrm{off}} + j\overline{B}_{2,\mathrm{off}}$ represents the normalized admittance of MOSFET 2 under off-state with respect to J, and $\overline{Y}_o$ is the normalized source (or load) admittance with respect to J. The IL from the input port 1 to the output port 2 when MOSFET 1 and MOSFET 2 are off and on, respectively, is given by the same equation but with $\overline{Y}_{2,\mathrm{off}} = \overline{G}_{2,\mathrm{off}} + j\overline{B}_{2,\mathrm{off}}$ representing the normalized admittance with respect to J looking into MOSFET 2 under on-state via the admittance inverter, and $\overline{Y}_{1,\mathrm{off}} = \overline{G}_{1,\mathrm{off}} + j\overline{B}_{1,\mathrm{off}}$ representing the normalized admittance of MOSFET 1 under off-state with respect to J.

For the isolation formulation, we show in Figure 14.17 an equivalent circuit for the off-path and that looking into MOSFET 2 under off-state through the admittance inverter. The isolation of the SPDT switch from port 1 to port 2 can be obtained using (14.22) for the isolation of the SPST in Figure 14.13 as

$$\mathrm{ISO}_{21} = \frac{4}{\begin{array}{c}[\overline{Y}_o + 1/\overline{Y}_o + \overline{G}_{1,\mathrm{on}} + \overline{G}_{2,\mathrm{on}} + (\overline{G}_{1,\mathrm{on}}\overline{G}_{2,\mathrm{on}} - \overline{B}_{1,\mathrm{on}}\overline{B}_{2,\mathrm{on}})/\overline{Y}_o]^2 \\ +[\overline{B}_{1,\mathrm{on}} + \overline{B}_{2,\mathrm{on}} + (\overline{G}_{1,\mathrm{on}}\overline{B}_{2,\mathrm{on}} + \overline{G}_{2,\mathrm{on}}\overline{B}_{1,\mathrm{on}})/\overline{Y}_o]^2\end{array}} \tag{14.58}$$

where $\overline{Y}_{1,\mathrm{on}} = \overline{G}_{1,\mathrm{on}} + j\overline{B}_{1,\mathrm{on}}$ represents the normalized admittance of MOSFET 1 under on state with respect to J, and $\overline{Y}_{2,\mathrm{on}} = \overline{G}_{2,\mathrm{on}} + j\overline{B}_{2,\mathrm{on}}$ represents the normalized admittance with respect to J looking into MOSFET 2 under off-state via the admittance inverter. The isolation from the input port 1 to the output port 3 when MOSFET 1 and MOSFET 2 are off and on, respectively, is given by the same equation but with $\overline{Y}_{1,\mathrm{on}} = \overline{G}_{1,\mathrm{on}} + j\overline{B}_{1,\mathrm{on}}$ represents the normalized admittance with respect to J looking into MOSFET 1 under off-state via the admittance inverter, and $\overline{Y}_{2,\mathrm{on}} = \overline{G}_{2,\mathrm{on}} + j\overline{B}_{2,\mathrm{on}}$ represents the normalized admittance of MOSFET 2 under on-state with respect to J.

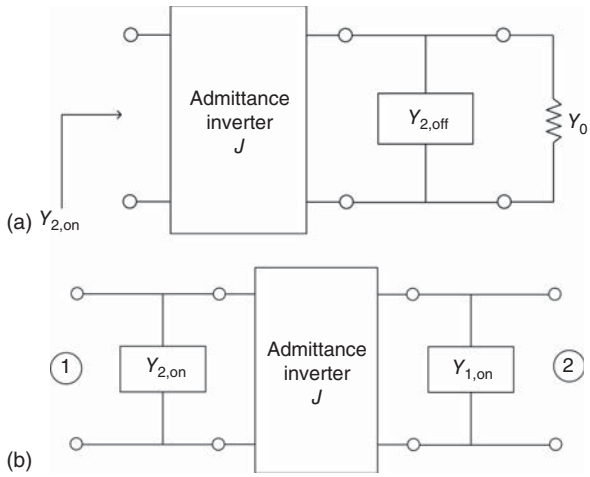


Figure 14.17. (a) Off-path equivalent circuit of the SPDT switch in Figure 14.15(c) and (b) equivalent circuit looking into MOSFET 2 through the admittance inverter. The source admittance at the input port 1 and load admittances at the output ports 2 and 3 are Y_o.

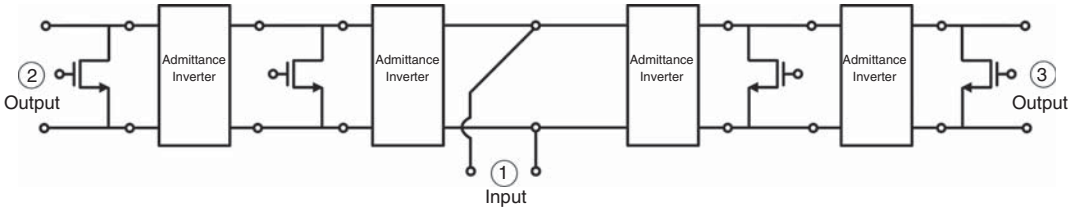


Figure 14.18. SPDT switch employing multiple shunt MOSFETs and admittance inverters.

Similar to the SPST switches employing shunt transistors and admittance inverters discussed in Section 14.3, the isolation of the SPDT switches based on shunt transistors and admittance inverters can be increased by adding more transistors separated by admittance inverters. Figure 14.18 shows an example of such SPDT switches. Expressions for the IL and isolation of this switch can be easily derived following the analysis for the corresponding SPST switches. Using similar analyses for the SPST switches in Section 14.3 and SPDT switch in this section, we can also derive the ILs and isolations for various SPDT switches, such as those in Figure 14.15, as well as for multiple-pole multiple-throw switches such as single-pole triple-throw (SP3T) and double-pole quadruple-throw (DP4T).

14.5 ULTRA-WIDEBAND SWITCHES

Ultra-wideband RF systems are important for communications, radar, and sensing – not only for the existing applications but also for newly emerging and future applications – for instance, multi-function for simultaneous communications and sensing within the same broad band or over different bands. These systems require ultra-wideband switches. As for other wide-band RFICs, various topologies can be conceived for wideband switches. Among them, the distributed switch topologies based on the synthetic transmission lines discussed in Chapter 4, similar to the distributed amplifiers covered in Chapter 11, are some of the most attractive ones since they can produce ultra-wide bandwidths with a relatively simple configuration and design. This section includes two distributed switches: an SPST switch working from DC to 11 GHz and a T/R switch operating from DC to 20 GHz.

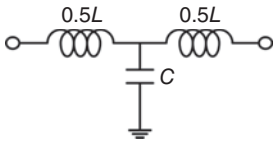


Figure 14.19. Single element of synthetic transmission lines.

14.5.1 Ultra-Wideband SPST Switch

We begin by considering a single element (or cell) of the synthetic transmission lines as shown in Figure 14.19. This synthetic transmission-line element is essentially equivalent to that consisting of only one series inductor L and one shunt capacitor C as seen in Figure 4.16(b), but herein the inductor is decomposed into two inductors, each having a half of the inductance ($L/2$) representing each of the two conductors of the transmission line. The characteristic impedance (Z_o), phase velocity (v_p) and cutoff frequency (f_c) of the synthetic transmission line based on Figure 14.19 are still given by the same equations (4.231), (4.232), and (4.241) or (4.242), respectively. For convenience, these equations are repeated below:

$$Z_o = \sqrt{\frac{L}{C}} \tag{14.59}$$

$$v_p = \frac{1}{\sqrt{LC}} \tag{14.60}$$

and

$$f_c = \frac{1}{\pi\sqrt{LC}} = \frac{1}{\pi C Z_o} \tag{14.61}$$

Figure 14.20 shows the schematic of the ultra-wideband SPST switch [1]. Under the switch's on-state, this topology simulates a synthetic transmission line based on the single synthetic transmission-line element represented in Figure 14.19. Specifically, the shunt transistors under their off-state, each behaving approximately as a capacitor, are used along with the on-chip inductors between adjacent transistors to form a synthetic transmission line between the input and output ports of the SPST switch to achieve low loss across an extremely wide bandwidth when the switch is on. One series and two shunt nMOSFETs are used to provide compromise between IL and isolation. Biasing resistors (R_G's) are used in lieu of RF chokes to minimize the chip area. It is noted that, ideally, a synthetic transmission line consists of many identical LC sections. A practical synthetic transmission line, however, has a limited number of sections – for example, two in our

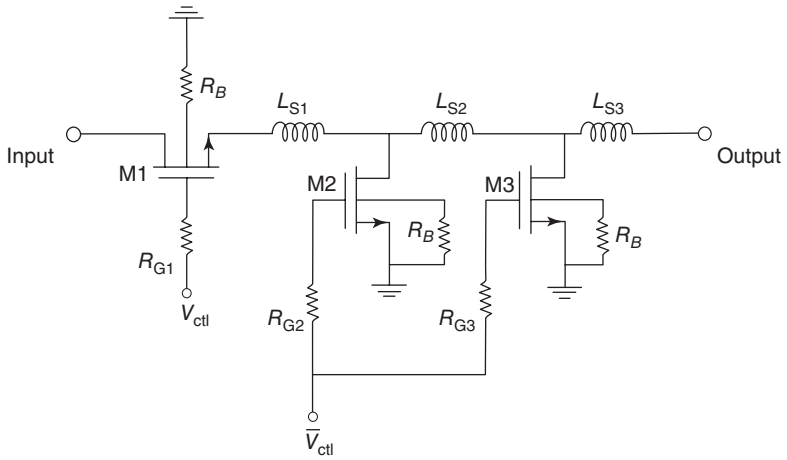


Figure 14.20. Schematic of the ultra-wideband SPST Switch. (After Rui, Jin, and Nguyen [1]. Reprinted with permission of IEEE.)

design – so the transmission line's operating frequency range, within which the IL is reasonably low and return loss is sufficiently large, is much lower than the cutoff frequency given in (14.61). Our analysis based on various transistors shows that the cutoff frequency should be approximately at least three times the highest operating frequency, so that a synthetic transmission line of limited number of sections would have low IL and high return loss.

To ensure wideband performance, the sizes of the series and shunt transistors need to be carefully determined. As mentioned before, the series transistor particularly plays an important role in the switch's IL, while the shunt transistors dominate the isolation. Shunt devices, however, also inadvertently aggravate the IL due to their parasitics. For the series-connected transistor, as the gate width is increased, the on-resistance reduces, resulting in low IL in the low-frequency region. The gate–source capacitance, however, increases, and hence increasing the IL at high frequencies. Figure 14.21 compares the IL of two 0.18-μm MOSFETs available in the employed CMOS process, each in series configuration, having two different gate widths (64- and 192-μm). As can be seen, the smaller-size transistor has higher loss at lower frequencies but its IL maintains relatively constant over a wide frequency range.

The main limitation of the CMOS transistors when used in switches is due to their parasitic capacitances, C_{gd} (gate–drain capacitance), C_{gs} (gate–source capacitance), C_{sb} (source–bulk capacitance), and C_{db} (drain–bulk capacitance). The effects of C_{sb} and C_{db} are more significant than those of C_{gs} and C_{gd} at millimeter-wave frequencies due to their larger values and the conductive Si substrate. RF signals can penetrate into the lossy Si substrate via C_{sb} and C_{db}, thus leading to increased IL. To reduce the effects of these capacitances, the floating-body technique with deep n-well transistors employed in [2] can be used. The bulk terminal of the deep n-well transistors, which are triple-well devices, is floated by a large resistor. With deep n-well, large resistors can be applied directly to the bulk of nMOS devices, making it floated at high frequencies without latch-up problems that would result in RFICs consisting of both nMOS and pMOS without deep n-well. Using deep n-well thus allows a switch to be fully integrated with other RFICs designed using both nMOS and pMOS transistors in a single chip. Floating the bulk forces the bulk resistance underneath the source and drain junctions open with respect to the ground, leading to a much smaller resistive loss in the conductive p-bulk than with the bulk grounded.

Figure 14.22 shows a simplified cross sectional view of a deep n-well nMOS transistor and its schematic and model. R_{sub} represents the losses of the conductive substrate and substrate-ground connection, which are normally small for good substrate and substrate-ground connection. Since the Si substrate is lossy, the effects of the substrate should be included in a transistor model and depend strongly on the layout of circuits. The deep n-well structure is used to isolate the bulk of the transistor from the p-substrate which is connected to the ground, thereby preventing the leakage (and hence loss) of RF signals from the transistor's conducting channel to the ground. However, the deep n-well layer also inadvertently creates a pair of parasitic junction diodes from the junctions between n-well/p-well and n-well/p-substrate, which are effectively equivalent to a parasitic capacitance C_{dnw} as shown in Figure 14.22. The value of this capacitance is approximately proportional to the size of the deep n-well layer needed to enclose the transistors. It is this capacitance that possibly

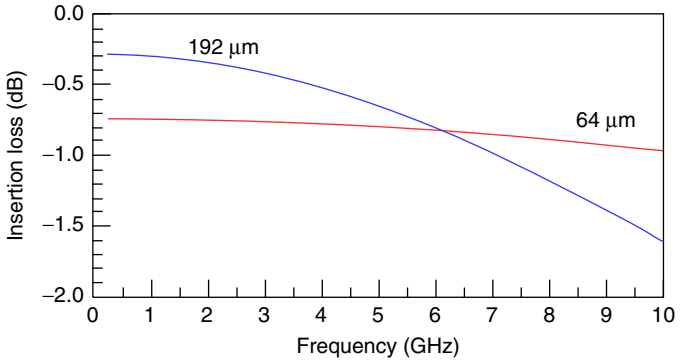


Figure 14.21. Insertion loss comparison of two 0.18-μm MOSFETs having 64- and 192-μm gate width. (After Rui, Jin, and Nguyen [1]. Reprinted with permission of IEEE.)

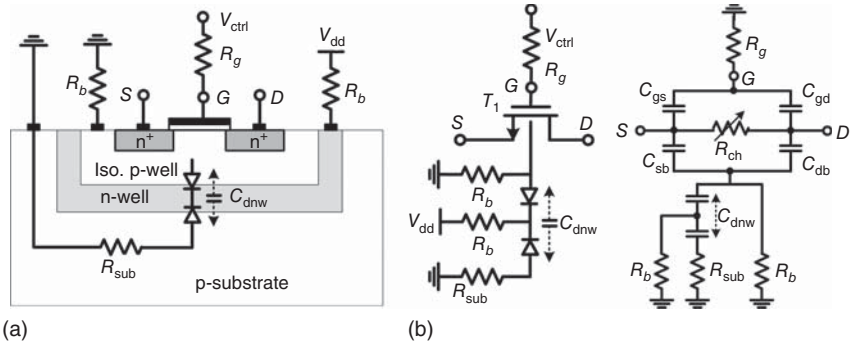


Figure 14.22. Simplified cross sectional view of a deep n-well transistor (a) and its schematic (b) and model (c).

limits the switch performance at high frequencies, especially in the millimeter-wave region. As the operating frequency increases, or the transistor size is larger, the impedance of this capacitance becomes smaller, hence degrading the inherent isolation (to the lossy Si substrate) provided by the deep n-well structure. As a result, for a given device technology, up to some certain frequency extent, the deep n-well layer becomes less effective in improving IL and isolation. To maximize the isolation to the substrate provided by the deep n-well layer (and hence minimizing the IL of RF signals), the isolated bulk of the transistor is biased at 0 V with a large resistor (e.g., 10 kΩ) to make the transistor body floating and the deep n-well is biased at V_{DD} (e.g., 1.8 V) through a large resistor to establish a reverse-bias for the two p–n deep n-well junctions to reduce the value of C_{dnw}.

Floating the bulk terminal of transistors in a switch also improves the power-handling ability. Referring back to the deep n-well transistor geometry as shown in Figure 14.22, it is apparent that the p–n junctions between the p-well (or p-bulk) and n+ regions (under the drain and source terminals) also form a pair of parasitic drain–bulk (or drain) and source–bulk (or source) diodes. The linearity of MOSFET switches is limited for large signal swings due to the conductivity modulation of the drain and source parasitic diodes caused by changes in the gate–source and drain–source voltages for large input RF signals. Another cause of the non-linearity is the parasitic source and drain junction diodes which can clip the RF signal at about 0.7 V above the biased DC voltage (i.e., $0.7 + V + V_{DD}$) or 0.7 V below ground.

14.5.1.1 On-State Operation. When the bias control voltages V_{ctl} and $\overline{V}_{ctl}$, as indicated in Figure 14.20, are set to V_{DD} and zero, respectively, the SPST switch works in its on-state, passing RF signals from the input to output ports. Figure 14.23 shows a simplified equivalent circuit of the on-state switch when both the gate and bulk of each transistor are floated by large resistors, making use of the approximate models for the on- and off-MOSFET based on the simplified small-signal equivalent-circuit models shown in Figure 9.27. $R_{on\text{-}M1}$ is the on-resistance R_{on} in Figure 9.27(a) for the transistor M1, and $C_{off\text{-}M2}$ and $C_{off\text{-}M3}$ are the off-capacitance C_{off} in Figure 9.27(b) for the transistors M2 and M3, respectively. The on-capacitance C_{on} in Figure 9.27(a) is neglected since its impedance is much larger than R_{on} under 20 GHz. Note that C_{off} of each transistor is an effective capacitor representing a lossy capacitor comprising the total capacitance C_g seen at the gate, including C_{gb}, C_{gd}, and C_{gs}, the off-state bulk capacitance $C_{b\text{-}off}$, consisting of C_{db} and C_{sb}, the drain–source capacitance C_{ds}, and the drain–source resistance R_{ds}.

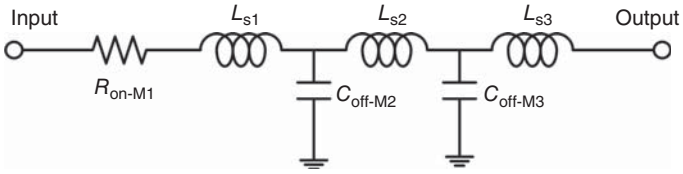


Figure 14.23. Equivalent network of the SPST switch under on-state.

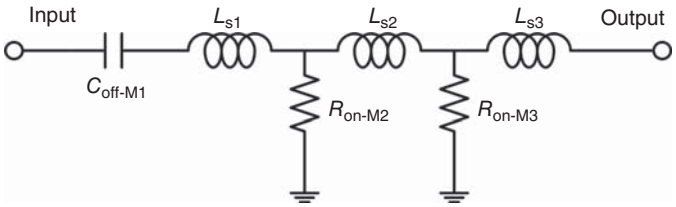


Figure 14.24. Equivalent network of the SPST switch working at off state.

To create a synthetic transmission line based on the model in Figure 14.19 we choose transistors M2 and M3 of the same size, which gives the same off-capacitance, and the three series inductors as $L_{S1} = L_{S3}$ and $L_{S2} = 2L_{S1}$. The on-state equivalent circuit of the SPST switch described in Figure 14.23 clearly shows a synthetic transmission line consisting of two basic (synthetic) segments in series with a resistor ($R_{on\text{-}M1}$). In the design, a proper size for M2 and M3 can be selected to give low IL. These devices possess a certain value for C_{off}. Assume typical 50-Ω characteristic impedance for the synthetic transmission line, we can then determine the corresponding values for the inductors according to the characteristic-impedance equation of (14.59). The inductors of these values along with the capacitors of C_{off} form a 50-Ω synthetic transmission line which works over an extremely wide bandwidth. We can now see that low IL over an ultra-wide bandwidth can be achieved for the designed SPST switch. It is noted that the bandwidth of this switch is limited by the cutoff frequency given approximately by (14.61). Moreover, according to this equation, the bandwidth of the switch is inversely proportional to the off-capacitance of M2 and M3, giving valuable information for the selection of transistor size for the switch design.

14.5.1.2 Off-State Operation. When the bias control voltages V_{ctl} and $\overline{V}_{ctl}$, as indicated in Figure 14.20, are set to zero and V_{dd}, respectively, the SPST switch works in off-state and rejects RF signals from the input to output ports. Figure 14.24 shows a simplified equivalent circuit for the switch under off-state using the approximate models for the on- and off-MOSFET. $R_{on\text{-}M2}$ and $R_{on\text{-}M2}$ are the on-resistances for the transistors M2 and M3, respectively, and $C_{off\text{-}M1}$ is the off-capacitance for the transistor M1. It is noted that the isolation contribution by the off-capacitance $C_{off\text{-}M1}$ is more at low frequencies than at high frequencies due to its impedance inversely proportional to the frequency. The on resistances $R_{on\text{-}M2}$ and $R_{on\text{-}M2}$ result in high isolation at high frequencies.

14.5.1.3 Switching Time. In typical transistor switches with gate biasing resistors, large resistors are normally used to make the gate open at RF so that the switch performance is not affected, thus making them suitable for slow switching only. The large gate biasing resistor leads to a large RC constant, which effectively slows down the control signal applied to the control terminal connecting to the gate via the resistor, which inadvertently reduces the switching speed. Figure 14.25 demonstrates the effect of the gate-biasing resistor, assuming an ideal step-signal is applied to the control terminal. Due to the gate resistance R_g and the total gate-to-ground capacitance C_p, which includes the parasitic and intrinsic capacitances seen at the gate, the rising edge of the resultant gate voltage V_g is slowed down. This is actually an RC charging effect and the falling edge is slowed down as well since the gate voltage is discharging with the same time constant. As can be seen in the figure, for a given size of MOSFET, a larger gate resistor gives slower rising edge, leading to slower switching speed. From Figure 14.25, it is apparent that in order to maintain the fidelity of the control narrow-pulse signals (in order of sub-nanoseconds), the gate resistance should be less than a few hundred ohms. Although this may aggravate the return loss and IL of the switch, through optimization of the switch's circuit and components, it can have reasonably good performance while maintaining the rising/falling edge of the control pulse signal, hence causing less effect to the switching speed.

For the MOSFETs employed in the ultra-wideband SPST switch, the gate-to-ground capacitance is estimated as 0.5 pF, whereas the calculated gate–source (C_{gs}) and gate–drain (C_{gd}) capacitances are each between 0.2 and 0.3 pF. Numerically, C_p is thus approximately equal to C_{gs} and C_{gd} in parallel, which is actually expected from the MOSFET's equivalent circuit. Using the estimated value of C_p and the input

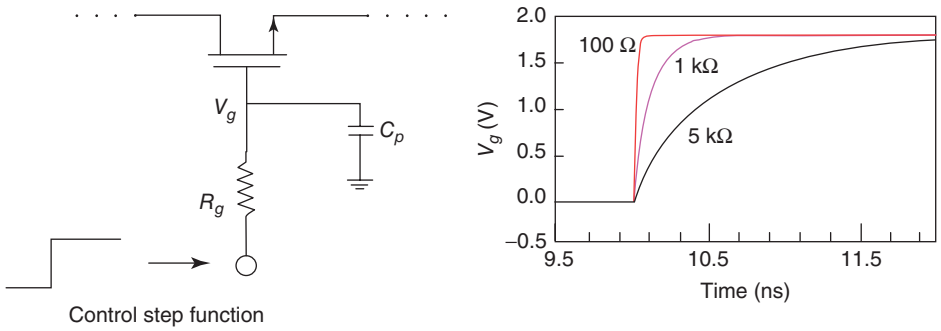


Figure 14.25. MOSFET with a step function applied to the gate and behavior of the gate voltage V_g for different gate bias resistances. (After Rui, Jin, and Nguyen [1]. Reprinted with permission of IEEE.)

TABLE 14.1. Summary of the Designed Ultra-Wideband SPST's Components

Circuit element	Element value
M1	512-μm gate width
M2, M3	368-μm gate width
L_{S1}	0.3 nH
L_{S2}	0.6 nH
L_{S3}	0.3 nH
R_{G1}	100 Ω
R_{G2}	100 Ω
R_{G3}	100 Ω
R_B	20 kΩ

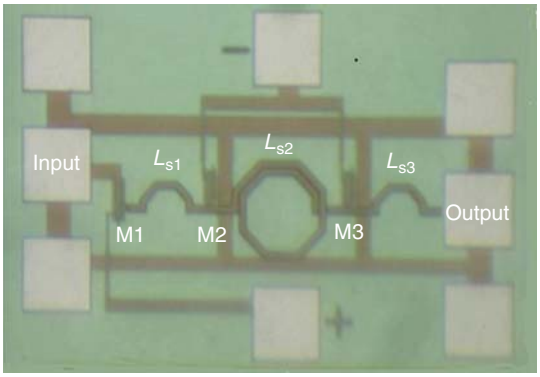


Figure 14.26. Microphotograph of the ultra-wideband SPST switch.

control signal, the optimal value for the gate resistor, that produces 300-ps switching time for the SPST switch, can thus be determined as 100 Ω.

The ultra-wideband SPST switch was designed and fabricated using the TSMC 0.18-μm CMOS process [3]. Table 14.1 summarizes the designed circuit elements.

Figure 14.26 shows a microphotograph of the SPST switch. The die area without pads is 0.08 mm^2. Transistors M1, M2, and M3 are implemented by multi-finger structures. On-chip inductors (L_{S1}, L_{S2}, and L_{S3}) are on the topmost thickest metallization, resulting in smaller series resistance and smaller parasitic capacitance. The inductors' physical parameters, including diameters and spiral widths, are optimized for high Q. To achieve a very low IL for the switch, the total resistance of the switch's on-path, consisting of both on-resistances of the MOSFETs and self-resistances of the on-chip inductors, needs to be designed to be as

small as possible. It is recalled that on-chip inductors in silicon-based RFICs contribute considerable IL and size because of their limited quality factor (Q) and relatively large size. The gate bias resistors and bulk bias resistors are all implemented with poly-silicon resistors.

14.5.1.4 Performance. The ultra-wideband SPST switch was measured on-wafer using a probe station and vector network analyzer. Short, open, load and thru calibration standards were designed and fabricated on-wafer using the same RF pads as those implemented by the switch. On-wafer calibration was performed to correct for all the effects including the RF-pad capacitances, contact resistances, probes and cables at both input and output ports. The control and bias voltages were applied through DC probes directly.

Figure 14.27 shows the measured results of the SPST switch under on- and off-state. As seen in Figure 14.27(a), the SPST switch under on-state exhibits measured IL less than 2 dB and return loss greater than 18 dB from DC to 11 GHz. A 0.6 dB variation of IL from DC to 11 GHz is achieved and the roll off of the IL beyond 11 GHz is caused by the RF leakage through the gate of the transistors, which is terminated with a resistor of only 100 Ω, and the substrate loss. The achieved high return loss at the input and output ports from DC to 15 GHz helps reduce the reflections at these ports, hence increasing the delivery of the incident power across this frequency range. The SPST switch under off state exhibits isolation greater than 50 dB from DC to 9 GHz and more than 40 dB from DC to 11 GHz, as shown in Figure 14.27(b). It is also shown in Figure 14.27(b) that both the measured input and output return losses are close to 0 dB, suggesting that most of the incident signal is reflected back by the off-switch.

The 1-dB input power compression, $P_{1\,\mathrm{dB}}$, that characterizes the switch's linearity, was also measured on-wafer using a spectrum analyzer. Figure 14.28 shows the measured and simulated $P_{1\,\mathrm{dB}}$ of 22 and 26.5 dBm

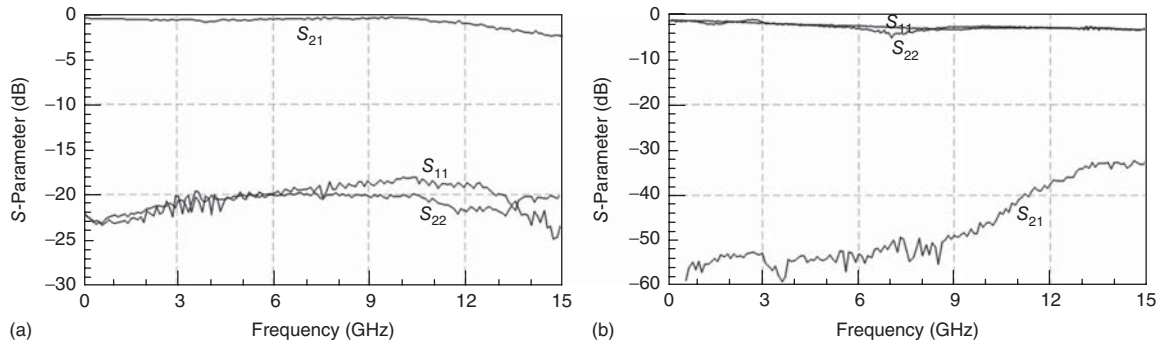


Figure 14.27. Measured S-parameters of the ultra-wideband SPST switch under on-state (a) and off-state. (b). S_{11} and S_{22} are the input and output return loss, respectively, and S_{21} is the insertion loss or isolation under on- and off-state, respectively. (After Rui, Jin, and Nguyen [1]. Reprinted with permission of IEEE.)

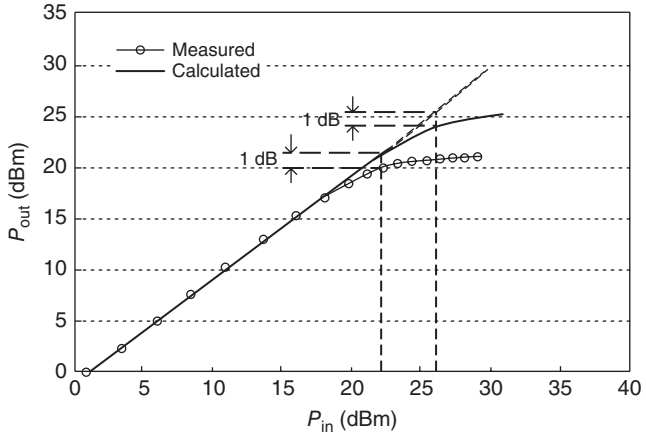


Figure 14.28. Measured and simulated $P_{1\,\mathrm{dB}}$ of the SPST switch under on-state at 5.8 GHz.

at 5.8 GHz, respectively. The discrepancy between the measured and calculated $P_{1\,\mathrm{dB}}$ is due to the inaccuracy of the large-signal model of the triple-well nMOS, especially when the bulk of the transistors is disconnected from the source, that is, large bulk resistors are used to float the bulk. Moreover, the measured 22-dBm $P_{1\,\mathrm{dB}}$ is without any bulk bias, which is 4 dB lower than the $P_{1\,\mathrm{dB}}$ of the T/R switch with negative bulk bias (as described in Section 14.5.2), confirming the fact that the major source of the non-linearity is the forward-biased drain diode of the shunt nMOS devices.

14.5.2 Ultra-Wideband T/R Switch

14.5.2.1 T/R Switch Topologies. T/R switches with two identical arms consisting of series and shunt transistors described in Section 14.2.3, such as that shown in Figure 14.29, are perhaps the most commonly used topology. In this topology, the path from the antenna (ANT) to receiver (RX) is on and from the transmitter (TX) to ANT is off when V_{C1} and V_{C2} are set to 0 and V_{DD}, respectively. The series (M1, M2) and shunt (M3 and M4) MOSFETs dominate the IL and isolation, respectively. As mentioned earlier for the device selection, low IL and high isolation can be achieved by using properly compromised large devices due to their small on-state resistances, which are scaled down approximately as L/W with L and W being the gate length and width, respectively, in advanced sub-micrometer CMOS processes. Large devices, however, have significant parasitic capacitances, causing considerable effects to circuit matching and eventually limiting the switch's bandwidth, especially in the high frequency regions. These parasitic capacitances are more pronounced in sub-micrometer CMOS processes. In practical switching circuits, the effects of parasitic capacitances are much more severe than the on-resistances in large sub-micrometer CMOS devices, particularly at high frequencies. Synthetic transmission lines utilize these parasitic capacitances as a constituent part of the transmission lines and hence can be used to alleviate the bandwidth limitation in CMOS T/R switches, as implemented for the CMOS ultra-wideband SPST switch discussed previously.

Figure 14.30 shows the schematic of the ultra-wideband CMOS T/R switch [4]. These topologies are identical except for the different bias schemes for the shunt transistors, bulk bias in Figure 14.30(a) and drain bias in Figure 14.30(b), used to enhance the switch's linearity and power handling to be discussed in the next section, and can be implemented using the same switch. Synthetic transmission lines, realized using two series (M1 and M2) and two shunt (M3 and M4) nMOS transistors, and three series on-chip spiral inductors (L_1, L_2, and L_3), are used to achieve a very wide bandwidth. The bias resistors (R_{G1}, R_{G2}, R_{G3}, R_{G4}, R_{D1}, and R_{D2}) have large resistances in order to isolate the DC bias voltages from the RF signals. Large bulk resistors (R_{B1}, R_{B2}, R_{B3}, and R_{B4}) are used to make the transistors floating at high frequencies to reduce the substrate loss and increase the switch's linearity and power handling capability. Large series and shunt nMOS transistors with gate widths in the order of several hundred micrometers are used to obtain small on-resistances and hence low IL and high isolation for the switch, besides enhancing the linearity and power handling capability. These devices, although having large sizes and correspondingly large parasitic capacitances, still result in extremely wide bandwidth due to the use of the synthetic transmission-line technique.

Figure 14.31(a) shows the small-signal equivalent circuit of the T/R switch assuming the ANT-RX and ANT-TX paths are on and off, respectively, where R_{on}'s and C_{off}'s are the on-resistances and off-capacitances seen in Figure 9.27. Figure 14.31(a) takes into account the fact that the on-capacitances C_{on}'s, as shown in

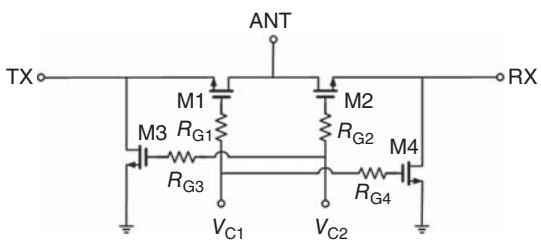


Figure 14.29. Typical series-shunt T/R switch topology. ANT, TX, and RX stand for the (input) antenna, (output) transmitter, and (output) receiver ports, respectively.

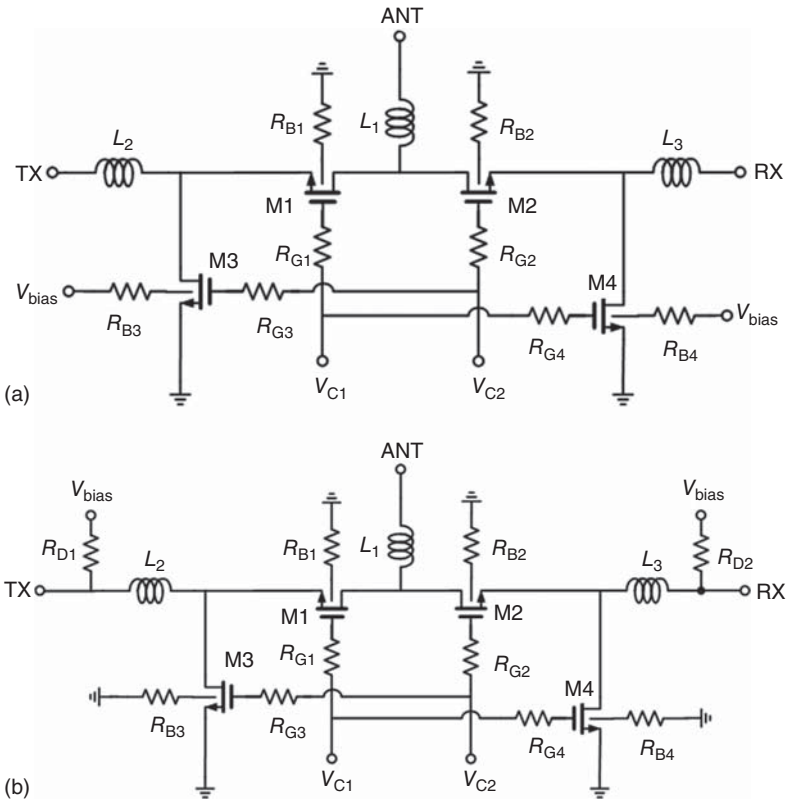


Figure 14.30. Schematics of the CMOS T/R switch implemented using synthetic transmission lines with negative bias to the bulk (a) and positive bias to the drain (b) of the shunt nMOS devices. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

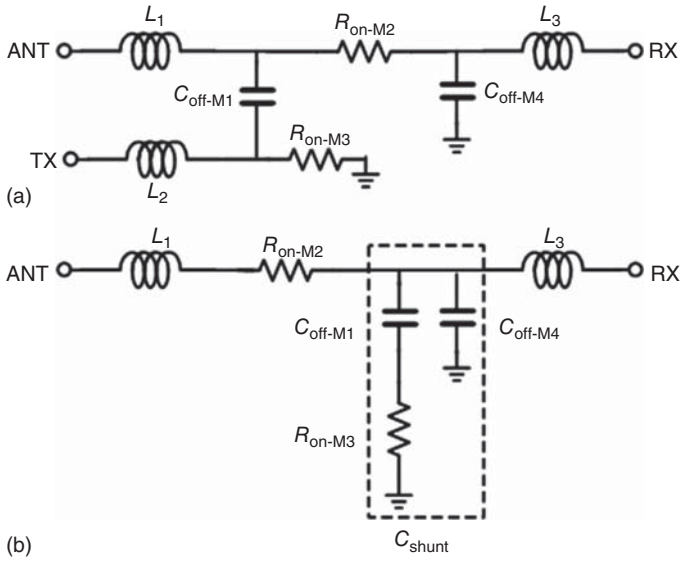


Figure 14.31. Small-signal equivalent-circuit models of the T/R switch with ANT-RX on and ANT-TX off (a) and the ANT-RX on-path (b). (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

Figure 9.27(a), have relatively large impedances as compared to R_{on}'s below 20 GHz. As can be seen, the isolation between the ANT and TX ports is mainly determined by the on-resistance $R_{on\text{-}M3}$ of the shunt MOSFET M3. Figure 14.31(b) shows the small-signal equivalent circuit of the on-path between the ANT and RX ports, where C_{shunt} represents the combined off-capacitances $C_{off\text{-}M1}$ of M1 and $C_{off\text{-}M4}$ of M4, and $R_{on\text{-}M3}$ of M3.

Considering a single synthetic transmission-line element as shown in Figure 14.19, we can see that the small-signal equivalent circuit of the ANT-RX on-path indeed represents a segment of a lossy synthetic transmission line, where the loss is contributed by $R_{on\text{-}M2}$ and $R_{on\text{-}M3}$ of the series transistor, M2, and shunt transistor, M3, respectively. This indicates that the IL of the switch can be minimized by increasing the sizes of the series and shunt MOSFETs, which effectively reduces the on-resistances. Meanwhile, the impacts of the large parasitic capacitances resulted from large-size MOSFETs can be alleviated by using well-designed on-chip inductors.

14.5.2.2 Linearity and Power Handling Enhancement. Typical CMOS switches have poor linearity and power handling, primarily due to the resultant forward-bias of the drain and source parasitic diodes of the shunt MOSFET device under operation, even with small transient voltage swings. In order to overcome the forward-bias problem, and hence increasing the switch's linearity and power handling capability, the bulk can be floated and negatively biased simultaneously as shown in Figure 14.30(a). Due to the small on-resistance of the series MOSFETs, the source and drain are kept approximately at the same potential under the on-state. The parasitic drain and source diodes are thus always kept reverse-biased even when there are strong voltage swings at the drain and source, respectively. However, because the sources of the shunt MOSFETs are grounded, a negative voltage swing on the drain can push the two back-to-back parasitic diodes into a forward-bias region. Consequently, the voltage on the drain is clamped to a certain value in the negative region by these forward-biased diodes, leading to distortion in the output signal and, consequently, degrading the IL, linearity, and power handling.

Figure 14.32 illustrates three conditions for a shunt MOSFET, which result in different linearity and power handling capabilities. In Figure 14.32(a), the bulk of the MOSFET is grounded directly. When the RF voltage swing at the drain is lower than $-2V_B$, where V_B is the forward pinch-on voltage of the parasitic diodes, the back-to-back parasitic diodes are all forward-biased and, consequently, the MOSFET functions as a small forward-biased resistor in parallel with the capacitor C_{off}. The output voltage of the switch is then approximately clamped to $-2V_B$. Figure 14.32(b) shows the MOSFET with the bulk floated through a grounded resistor. A voltage swing reaching $-V_B$ at the drain would push the drain parasitic diode forward biased, but the large resistor at the bulk keeps a high impedance between the drain and ground and the source parasitic diode reverse-biased, thus improving the power handling capability. Figure 14.32(c) demonstrates another technique to further improve the linearity and power handling. It shows the floating bulk of the shunt MOSFET is biased using a negative DC voltage via the bulk bias resistor R_B. This technique is implemented in the topology shown Figure 14.30(a). Since there is no current flow through R_B, the DC potential of the bulk node is kept the same as the negative bias. Therefore, the source parasitic diode is always in the reverse bias

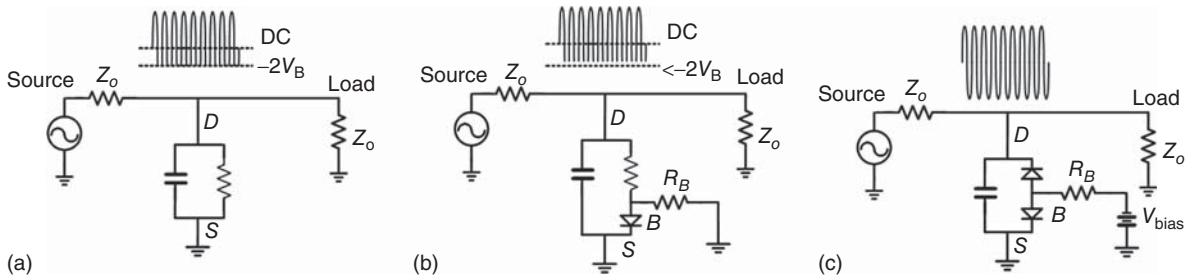


Figure 14.32. Simplified large signal models for a shunt MOSFET with bulk grounded (a), floated (b), and negatively biased (c) for different linearity and power handling conditions. The sinusoidal signal is at the drain. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

and the drain parasitic diode can withstand a strong negative voltage swing to $-(V_B + V_{Bias})$. Using the negative bulk-bias technique can therefore lead to larger power handling for CMOS switches than the other two methods shown in Figure 14.32(a) and (b). It is noted that, due to no output current required for the negative voltage source, a negative voltage reference can be implemented in fully integrated systems without any noise and stability issues.

The linearity and power handling capability can also be improved by generating a positive DC potential between the drain and the bulk of the shunt MOSFETs. This is achieved by applying a positive bias to the drain of the shunt MOSFETs and grounding the bulk resistors, as seen in the topology shown in Figure 14.30(b). This positive-drain bias technique is especially attractive when the RF signal entering the receiver or leaving the transmitter port has a positive DC offset, because no DC blocks would be needed, thus making it very useful for integration with other RF ICs in a single chip. Using the same CMOS T/R switch, both the negative-bulk and positive-drain bias techniques give the same large signal equivalent circuits and thus have the same linearity and power handling.

14.5.2.3 T/R Switch Design. The CMOS T/R switch was designed and fabricated using the TSMC 0.18-μm CMOS triple-well process [3] with nMOS transistors and on-chip octagonal spiral inductors on the top most metallization layer, which is the thickest among the metallization layers. All the bias resistors are realized on the poly-silicon layer to achieve small layouts. The on-resistance R_{on} is about 4 and 11 Ω for the employed series and shunt nMOSFET, respectively. The combined off-capacitance of the series and shunt transistors is about 280 fF.

Several remarks need to be made at this point concerning the CMOS T/R switch design. As implied in Eq. (14.61), small transistors, which correspond to small off-state capacitances, should be employed to achieve a wide bandwidth for the switch. On the other hand, large series and shunt devices are needed to produce low IL and high isolation, respectively. A trade-off is thus needed not only in the switch topology but also in the device size, in order to achieve simultaneously an ultra-wide bandwidth along with low IL and high isolation. Particularly for the selected T/R switch topology, the inductance and Q of the on-chip spirals versus frequency should be optimized together with C_{shunt}, the combined off-capacitances of the series and shunt devices. The spiral inductance and C_{shunt} dictate the bandwidth while the Q affects the IL.

Table 14.2 lists the parameters of the designed CMOS T/R switch. Figure 14.33 shows a micrograph of the CMOS T/R switch, including on-wafer RF and DC probe pads, with the TX port terminated by an on-chip 50-Ω resistor. The actual area of the switch is measured 230 μm × 250 μm, with the inductors occupying approximately 60% of the chip area.

TABLE 14.2. Summary of the CMOS T/R Switch's Parameters

Circuit element		Element value
M1, M2		304-μm gate width
M3, M4		184-μm gate width
L_1, L_2, L_3		0.35 nH, 1.5 turns
$R_{G1}, R_{G2}, R_{G3}, R_{G4}$		10 kΩ
$R_{B1}, R_{B2}, R_{B3}, R_{B4}$		10 kΩ
With bulk bias	V_{bias}	−1.8 V
	V_{C1}	1.8 V (on), −1.8 V (off)
	V_{C2}	−1.8 V (on), 1.8 V (off)
With drain bias	V_{bias}	1.8 V
	V_{C1}	3.6 V (on), 0 V (off)
	V_{C2}	0 V (on), 3.6 V (off)
Without bulk/drain biases	V_{bias}	0 V
	V_{C1}	1.8 V (on), 0 V (off)
	V_{C2}	0 V (on), 1.8 V (off)

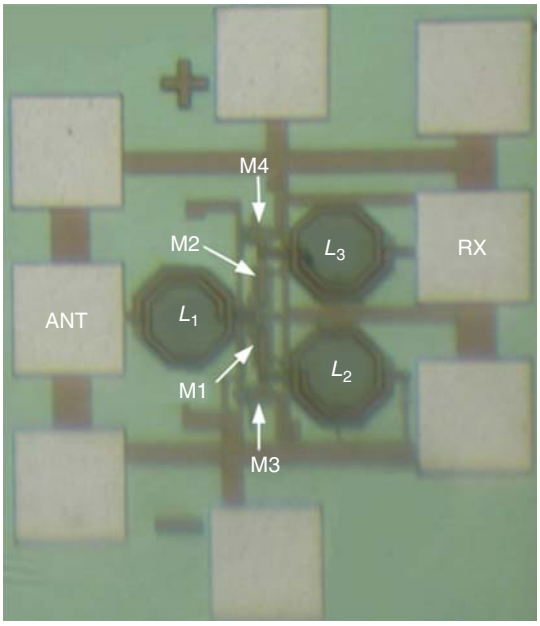


Figure 14.33. Microphotograph of the ultra-wideband CMOS T/R switch. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

14.5.2.4 T/R Switch Performance. The measurements of the CMOS T/R switch were conducted on-wafer using a probe station, vector network analyzer, frequency synthesizers, and on-wafer calibration standards (short, open, load, and thru) built on the same chip.

Small-Signal *S*-Parameter Measurement

The measurement was carried out between the ANT port (port 1) and RX port (port 2). The TX port is terminated with a 50-Ω on-chip resistor. Due to the symmetric structure of the switch, it is expected that the path from the ANT port to TX port exhibits the same characteristics as those between ANT and RX ports. Figure 14.34 shows the measured and calculated IL, isolation, and return loss of the ultra-wideband T/R switch. It exhibits ILes of less than 0.7, 1, and 2.5 dB from DC to 10, 10 to 18, and 18 to 20 GHz, respectively. The measured isolation varies from 32 to 60, 25 to 32, and 25 to 27 dB between DC to 10, 10 – 18, and 18 – 20 GHz, respectively. The measured return losses at the antenna and receiver/transmitter ports are better than 15 and 10 dB from DC to 10 GHz and 10 to 18 GHz, respectively. These results are the same for the three different bias conditions listed in Table 14.2.

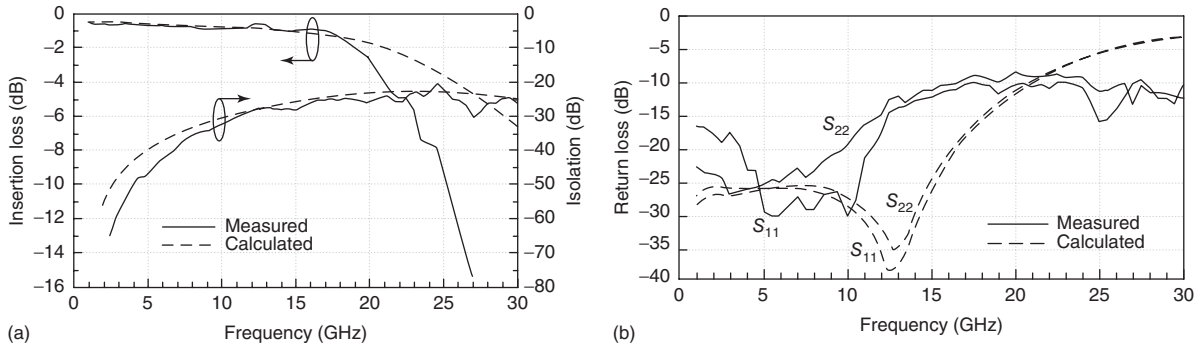


Figure 14.34. Measured and calculated insertion loss and isolation (a) and return loss at the antenna port (S_{11}), and RX or TX port (S_{22}) (b). (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

As can be seen in Figure 14.34(a), the measured and calculated ILes agree very well to approximately 18 GHz, beyond which the measured IL significantly drops below the simulated result. This discrepancy above 18 GHz has been observed in several chips and is mainly caused by the inaccuracy of the transistors' available model, specifically, the parameters of the model which characterize the silicon substrate loss. The employed transistor model also leads to the discrepancy between the measured and simulated isolation and return loss seen in Figure 14.34.

Linearity and Power Handling Measurement

Linearity of the switch is characterized by either the third-order intercept point (IP3) or the 1-dB compression point ($P_{1\,\mathrm{dB}}$). Figure 14.35 shows the measured IP3 at 5.8 GHz with −1.8 V bulk bias voltage. Figure 14.36 shows the measured input IP3 (IIP3) variation versus bulk bias voltage at 5.8 GHz. As can be seen, the IIP3 can reach as much as 41 dBm by increasing the negative bias voltage. Figure 14.37 compares the simulated and measured output power versus input power at 5.8 GHz with −1.8 V bulk bias, from which $P_{1\,\mathrm{dB}}$ can be determined. There is a 3-dB discrepancy between the measured and calculated output powers when the input power is around 1 W due to the inaccuracy of the transistor model. Figure 14.38 compares the $P_{1\,\mathrm{dB}}$ results without bulk bias (bulk bias resistor is grounded) and with bulk bias at −1.8 V. The $P_{1\,\mathrm{dB}}$ with the bulk

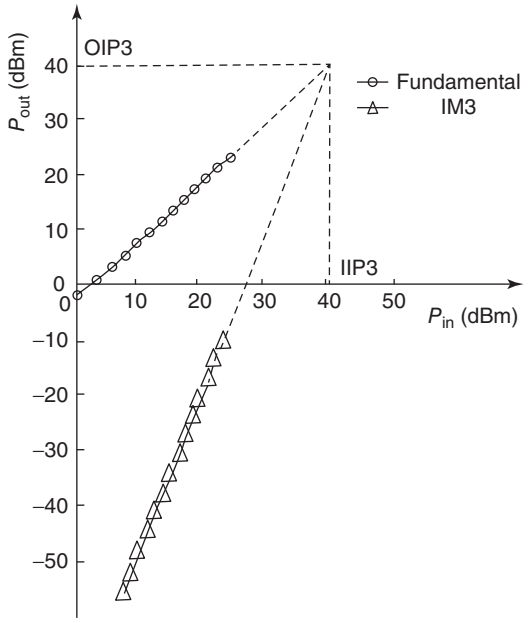


Figure 14.35. Measured IP3 at 5.8 GHz by a two-tone test. Two tones have frequencies of 5.8 and 5.85 GHz with same input power. Bulk bias voltage is −1.8 V. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

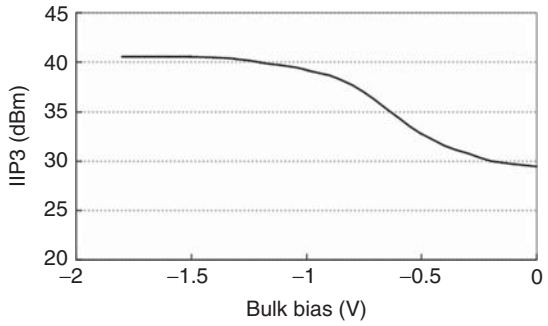


Figure 14.36. Measured IIP3 at 5.8 GHz. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

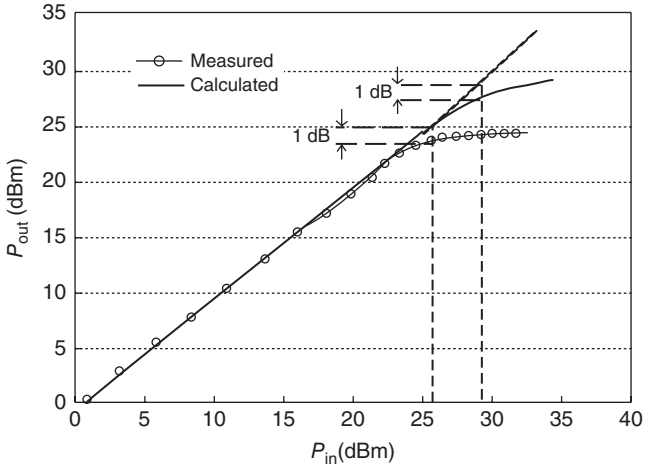


Figure 14.37. Calculated and measured output powers at 5.8 GHz. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

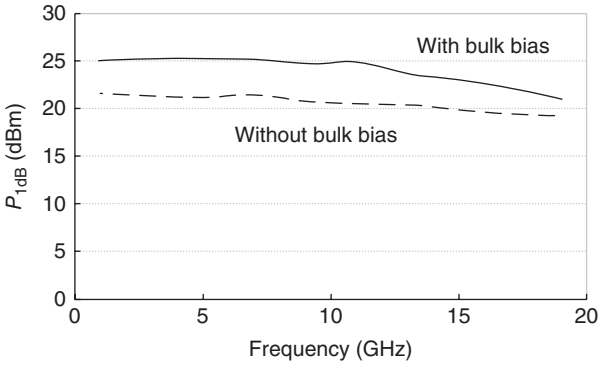


Figure 14.38. Measured $P_{1\,\mathrm{dB}}$ with 0- and − 1.8-V bulk bias voltage. (After Jin and Nguyen [4]. Reprinted with permission of IEEE.)

biased is more than 25 and 20 dBm from 1 to 11.5 and 11.5 to 19 GHz, respectively, reaching 26.2 dBm at around 4 GHz. It is seen that by applying a negative bias to the bulk, the power handling capability of the switch is increased by approximately 4 dBm. Similar performance for the IIP3 and $P_{1\,\mathrm{dB}}$ were also obtained when the drain was biased with a positive voltage as discussed in Section 14.5.2.2.

14.6 ULTRA-HIGH-ISOLATION SWITCHES

Isolation is an important parameter in switches and RF systems, and high isolation is especially crucial in order to reduce or prevent RF leakage. Undesired RF leakage not only causes external effects, such as harm to other coexisting systems but also causes internal effects such as reduced dynamic range. This section addresses an SPST switch architecture that not only can increase the isolation significantly but also can lower the IL or potentially produce gain. Particularly, the design of an ultra-high-isolation SPST switch based on this architecture that operates from 10 to 38 GHz is described.

14.6.1 Ultra-High-Isolation Switch Architecture and Analysis

14.6.1.1 Switch Architecture and Operation. Figure 14.39 shows the ultra-high-isolation SPST switch architecture [5]. It consists of two identical SPST switches and an active balun. The active balun can also be replaced with a passive balun. The two identical switches are named "Off-SPST" switch (always operating in the off-mode) and "Core-SPST" switch operating as a normal switch (on/off). The Off-SPST and

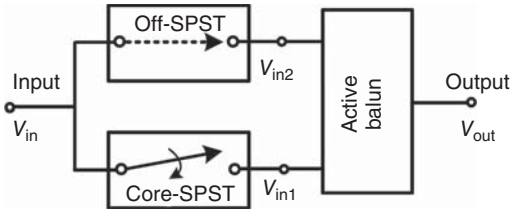


Figure 14.39. Ultra-high-isolation SPST RF switch architecture. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

Core-SPST switches can be implemented using any topology such as a simple conventional switch employing series and/or shunt transistors such as the series-shunt MOSFET switch used in one path of the T/R switch shown in Figure 14.29.

The ultra-high-isolation SPST switch is off or on according to the off-state or on-state of the Core-SPST switch, respectively. When the Core-SPST switch is off, an RF signal (V_{in1}) leaking through this switch, due to its limited isolation, appears at the input port 1 of the balun. At the same time, another RF leakage signal (V_{in2}) from the Off-SPST switch also appears at the balun's input port 2. These signals are theoretically equal since both switches are identical, and are cancelled in the balun, resulting in no RF signal at the output of the ultra-high-isolation SPST switch, and hence (theoretically) infinite isolation at the design frequency of the balun.

It is noted that when the Core-SPST switch, and hence the ultra-high-isolation SPST switch, is on, there are still two RF signals passing through the balun, the main V_{in1} signal from the "on" Core-SPST switch and the leaking V_{in2} signal from the Off-SPST switch. However, as the balun's output signal (V_{out}) is a function of the difference between V_{in1} and V_{in2}, and V_{in1} is much larger than V_{in2}, V_{out} is not altered significantly. In fact, the active balun can also be designed to function as an amplifier to provide amplification for V_{in1} to compensate for the loss it suffers as going through the Core-SPST switch. This, in turn, results in an overall gain for the ultra-high-isolation SPST switch.

14.6.1.2 Active Balun Analysis. The performance of the ultra-high-isolation SPST switch, particularly its isolation, depends significantly on the balun and, hence, an analysis for the balun is important for the performance assessment and design of the switch. Figure 14.40(a) shows the schematic of the active balun

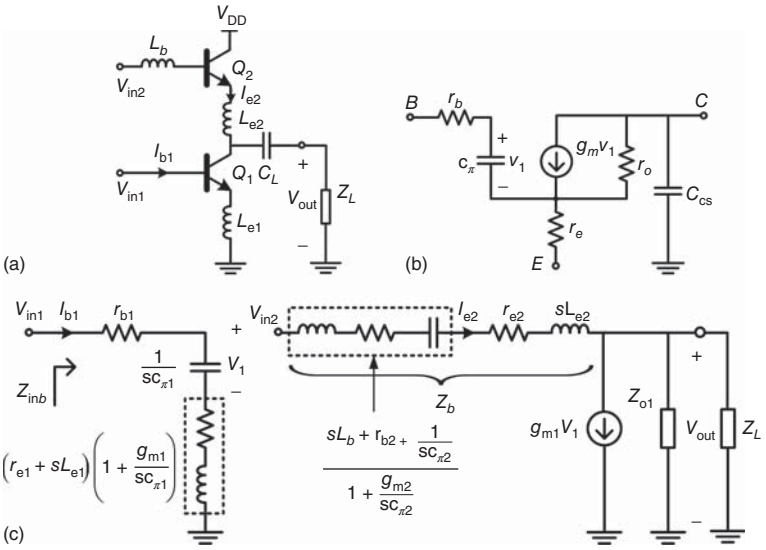


Figure 14.40. (a) Schematic of the active balun. (b) Small-signal transistor model with base resistance r_b, emitter resistance r_e, output resistance r_o, base–collector capacitance c_π, transconductance g_m, and collector–substrate capacitance c_{cs}. (c) Equivalent circuit of the active balun. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

used in the ultra-high-isolation SPST switch. It consists of two identical heterojunction bipolar transistor (HBT) transistors and three inductors L_{e1}, L_{e2}, and L_b which control the balance of the balun. These three inductors enable the balance of the balun to work up to a very high frequency, and hence across an extremely wide frequency range. In addition, L_{e1}, as a degenerative inductor, helps improve the linearity of the balun. The use of L_{e1} and L_{e2} also helps reduce the DC power consumption.

Figure 14.40(b) shows a simplified small-signal model of the HBT and Figure 14.40(c) shows the corresponding small-signal equivalent circuit of the active balun after several manipulations. The transistor's output resistance r_o is normally large, around several kilo-ohms, and is omitted for transistor Q_2. Z_{o1} is the output impedance looking into transistor Q_1 from its collector and can be derived as

$$Z_{o1} = \frac{1}{sc_{cs}} \Bigg\| \left\{ r_{o1} + \left[\left(r_{b1} + \frac{1}{sc_{\pi 1}} \right) \Bigg\| (r_{e1} + sL_{e1}) \right] \left(1 + \frac{r_{o1} g_{m1}}{1 + sc_{\pi 1} r_{b1}} \right) \right\} = R_{o1} \| jX_{o1} \tag{14.62}$$

where $\|$ denotes parallel operation and $s = j\omega$. This impedance, as indicated in (14.62), is represented by two equivalent elements, R_{o1} and X_{o1}, connected in parallel at the output of Q_1.

The balance of the balun is characterized by V_{out} as a function of V_{in1} and V_{in2} and, from Figure 14.40(c), we can derive

$$V_{out} = Z_L \| Z_{o1} (i_{e2} - g_{m1} v_1) \tag{14.63}$$

where

$$i_{e2} = \frac{V_{in2} - V_{out}}{Z_b} \tag{14.64}$$

$$V_1 = \frac{1}{sc_{\pi 1}} \frac{v_{in1}}{Z_{in\,b}} \tag{14.65}$$

with

$$Z_b = \frac{sL_b + r_{b2} + r_{e2} + (L_{e2} g_{m2}/c_{\pi 2}) + sL_{e2} + (1 + r_{e2} g_{m2}/sc_{\pi 2})}{1 + (g_{m2}/sc_{\pi 2})} \tag{14.66}$$

$$Z_{in\ \ b} = r_{b1} + r_{e1} + \frac{L_{e1} g_{m1}}{c_{\pi 1}} + sL_{e1} + \frac{1 + r_{e1} g_{m1}}{sc_{\pi 1}} \tag{14.67}$$

$Z_{in\,b}$ is the input impedance of the balun at the input port corresponding to V_{in1}.

Substituting (14.64) and (14.65) into (14.63) and making use of (14.66) and (14.67), we obtain

$$V_{out} = \frac{Z_L \| Z_{o1}}{Z_L \| Z_{o1} + Z_b} (V_{in2} - K V_{in1}) \tag{14.68}$$

where

$$K = \frac{Z_b}{Z_{in\,b}} \frac{g_{m1}}{sc_{\pi 1}} = \frac{(g_{m1}/c_{\pi 1})}{s + (g_{m2}/c_{\pi 2})} \frac{s + ((r_{b2} + r_{e2} + (L_{e2} g_{m2}/c_{\pi 2}))/L_b) + ((sL_{e2} + (1 + r_{e2} g_{m2}/sc_{\pi 2}))/L_b)}{((r_{b1} + r_{e1} + (L_{e1} g_{m1}/c_{\pi 1}))/L_b) + ((sL_{e1} + (1 + r_{e1} g_{m1}/sc_{\pi 1}))/L_b)} \tag{14.69}$$

Now we consider two cases corresponding to the on and off states of the ultra-high-isolation SPST switch which are obtained when the Core-SPST switch is turned on and off, respectively.

Case 1 (Ultra-High-Isolation SPST Switch Off)

Under this condition, V_{in1} and V_{in2} are equal and represent two identical leaking RF signals feeding the balun. Assuming the balun is perfectly balanced, its output signal V_{out} must be zero which, from (14.68), leads to

$$K = 1 \tag{14.70}$$

Equation (14.70) shows that the condition required to make the balun balanced, hence suppressing the leaking RF signal, is independent of the load Z_L of the balun, and hence the load of the ultra-high-isolation SPST switch, and the transistor's output elements r_o and c_{cs}. Using the balance condition in (14.70) and making use of (14.69), we get

$$sL_{\text{e1}} + \frac{1 + r_{\text{e1}}g_{\text{m1}}}{sc_{\pi 1}} = sL_{\text{e2}} + \frac{1 + r_{\text{e2}}g_{\text{m2}}}{sc_{\pi 2}} = 0 \tag{14.71}$$

and

$$\frac{g_{\text{m1}}}{c_{\pi 1}} = \frac{r_{\text{b1}} + r_{\text{e1}} + (L_{\text{e1}}g_{\text{m1}}/c_{\pi 1})}{L_b} = \frac{g_{\text{m2}}}{c_{\pi 2}} = \frac{r_{\text{b2}} + r_{\text{e2}} + (L_{\text{e2}}g_{\text{m2}}/c_{\pi 2})}{L_b} \tag{14.72}$$

from which, L_{e1}, L_{e2}, and L_b are obtained at the design frequency, ω_o, as

$$L_{\text{e1}} = L_{\text{e2}} = \frac{1 + r_{\text{e1}}g_{\text{m1}}}{\omega_0^2 c_{\pi 1}} = \frac{1 + r_{\text{e2}}g_{\text{m2}}}{\omega_0^2 c_{\pi 2}} \tag{14.73}$$

and

$$L_b = L_{\text{e1}} + \frac{c_{\pi 1}(r_{\text{b1}} + r_{\text{e1}})}{g_{\text{m1}}} = L_{\text{e2}} + \frac{c_{\pi 2}(r_{\text{b2}} + r_{\text{e2}})}{g_{\text{m2}}} \tag{14.74}$$

Equations (14.73) and (14.74) represent the required inductances for the balun in terms of the transistor's parameters that make it perfectly balanced (equal amplitude and 180° phase difference at outputs) at the design frequency.

The gain of the active balun when the ultra-high-isolation switch is off can be derived from (14.68) as

$$A_{\text{voff}} = \frac{V_{\text{out}}}{V_{\text{in2}}} = \frac{V_{\text{out}}}{V_{\text{in1}}} = \frac{Z_L \| Z_{\text{o1}}}{Z_L \| Z_{\text{o1}} + Z_b}(1 - K) \tag{14.75}$$

A_{voff} measures the suppression of the leaking RF signal V_{in1} as it traverses the active balun to the ultra-high-isolation SPST switch's output. It represents the additional isolation provided by the active balun to the ultra-high-isolation SPST switch besides the isolation obtained from the Core-SPST switch. This extra isolation is contributed by the factors $(1 - K)$ and $(Z_L \| Z_{\text{o1}})/(Z_L \| Z_{\text{o1}} + Z_b)$. The first represents the main isolation provided by the balun. The latter, on the other hand, represents the gain of the active balun when the ultra-high-isolation SPST switch is on, as will be seen later in Eq. (14.76); this gain degrades slightly the overall isolation of the ultra-high-isolation SPST switch.

Figure 14.41 shows the magnitude and phase of K versus frequency for the active balun employing identical transistors Q_1 and Q_2 under the same bias conditions, whose parameters are $r_{\text{b1}} = r_{\text{b2}} = 20\ \Omega, r_{\text{e1}} = r_{\text{e2}} = 1.7\ \Omega$, $g_{\text{m1}} = g_{\text{m2}} = 254$ mA/V, $c_{\pi 1} = c_{\pi 2} = 232$ fF, and $c_{\text{cs1}} = c_{\text{cs2}} = 50$ fF. The design frequency at which $K = 1$ is 35 GHz. However, as can be seen, the magnitude and phase of K vary only slightly over wide frequency ranges. For instance, across DC to 80 GHz, the magnitude of K changes from 0.975 to 1.09, while its phase varies from 0.06° to −4°. This shows that the active balun not only is perfectly balanced at the design frequency but also maintains good balance from DC to far beyond the design frequency. This unique characteristic of the active balun makes possible the design of ultra-high-isolation switches over extremely wide bandwidth.

The magnitudes of $(1 - K)$, representing the active balun's primary isolation contribution, and A_{voff}, characterizing the overall isolation contribution of the active balun to the ultra-high-isolation SPST switch, with

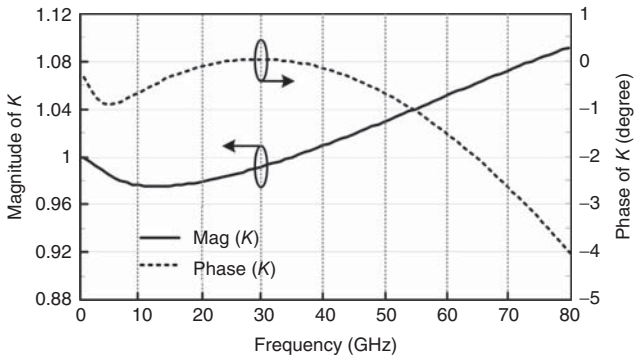


Figure 14.41. Magnitude and phase of K. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

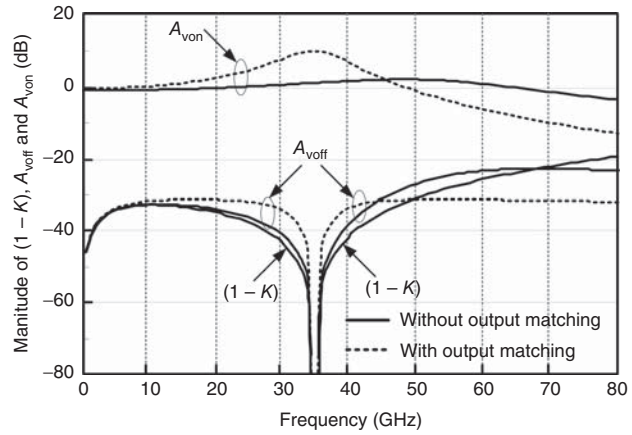


Figure 14.42. Magnitudes of $(1-K)$, A_{voff} and A_{von}. $(1-K)$ is not affected by the output matching. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

and without output matching are shown in Figure 14.42. As can be seen, the extra isolation provided by the active balun to the ultra-high-isolation SPST switch is significant, particularly around the design frequency, as compared to the conventional implementation of a single switch (i.e., Core-SPST switch alone). This effective isolation-enhancement technique is very useful for extending the isolation bandwidth of RF switches into the millimeter-wave range. Without the output matching, the active balun provides more than 32-dB isolation up to 40 GHz and (theoretically) infinite isolation at the design frequency of 35 GHz. When the output is matched, the active balun provides an isolation higher than 31 dB up to 40 GHz and also theoretically infinite isolation at 35 GHz. Good isolation improvement up to 80 GHz can also be observed from the plots.

Case 2 (Ultra-High-Isolation SPST Switch On)

When the Core-SPST switch is on, the main RF signal V_{in1} from the Core-SPST switch is much larger than the leaking RF signal V_{in2} from the Off-SPST switch. This, together with the fact that K is almost 1 over a wide frequency range as seen in Figure 14.41, enables V_{in2} to be ignored in Eq. (14.68). The active balun now works as an amplifier with its gain given from (14.68) as

$$A_{\text{von}} = \frac{V_{\text{out}}}{V_{\text{in1}}} = -\frac{Z_L \| Z_{\text{o1}}}{Z_L \| Z_{\text{o1}} + Z_b} K \approx -\frac{Z_L \| Z_{\text{o1}}}{Z_L \| Z_{\text{o1}} + Z_b} \tag{14.76}$$

At the design frequency $f = f_0$, $K = 1$, and we can obtain from (14.66) and (14.67):

$$Z_b = j\omega_0 L_b \tag{14.77}$$

$$Z_{\text{in}\,b} = R_{\text{in}} = r_{\text{b1}} + r_{\text{e1}} + \frac{L_{\text{e1}} g_{\text{m1}}}{c_{\pi 1}} = \frac{L_b g_{\text{m1}}}{c_{\pi 1}} \tag{14.78}$$

The matching at the input can be obtained by setting R_{in} in (14.78) equal to 50 Ω, while the output is matched by letting

$$Z_L = (Z_b \| Z_{\text{o1}})^* \tag{14.79}$$

where * indicates conjugate match. The gain of the balun with output matching at ω_o can be obtained from (14.62), (14.76), and (14.77) as

$$A_{\text{von}} = -\frac{1}{2}\frac{R_{\text{o1}}}{\omega_0 L_b} \tag{14.80}$$

or

$$\begin{aligned} |A_{\text{von}}| &= \frac{1}{2}\frac{R_{\text{o1}}}{\omega_0 (L_{\text{e2}} + ((c_{\pi 2}(r_{\text{b2}} + r_{\text{e2}}))/g_{\text{m2}}))} \\ &= \frac{1}{2}\frac{R_{o1}}{((1 + r_{\text{e2}} g_{\text{m2}})/\omega_0 c_{\pi 2}) + ((\omega_0 c_{\pi 2}(r_{\text{b2}} + r_{\text{e2}}))/g_{\text{m2}})} \end{aligned} \tag{14.81}$$

upon using (14.73) and (14.74).

The magnitude of A_{von} is also shown in Figure 14.42, demonstrating possible gain contribution to the ultra-high-isolation SPST switch. As can be seen, the active balun provides gain for the RF switch up to 50 GHz for both unmatched and matched output. At the design frequency of 35 GHz, it provides 1.3- and 10-dB gain with and without output matching, respectively. Of particular note is that, over a narrow frequency range, the ultra-high-isolation SPST switch can achieve both high gain and extremely high isolation at microwave and millimeter-wave frequencies with proper device technologies.

Letting the derivative of $|A_{\text{von}}|$ in (14.81) with respect to f_o equal to zero, we can find the maximum gain

$$|A_{\text{von}}|_{\text{max}} = \frac{R_{\text{o1}}}{4\sqrt{(r_{\text{b2}} + r_{\text{e2}})((1 + r_{\text{e2}} g_{\text{m2}})/g_{\text{m2}})}} \tag{14.82}$$

occurring at

$$f_{0\text{max}} = \frac{1}{2\pi c_{\pi 2}} \sqrt{\frac{g_{\text{m2}}(1 + r_{\text{e2}} g_{\text{m2}})}{(r_{\text{b2}} + r_{\text{e2}})}} \tag{14.83}$$

In this analysis, it is assumed that the outputs of the Core-SPST and Off-SPST switches do not affect the active balun operation. In reality, as can be inferred from the SPST switch architecture shown in Figure 14.39, these outputs can affect the balun's operation. In fact, the output impedances of these constituent switches would modify the derived equations (14.73) and (14.74) for the active balun's balanced conditions. This, however, will not cause loss of our analysis' generality. The foregoing formulation process still remains the same, except that the real and imaginary parts of these impedances, appearing at the input ports of the balun's equivalent circuit in Figure 14.40(c), will be absorbed into the impedances described in (14.66) and (14.67), and K may not be exactly equal to 1, hence slightly reducing the gain A_{von} when the RF switch is on.

14.6.1.3 Core- and Off-SPST Switch. The ultra-high-isolation SPST switch is made up of the core-SPST and off-SPST switches as shown in Figure 14.39. These constituent switches can be realized using any topology. For wide-band operation, we use a series-shunt SPST switch topology implemented using the synthetic transmission-line technique. Figure 14.43 shows the schematic of these switches using the schematic for deep n-well nMOS transistors as shown in Figure 14.22(b). As mentioned in the design of an ultra-wideband SPST switch in Section 14.5.1, the cutoff frequency of a synthetic transmission line should be at least three times the highest operating frequency, so that a practical synthetic transmission line with limited number of sections

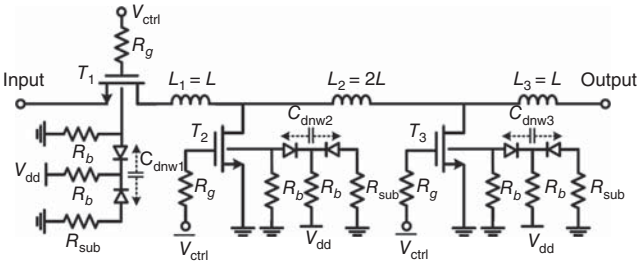


Figure 14.43. Schematic of the core- and off-SPST switches. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

would have low IL and high return loss. Therefore, in this design, we assume a cutoff frequency equal to at least three times of the highest operating frequency. The value of L and the size for T_2 and T_3 can then be determined as follows. L and C_{off} can be determined by solving (14.59) and (14.61), and, from C_{off}, the size of T_2 and T_3 can be selected. It is noted that, if very high cutoff frequency is chosen, L may be too small to be realized and C_{off}, hence T_2 and T_3 size, may be so small that gives poor isolation for the SPST switch. The size of T_1 is finally selected to produce an IL as low as possible while not affecting the isolation significantly.

14.6.2 Ultra-High-Isolation SPST Switch Design

The ultra-high-isolation SPST switch was designed and fabricated using Jazz 0.18-μm BiCMOS process [6]. All the on-chip inductors and interconnecting lines were designed and simulated using the full-wave electromagnetic simulator IE3D [7].

14.6.2.1 Core-SPST and Off-SPST Switch Design. The Core-SPST and Off-SPST switches are identical and their schematics are given in Figure 14.43. A small gate-bias resistor (R_g) of 200 Ω is chosen for possible high-speed switching. All other bias resistors (R_b) have the same value of 10 kΩ. R_{sub} is assumed to be very small, around 0.1 Ω, which is reasonable with a very good ground-to-substrate contact implemented in the layout.

Since the Core- and Off-SPST switches are designed to have a wide bandwidth up to 40 GHz, the cutoff frequency of the synthetic transmission line, f_c, is chosen to be 150 GHz. Solving Eqs. (14.59) and (14.61) with $Z_o = 50$ Ω and $f_c = 150$ GHz, we obtain $L = 106$ pH, which is used for all three inductors L_1, L_2, and L_3, and $C_{\text{off}} = 42.2$ fF. From C_{off}, the size of transistors T_2 and T_3 is found to be 75 μm, consisting of 15 fingers of 5-μm width. The size of transistor T_1 is chosen as 40 μm for low IL. The small size selected for T_1 also leads to a large input impedance for the Off-SPST switch to facilitate a good matching at the input of the proposed RF switch. Figure 14.44 shows the simulated IL, return loss and isolation of the core (or off) SPST

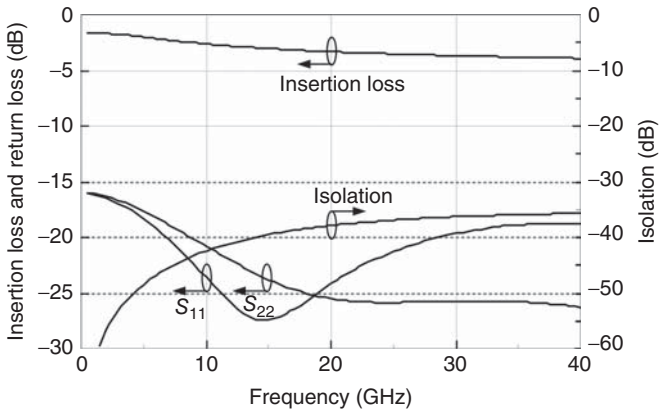


Figure 14.44. Simulated insertion loss, return loss, and isolation of the core (or off) SPST switch. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

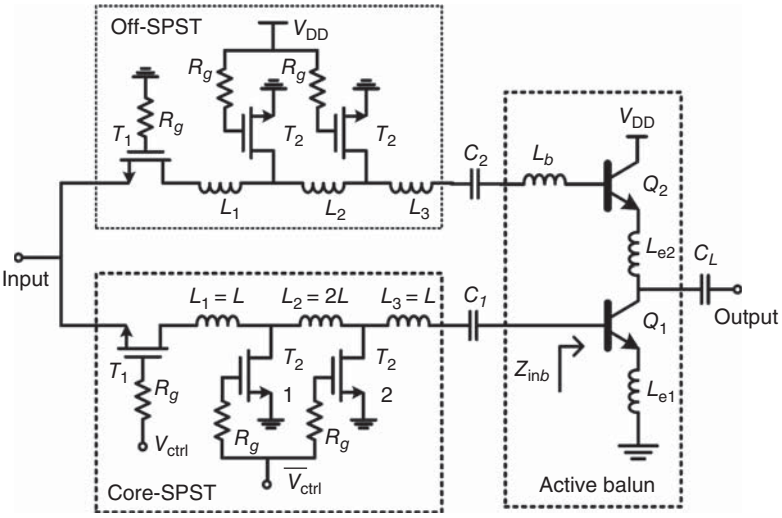


Figure 14.45. Complete schematic of the ultra-high-isolation SPST switch. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

switch. The simulated results show that the core (or off) SPST switch has an IL less than 3.9 dB, isolation higher 35 dB, and return losses more than 16 dB from DC up to 40 GHz.

14.6.2.2 Ultra-High-Isolation SPST Switch Design. Figure 14.45 shows the complete schematic of the ultra-high-isolation SPST switch. It consists of two identical SPST switches (core-SPST and off-SPST) described in Section 14.6.2.1 and an active balun described in Section 14.6.1.2. The Off-SPST switch is always in off-state while the Core-SPST switch functions as a normal on-off switch controlled by V_{ctrl} and $\overline{V}_{\text{ctrl}}$. The two HBT transistors Q_1 and Q_2 in the active balun have the same size with an emitter area of 2 μm × 8.28 μm and are biased at a DC current of 8 mA. The bases of Q_1 and Q_2 are biased at 0.92 and 1.8 V, respectively, through large resistors which do not affect the balun's RF operation. The collector of Q_2 is connected to V_{DD} of 1.8 V. The values of inductors L_{e1}, L_{e2}, and L_b are first estimated from Eqs. (14.73) and (14.74) and then optimized in Cadence simulation [8] at 35 GHz. Their final values are 130, 130, and 145 pH, respectively.

The input impedance of the Off-SPST switch in Figure 14.45 has to be much larger than that of the Core-SPST switch in its on-state, so when the ultra-high-isolation SPST switch is on, the Off-SPST switch will not degrade the overall SPST switch's input matching and hence its loss or gain. This requires a small size for the two series transistors T_1. The designed gate width of 40 μm for T_1 meets this requirement.

The active balun acts as a load to the Core-SPST switch and hence its input impedance $Z_{\text{in}\,b}$, as given in Eq. (14.67), directly affects the RF switch's input matching. Since the Core-SPST switch is designed as a 50-Ω synthetic transmission line and the frequency-dependent $Z_{\text{in}\,b}$ can be real only at a single design frequency, as can be seen from Eq. (14.78), a perfect match can only be possible at a single frequency. In order to obtain a good input matching for the RF switch across a wide bandwidth up to 40 GHz, the value of L, and hence L_1, L_2, and L_3, is optimized to get the best input matching at 25 GHz, instead of at the design frequency of 35 GHz. $L = 140$ pH is finally chosen. Table 14.3 lists the parameters of the ultra-high-isolation SPST switch.

The microphotograph of the ultra-high-isolation SPST switch is shown in Figure 14.46 with a chip area of 440 μm × 520 μm. The layout is made as symmetric as possible to enhance the balance of the active balun, which helps suppression of the leaking RF signal as much as possible.

14.6.2.3 Performance of the Ultra-High-Isolation SPST Switch. The unconditional stability of the ultra-high-isolation SPST switch is confirmed through simulations. Over 1-50 GHz, the switch's stability K and $|\Delta|$ factors are larger than 1.5 and smaller than 0.83, respectively. The ultra-high-isolation SPST switch was measured on-chip using a vector network analyzer (VNA). The measurements were calibrated using

TABLE 14.3. Ultra-High-Isolation SPST Switch's Parameters

Circuit elements and values			
T_1		40 μm gate width	
T_2, T_3		70 μm gate width	
Q_1, Q_2		2 μm × 8.28 μm emitter area	
L_1, L_3	140 pH	L_b	145 pH
L_2	280 pH	R_g	200 Ω
L_{e1}, L_{e2}	130 pH	R_b	10 kΩ
C_1, C_2, C_L	2 pF	V_{DD}	1.8 V

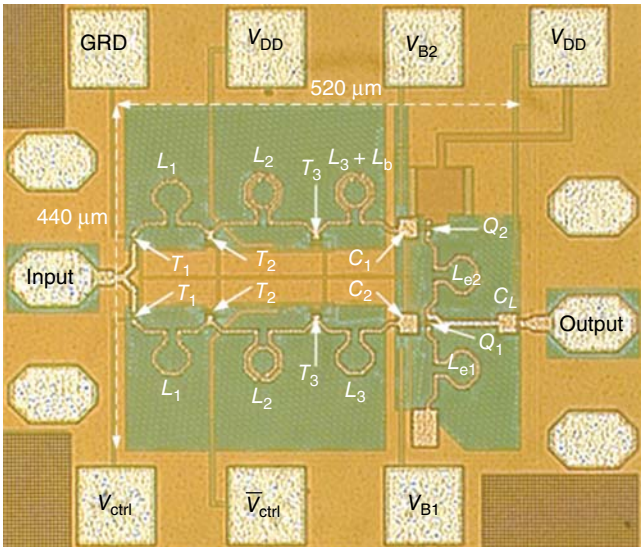


Figure 14.46. Microphotograph of the ultra-high-isolation SPST switch. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

the Short-Open-Load-Thru calibration method along with the impedance standard substrate calibration standards [9].

Figure 14.47 shows the simulated and measured IL/gain, reverse isolation (S12), input and output return loss, and isolation of the ultra-high-isolation SPST switch under small signal conditions. The measured results show that the switch exhibits an ultra-wideband, ultra-high-isolation performance while consuming a DC current of only 8 mA from a 1.8 V source. From 10 to 38 GHz, the loss/gain is from −2.6 (loss) to 0.4 dB (gain), the input return loss is from 8 to 20 dB, the output return loss is from 1 to 5 dB, the reverse isolation under on operation is from 21 to 35 dB, and the isolation goes from 40 to about 70 dB. The isolation approaches 75 dB as the frequency is reduced to below 3 GHz. The switch provides the highest gain of 0.4 dB at 30 GHz, maximum input return loss of 20 dB at 23.5 GHz and, especially, an extremely high isolation around 70 dB at 36 GHz. It is noted that the switch was designed without an output matching network, which results in a very poor output return loss. This, however, does not affect the balun performance. As noted in the balun analysis, the active balun can achieve its balance, hence RF leakage suppression leading to very high switching isolation, regardless of its load. Nevertheless, an output matching network should be included to improve the output return loss and gain for the switch.

Within the narrow band around 35.5 – 38.5 GHz, the isolation of the ultra-high-isolation SPST switch, reaching ultra-high values with the highest value of about 70 dB at 36 GHz, is found to be limited by the measured isolation limitation of the VNA as shown in Figure 14.47. The VNA's highest isolation level was measured by lifting the two RF probes into air and far away from each other. The switch may provide higher than 70 dB isolation at 36 GHz if measured with a VNA having higher isolation capability.

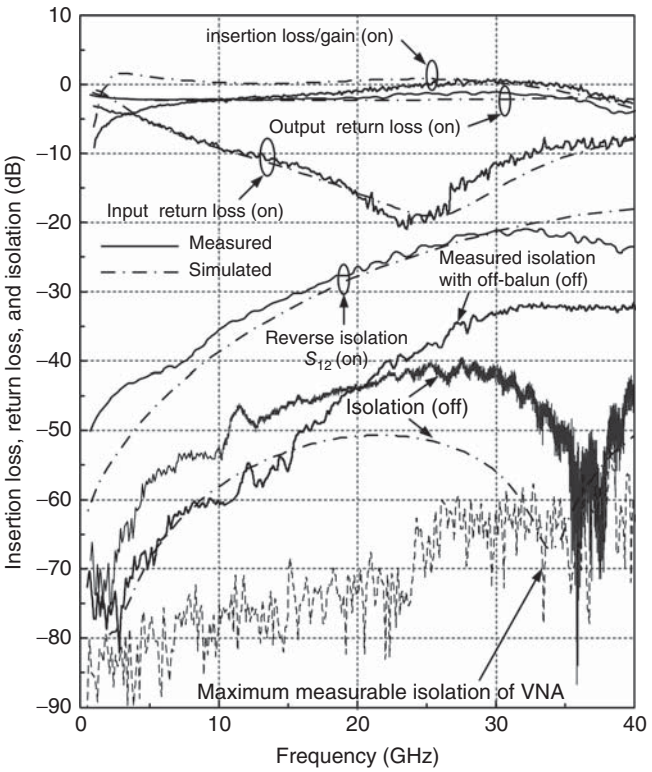


Figure 14.47. Simulated and measured insertion loss/gain, return loss, and isolation. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

The calculated isolation of the ultra-high-isolation SPST switch, as seen in Figure 14.47, is substantially higher than that of a conventional switch topology without the RF-leakage cancellation as shown in Figure 14.44, particularly at the design center frequency, demonstrating the superior isolation characteristic of the switch. At 34 GHz, it shows a calculated isolation of 67 dB as compared to 36 dB obtained from a conventional switch. It is noted that, the isolation of a conventional switch usually keeps decreasing as the frequency is increased, but the isolation of the ultra-high-isolation SPST switch, while also following a similar trend, has a particular window within which the isolation is increased substantially, reflecting the RF leakage suppression characteristic provided by the active balun.

Figure 14.47 also shows the measured isolation of the ultra-high-isolation SPST switch when the active balun is turned off by setting all of its DC-bias voltages to 0 V. It is observed that, from 25 to 40 GHz, the isolation of the switch with the active balun off is reduced to below 23 dB, while that with the active balun on is increased and, within the frequency band around the design frequency, reaches extremely high values, approaching more than 70 dB at 36 GHz. This clearly reveals the working function of the active balun in suppressing the RF leaking signal. When the active balun is off, both the leaking RF signals through the Off-SPST and Core-SPST switches go to the ultra-high-isolation SPST switch's output through the parasitic capacitors of the HBTs instead of being cancelled. At 36 GHz, the ultra-high-isolation SPST switch in its normal operation improves more than 38 dB of isolation than in the off-active-balun condition, which can be effectively considered the minimum isolation improvement with respect to using just the (conventional) Core-SPST switch.

While the measured and simulation results of the IL/gain, reverse isolation (S12), and input and output return losses are in good agreement from 10 to 40 GHz, the simulated isolation result shows more than 5-dB difference with the measured one, except in the narrow frequency range of 35.5 – 40 GHz. The difference in isolation is mainly due to the imbalance of the fabricated active balun, the control-signal bonding wires for the Core-SPST and Off-SPST switches, and the inaccuracy of the models for the passive elements and active devices. The imbalance of the active balun is primarily due to the circuit layout as well as the layout and

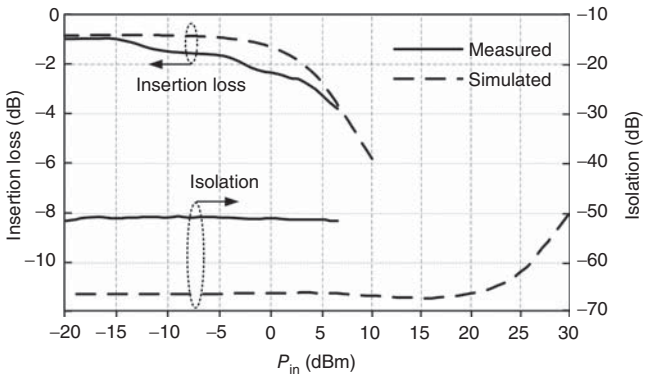


Figure 14.48. Measured and simulated insertion loss and isolation versus input power at 35 GHz. (After Huynh and Nguyen [5]. Reprinted with permission of IEEE.)

fabrication of the HBTs and inductors. This imbalance makes the conditions in (14.73) and (14.74) not very well satisfied. Moreover, although the constituent switches are laid-out symmetrically on-chip, the bonding wires at their control pins connected to the gates of the CMOS transistors through small resistors of 200 Ω affect their symmetry. As the ultra-high-isolation SPST switch is off, the asymmetry leads to two unequal leaking RF signals produced by the Core-SPST and Off-SPST switches. These unequal signals cannot be cancelled completely by the balun, hence reducing the isolation.

Figure 14.48 shows the measured and calculated IL and isolation of the ultra-high-isolation SPST switch versus input power at 35 GHz using a spectrum analyzer. The measured results show that, in the on-state, the switch has an IL from 1 to 2 dB and an input 1-dB compression power of −2.5 dBm. This low power compression is due to the fact that the linearity of the switch is primarily determined by the active balun's linearity. Either a higher-linearity active balun or a passive balun is needed to improve the switch's linearity. On the other hand, the linearity of the switch in its off-state is relatively high. The measured results show that the isolation of the switch varies from 51.3 to 51.85 dB for the input power from −20 to 6.5 dBm which is the maximum power available to the chip from our synthesizer. Within this input power range, the off-state 1-dB power compression does not occur. The simulated isolation does not compress until 22-dBm input power. This high power compression results since, up to 22 dBm, the balun is still operating in its linear region due to the small input power caused by the switch's isolation. The measured large-signal results are consistent with those measured under small-signal conditions.

14.7 FILTER SWITCHES

Switching and filtering functions are inevitable in RF systems and they are normally executed independently using separate switches and filters, respectively. For instance, a T/R switch is typically employed between the transmitter and receiver to share a common antenna, and a band-pass filter is normally used in front of the receiver's LNA. This typically employed approach requires spaces for both the T/R switch and band-pass filter (and hence large chip size) and suffers losses from these components, which directly affect the noise figure of the receiver. Combining both the filtering and switching functions into a single component working as a filter switch would help lessen these problems. This section addresses the principle and design of SPST filter switches, which form the basics for designing other multi-pole, multi-throw filter switches such as SPDT filter switches.

The principle of filter switches is relatively simple. A filter switch can be realized by imposing a filter function, such as low-pass, high-pass, band-pass and band-stop, on a switch. This very simple concept leads to possibilities of realizing filter switches using any filter and switch types and configurations (e.g., an elliptical band-pass filter and an SPDT with impedance inverters). These filter switches are made possible by recognizing that an off- and on-state MOSFET behaves primarily as a capacitor and resistor, respectively.

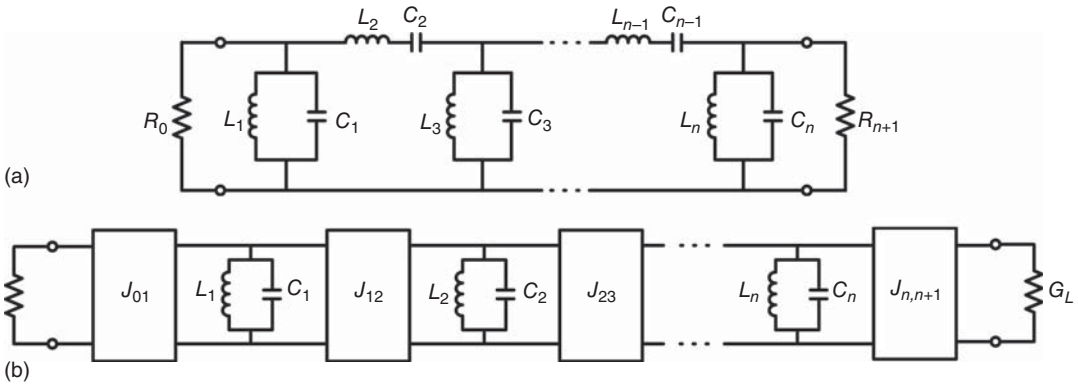


Figure 14.49. (a) Basic band-pass filter and (b) band-pass filter consisting of only shunt resonators and admittance inverters. J_{ok}, $k = 1, 2, \ldots, n+1$, are the admittance inverter parameters. $G_S = 1/R_S$ and $G_L = 1/R_L$ are the source and load conductance, respectively.

For illustration purposes, we consider a basic band-pass filter and a band-pass filter using admittance inverters as shown in Figure 14.49, which are discussed in Sections 8.5.4 and 8.5.6, respectively. In order to introduce the switching function into the filter responses, we replace each capacitor in the filters with a proper MOSFET, whose off-state capacitance is approximately equal to the capacitor's capacitance. Accordingly, the band-pass filters shown in Figure 14.49 become the SPST band-pass switches as shown in Figure 14.50. The SPST band-pass switches would have a band-pass filter response during their on-state and provide an isolation between the input and output ports when they are off by controlling the bias voltages to the transistors. Under the switches' on-state, the series and shunt MOSFETs are turned off, whereas, under the switches' off-state, the shunt MOSFETs are turned on and the series MOSFETs are turned either on or off. When the series MOSFET is off, it resonates with the series inductor, while, under its off state, it behaves approximately as a small resistor. Turning the series MOSFETs on, during the switch's on-state, is thus expected to increase the mismatch of the switch, and hence resulting in more isolation for the switch than turning them off. As discussed previously, series transistors affect more on the IL than the isolation of a switch. Therefore, for low IL, the series capacitors of the band-pass filter in Figure 14.49(a) should not be replaced with a MOSFET, particularly when the SPST band-pass switch is used in front of a LNA. The design of an SPST band-pass switch then starts with the design of a band-pass filter to meet certain specifications of pass-band frequency range, IL and rejections. Sections 8.5.4 and 8.5.6 provide information for designing some band-pass filters. Appropriate MOSFETs can then be used to simulate the resultant capacitors while considering loss

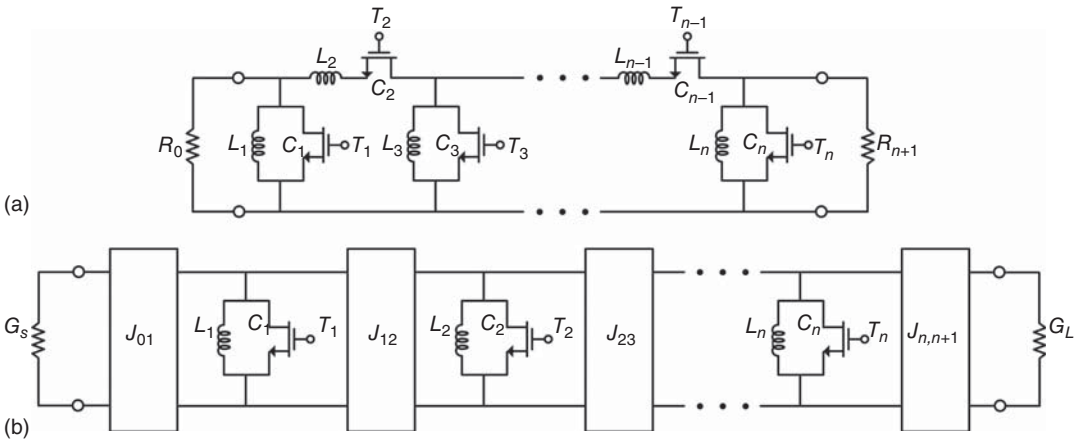


Figure 14.50. (a) Basic SPST band-pass switch and (b) SPST band-pass switch using admittance inverters corresponding to the band-pass filters in Figure 14.51. MIM capacitors may be used for series capacitors instead of MOSFETs.

and isolation for the switch. To minimize the IL, only shunt capacitors should be replaced with MOSFETs, whereas MIM capacitors, instead of MOSFETs, are used for series capacitors. Spiral inductors are typically used for all inductors.

REFERENCES

1. R. Xu, Y. Jin, and C. Nguyen, "Power-Efficient Switching-Based CMOS UWB Transmitters for UWB Communications and Radar Systems," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-54, No. 8, pp. 3271–3277, Aug. 2006.
2. M. -C. Yeh, Z. -M. Tsai, R. -C. Liu, K. -Y. Lin, Y. -T. Chang, and H. Wang, "Design and Analysis for a Miniature CMOS SPDT Switch Using Body-Floating Technique to Improve Power Performance," *IEEE Trans. Microwave Theory Tech.*, Vol. 54, No. 1, pp. 31–39, Jan. 2006.
3. TSMC 0.18-μm CMOS Process, MOSIS Foundry, Marina del Rey, California, USA.
4. Y. Jin and C. Nguyen, "Ultra-Compact, High-Linearity, High-Power Fully-Integrated DC-20 GHz 0.18-μm CMOS T/R Switch," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-55, No. 1, pp. 30–36, Jan. 2007.
5. C. Huynh and C. Nguyen, "New Ultra-High-Isolation RF Switch Architecture and Its Use for a 10–38 GHz 0.18-μm BiCMOS Ultra-Wideband Switch," *IEEE Trans. Microwave Theory Tech.*, Vol. MTT-59, pp. 345–353, Feb. 2011.
6. Jazz Semiconductor, 4321 Jamboree Road, Newport Beach, California, USA.
7. IE3D, IC Nanometer Design Tool, Mentor Graphics, Wilsonville, Oregon, USA.
8. Cadence Design Systems, San Jose, California, USA.
9. Cascade Microtech, Beaverton, Oregon, USA.

PROBLEMS

14.1 Derive Eq. (14.5).

14.2 Derive Eq. (14.8).

14.3 Derive Eqs. (14.9) and (14.10).

14.4 Consider MOSFETs in any available submicron CMOS process suitable for 20 GHz operation. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation (in decibel) at 20 GHz of a shunt MOSFET versus the numbers of fingers or size of the transistor. Any available finger width can be used. Use a large resistor (e.g., 2 kΩ) for the gate bias.

14.5 Repeat Problem 14.4 using the MOSFET's on- and off-state impedances along with (14.3) and (14.4). Compare and comment on the results with those of Problem 14.4.

14.6 Repeat Problem 14.4 for a series MOSFET.

14.7 Repeat Problem 16.6 using the MOSFET's on- and off-state impedances along with (14.6) and (14.7). Compare and comment on the results with those of Problem 14.6.

14.8 Consider MOSFETs in any available submicron CMOS process suitable for 20 GHz operation. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation contours at 20 GHz for a pair of series and shunt MOSFET for different finger numbers of series and shunt the transistor (similar to Figure 14.7). Any available finger width can be used. Use a large resistor (e.g., 2 kΩ) for the gate bias. Discuss the results. Use this figure to choose optimum sizes for the series and shunt MOSFET according to certain isolation and IL of your choice.

14.9 Repeat Problem 16.8 using the MOSFET's on- and off-state impedances along with (14.9) and (14.10). Compare and comment on the results with those of Problem 14.8.

14.10 Consider MOSFETs in any available submicron CMOS process suitable for operation up to at least 60 GHz. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation (in decibel) versus frequency from 1 to 40 GHz in 0.5-GHz steps for one shunt MOSFET and two, three, and four shunt MOSFETs. Comment on the results with respect to switch design.

14.11 Repeat Problem 14.10 for one series MOSFET and two and three series MOSFETs. Comment on the results with respect to switch design.

14.12 Repeat Problem 14.10 for two cases: Case 1: one series MOSFET and one shunt MOSFET (next to each other) and Case 2: two series MOSFETs and two shunt MOSFETs (next to each other). Comment on the results with respect to switch design.

14.13 Derive Eqs. (14.21) and (14.22).

14.14 Consider MOSFETs in any available submicron CMOS process suitable for operation up to at least 60 GHz and an SPST switch consisting of two identical shunt MOSFETs spaced apart by an ideal admittance inverter having $J = Y_o$ as shown in Figure 14.11. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation (in decibel) versus frequency from 1 to 40 GHz in 0.5-GHz steps for three cases: Case 1: IL and isolation of the SPST switch. Case 2: IL and isolation of one shunt MOSFET. Case 3: isolation given by $(1/4)\text{ISO}_1\text{ISO}_2$, where ISO_1 and ISO_2 are the isolations of the individual MOSFETs. Compare and comment on the results.

14.15 Repeat Problem 14.14 for the isolation using Eqs. (14.28), (14.30), and (14.31). Compare and comment on the results. If you also do Problem 14.14, then compare and comment on the results obtained in this problem and those in Problem 14.14.

14.16 Derive Eq. (14.33).

14.17 Derive Eqs. (14.36) and (14.37).

14.18 Consider MOSFETs in any available submicron CMOS process suitable for operation up to at least 60 GHz and an SPST switch consisting of two identical series MOSFETs spaced apart by an ideal impedance inverter having $K = Z_o$ as shown in Figure 14.13. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation (in decibel) versus frequency from 1 to 40 GHz in 0.5-GHz steps for three cases: Case 1: IL and isolation of the SPST switch. Case 2: IL and isolation of one series MOSFET. Case 3: isolation given by $(1/4)\text{ISO}_1\text{ISO}_2$, where ISO_1 and ISO_2 are the isolations of the individual MOSFETs. Compare and comment on the results.

14.19 Repeat Problem 14.18 for the isolation using Eqs. (14.44), (14.46), and (14.31). Compare and comment on the results. If you also do Problem 14.18, then compare and comment on the results obtained in this problem and those in Problem 14.18.

14.20 Derive Eqs. (14.48)–(14.50).

14.21 Derive Eqs. (14.51)–(14.56).

14.22 Derive Eqs. (14.31) and (14.32) for an SPST switch consisting of two pairs of series/shunt MOSFETs separated by an impedance inverter as shown in Figure 14.14, where ISO_1 and ISO_2 are the isolations of the two constituent switches, each consisting of MOSFET 1 in series and MOSFT 2 in shunt.

14.23 Consider MOSFETs in any available submicron CMOS process suitable for operation up to at least 60 GHz and an SPST switch consisting of two pairs of series/shunt MOSFET separated by an impedance inverter having $K = Z_o$ as shown in Figure 14.14. Use the MOSFET's equivalent-circuit model available in the PDK and a commercial computer program, simulate and plot the IL and isolation (in decibel) versus frequency from 1 to 40 GHz in 0.5-GHz steps for three cases: Case 1:

IL and isolation of the SPST switch. Case 2: IL and isolation of one pair of series/shunt MOSFET. Case 3: isolation given by $(1/4)ISO_1ISO_2$, where ISO_1 and ISO_2 are the isolations of the individual pairs. Compare and comment on the results.

14.24 Repeat Problem 14.23 for the isolation using Eqs. (14.52) and (14.31). Compare and comment on the results. If you also do Problem 14.23, then compare and comment on the results obtained in this problem and those in Problem 14.23.

14.25 Derive expressions for the IL and isolation of the SPDT switch as shown in Figure 14.20.

14.26 Design an SPST switch consisting of series and shunt MOSFETs as shown in Figure 14.5 to achieve a maximum IL of 4 dB, minimum isolation of 20 dB and minimum return loss of 10 dB across 33 – 37 GHz using any available suitable submicron CMOS process. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation using a procedure similar to that presented in Figure 14.7. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.
c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 35 GHz when the switch is on and, from which, identify the 1-dB input compression power.
e) Plot and discuss the IL and isolation versus input power at 35 GHz.
f) Simulate the RF pulse of a CW signal at 35 GHz by gating the switch, and from which determine the rising and falling times, and hence the switching time, at 35 GHz.

14.27 Repeat Problem 14.26 for an SPST switch consisting of two shunt MOSFETs spaced apart by a quarter-wavelength admittance inverter as shown in Figure 14.11 to achieve a maximum IL of 5 dB, minimum isolation of 25 dB and minimum return loss of 10 dB across 33 – 37 GHz.

14.28 Repeat Problem 14.27 using an inverter as shown in Figure 7.41(b) or Figure P5.5, whichever is more suitable.

14.29 Repeat Problem 14.26 for an SPST switch consisting of two pairs of series/shunt MOSFET spaced apart by an impedance inverter as shown in Figure 14.14 to achieve a maximum IL of 6 dB, minimum isolation of 40 dB and minimum return loss of 10 dB across 33 – 37 GHz. Any inverter can be used for the design.

14.30 Design a T/R switch consisting of shunt MOSFETs and admittance inverters as shown in Figure 14.15(c) to achieve a maximum IL of 4 dB (in on-path), minimum isolation of 20 dB (in off-path), and minimum return loss of 10 dB across 33 – 37 GHz using any available suitable submicron CMOS process. Any inverter can be used. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.

c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 35 GHz for the on-path with another path off, and from which identify the 1-dB input compression power.
e) Plot and discuss the IL and isolation versus input power at 35 GHz.

14.31 Repeat Problem 14.30 for a T/R switch consisting of series MOSFETs and series/shunt MOSFETs spaced apart by impedance inverters as shown in Figure 14.15(d) to achieve a maximum IL of 7 dB (in on-path), minimum isolation of 40 dB (in off-path), and minimum return loss of 10 dB across 33 – 37 GHz.

14.32 Design an ultra-wideband SPST switch, consisting of one series and two shunt MOSFETs, using synthetic transmission lines such as that shown in Figure 14.20 to achieve a maximum IL of 5 dB, minimum isolation of 35 dB and minimum return loss of 15 dB across 1 – 40 GHz using any available suitable submicron CMOS process. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.
c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 10, 20, and 40 GHz when the switch is on and, from which, identify the 1-dB input compression powers.
e) Plot and discuss the IL and isolation versus input power at 10, 20, and 40 GHz.
f) Simulate the RF pulse of a CW signal at 35 GHz by gating the switch, and from which determine the rising and falling times, and hence the switching time, at 35 GHz.

14.33 Design an ultra-wideband T/R switch, consisting of one series and one shunt MOSFET in each arm, using synthetic transmission lines such as that shown in Figure 14.30 to achieve a maximum IL of 5 dB, minimum isolation of 25 dB, and minimum return loss of 10 dB across 1 – 40 GHz using any available suitable submicron CMOS process. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.
c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 10, 20, and 40 GHz for the on-path with another path off, and from which identify the 1-dB input compression powers.
e) Plot and discuss the IL and isolation versus input power at 10, 20, and 40 GHz.

14.34 The ultra-high-isolation SPST switch presented in Section 14.6.1 uses an active balun. Discuss the advantages and disadvantages if a passive balun is used in place of the active balun.

14.35 Design an ultra-high-isolation SPST switch based on the architecture as shown in Figure 14.39 using any available suitable submicron CMOS process. The core-SPST and off-SPST switches consist of

series and shunt MOSFETs, which are the same as the SPST switch designed in Problem 14.24. Either active or passive balun can be used. The switch is designed to operate from 33 to 37 GHz with a maximum IL of 5 dB, minimum isolation of 55 dB, and minimum return loss of 10 dB. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.
c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 35 GHz when the switch is on and, from which, identify the 1-dB input compression power.
e) Plot and discuss the IL and isolation versus input power at 35 GHz.
f) Simulate the RF pulse of a CW signal at 35 GHz by gating the switch, and from which determine the rising and falling times, and hence the switching time, at 35 GHz.

14.36 Consider an ultra-high-isolation SPDT switch as shown in Figure P14.1, which is based on the ultrahigh-isolation SPST switch architecture in Figure 14.39. Describe its operation and discuss its advantage as compared to a conventional SPDT switch consisting of a single SPST switch in each arm. Which systems and where in the systems do you see this SPDT switch could be used with significant advantages to the systems?

14.37 Consider an ultra-high-isolation T/R switch as shown in Figure P14.2 with TX and RX ports representing the transmitter and receiver ports, respectively. The Core-SPST and Off-SPST switches are the same for both transmit and receive paths. However, the active baluns are different. The baluns in the transmit and receive paths are designed to behave as a PA and LNA, respectively. Describe the T/R switch's operation and discuss its advantage as compared to a conventional T/R switch consisting of a single SPST switch in each arm. Why an extra SPST switch is used after the active balun in the transmit path? What would happen if it is removed?

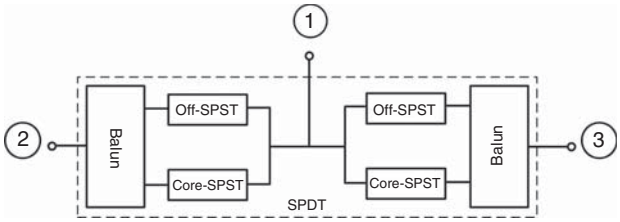


Figure P14.1.

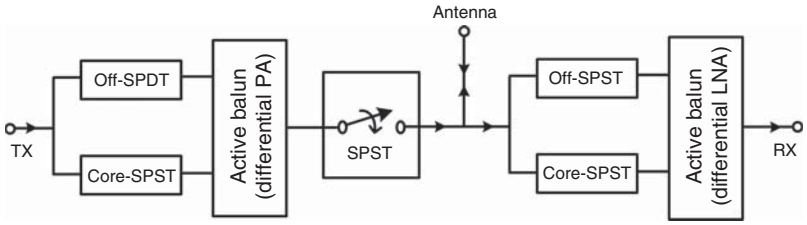


Figure P14.2.

14.38 Design an ultra-high-isolation SPDT switch based on the architecture as shown in Figure P14.1 using any available suitable submicron CMOS process. The core-SPST and off-SPST switches consist of series and shunt MOSFETs, which are the same as the SPST switch designed in Problem 14.26. Either active or passive balun can be used. The switch is designed to operate from 33 to 37 GHz with a maximum IL of 5 dB, minimum isolation of 55 dB, and minimum return loss of 10 dB. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.
c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.
d) Plot the output power versus input power at 35 GHz for the on-path with another path off, and from which identify the 1-dB input compression power.
e) Plot and discuss the IL and isolation versus input power at 35 GHz.

14.39 Draw the schematic of the SPST band-stop filter corresponding to the band-stop filter as shown in Figure 8.59.

14.40 Draw the schematic of the SPST low-pass filter corresponding to the low-pass filter as shown in Figure 8.51(b).

14.41 Design an SPST low-pass switch based of the low-pass prototype as shown in Figure 8.51(b) using any available suitable submicron CMOS process. The switch has the following specifications: 0.1-dB pass-band ripple, 10-GHz cutoff frequency, 30-dB minimum rejection at 20 GHz, 5-dB maximum IL, 25-dB minimum isolation, and 10-dB minimum return loss. The transistors should be selected to produce proper IL and isolation. Ideal inductors and capacitors can be used.

a) Draw the switch's schematic with all element values.
b) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation.

14.42 Design an SPST band-pass switch based on the Chebyshev band-pass filter having two resonators and three admittance inverters as shown in Figure P14.3 using any available suitable submicron CMOS process. Any kind of admittance inverters can be used. The switch has the following specifications: 0.1-dB pass-band ripple, 30 – 40 GHz pass-band, 5-dB maximum IL, 35-dB minimum isolation, and 10-dB minimum return loss. There is no requirement for switching speed. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.
b) Prepare a layout of the switch ready for tape-out for fabrication.

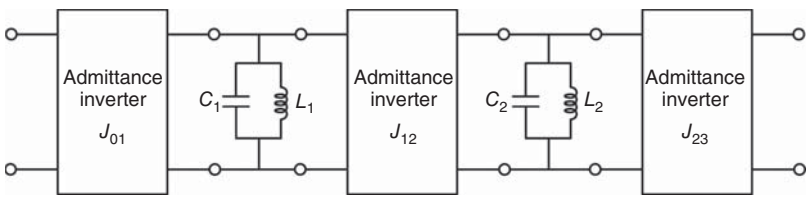


Figure P14.3.

c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.

d) Plot the output power versus input power at 30, 35, and 40 GHz when the switch is on and, from which, identify the 1-dB input compression powers.

e) Plot and discuss the IL and isolation versus input power at 30, 35, and 40 GHz.

f) Simulate the RF pulse of a CW signal at 35 GHz by gating the switch, and from which determine the rising and falling times, and hence the switching time, at 35 GHz.

14.43 Consider an SPDT band-pass switch topology having two shunt resonators and three admittance inverters in each arm as shown in Figure P14.4. Describe its operation.

14.44 Consider an SPDT band-pass switch topology as shown in Figure P14.5, in which the first resonator is shared between two paths. Design this switch using any available suitable submicron CMOS process for the following specifications: 0.1-dB pass-band ripple, 30-40 GHz pass-band, 5-dB maximum IL, 30-dB minimum isolation, and 10-dB minimum return loss. There is no requirement for switching speed. Any suitable inverter can be used. The design should include input and output lumped-element matching networks as needed with real inductors designed using an EM simulator. Capacitors available in the PDK can be used. The transistors should be selected to produce proper IL and isolation. Both circuit and EM simulators need to be used in the design. Perform post-layout simulations and reoptimize the circuit, after the post-layout simulations, if needed to improve the performance.

a) Draw the switch's schematic with all element values.

b) Prepare a layout of the switch ready for tape-out for fabrication.

c) Plot and discuss the switch performance versus frequency including IL, return loss, and isolation under small- and large-signal conditions.

d) Plot the output power versus input power at 20, 30, and 40 GHz for the on-path with another path off, and from which identify the 1-dB input compression powers.

e) Plot and discuss the IL and isolation versus input power at 20, 30, and 40 GHz.

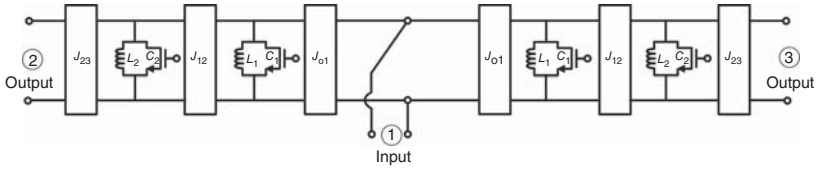


Figure P14.4.

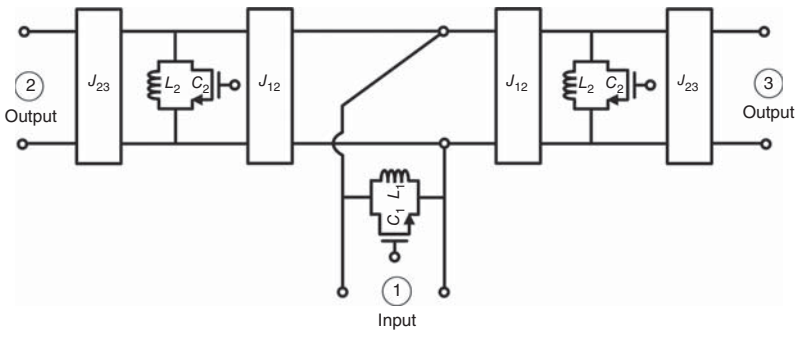


Figure P14.5.

14.45 Consider an SPST switch employing a single MOSFET modeled as a 1-Ω resistor (on-state) and a 1-Ω resistor in series with a 0.1-pF capacitor (off-state) between the drain and source terminals in a 50-Ω system.

a) Calculate the IL and isolation (in decibel) at 6 GHz when the MOSFET is connected in series.

b) Calculate the IL and isolation (in decibel) at 6 GHz when the MOSFET is in shunt.

14.46 Repeat Problem 14.45 with two MOSFETs connected in series and shunt. Compare and comment on the results with those of Problem 14.45.

14.47 A SPST switch consists of two shunt MOSFETs spaced apart by a 50-Ω quarter-wave transmission line acting as an admittance inverter. The MOSFET is modeled as a 2-Ω resistor (on-state) and a 3-kΩ resistor in parallel with a 0.2-pF capacitor (off-state) between the drain and source terminals in a 50-Ω system.

a) Calculate the IL and isolation at 10 GHz.

b) Compare and comment on the performance with that of an SPST switch using only one MOSFET in shunt.

c) Determine the shunt inductance to compensate for the off-state capacitances of the two shunt MOSFETs. Compare and comment on the performance in Part (a) with that obtained when the compensating inductance is used.

CHAPTER 15

RFIC SIMULATION, LAYOUT, AND TEST

The design of RFICs involves five main tasks: circuit design, simulation, layout, fabrication, and test. Each of these tasks is time-consuming and needs to be conducted carefully in order to achieve a successful design. Particularly, the design, simulation and layout are a combined interactive and iterative process, in which, several iterative circuit designs, simulations using circuit and EM simulators, and layouts need to be conducted, often in an interactive manner, in order to arrive at a successful design meeting required specifications. The following steps show a general design procedure for RFICs including circuit design, simulation, layout, post-layout simulation, fabrication and test:

General RFIC Design Procedure

1. Determine proper circuit topology.
2. Design circuit based on ideal passive elements (e.g., ideal inductors) and active devices (diodes, transistors) available in the process design kit (PDK) provided by the fabrication foundry.
3. Simulate and optimize circuit with ideal passive elements and active devices.
4. Check if meeting circuit specifications. If not, repeat steps 2, 3, and 4 until meeting specifications and proceed to step 5.
5. Replace ideal passive elements with real elements from the PDK and/or real elements designed and modeled using an EM simulator or actual measurement.
6. Simulate and optimize circuit with real passive elements and PDK transistor models.
7. Check if meeting circuit specifications. If not, redesign and optimize circuit until specifications are met and proceed to step 8. The final simulation at this step is considered the "schematic simulation.[1]"
8. Prepared circuit layout.
9. Perform EM simulations of all passive elements including inductors, capacitors, vias, transmission lines, interconnects, and so on based on their actual layouts.

[1]Schematic simulation is loosely defined. Depending on a designer, other schematic simulations such as simulation based only on ideal elements can also be specified.

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

10. Conduct post-layout simulation of the circuit incorporating EM-simulated data of passive elements and extracted parasitic components,[2] such as resistances, inductances, and capacitances, resulted from the layout of all passive elements.
11. Check if meeting circuit specifications. If not, redesign and optimize circuit and layout until specifications are met and proceed to step 13. The final simulation at this step is considered the "post-layout simulation.[3]"
12. Perform layout rule check and other tests required by the fabrication process.
13. Send final layout to foundry for fabrication.
14. Perform on-wafer measurement of fabricated chips. For off-chip tests, go to step 15.
15. Package chips and prepare printed-circuit boards (PCB) for mounting packaged chips for testing purposes.[4]
16. Compare measured and simulated results. Resimulate circuit using updated data such as on-chip measured data for transistors as needed to assess the circuit's calculated performance.

15.1 RFIC SIMULATION

Nonlinear, linear, and EM simulations play a dominant role in the design and analysis of RFICs. To facilitate the effective development of high-performance RF circuits, highly accurate and efficient computer-aided simulations are needed. This need is even more pronounced for the design of RFICs due to the extreme difficulty in tuning these circuits after fabrication. RFIC simulations are typically done using commercially available computer-aided design (CAD) programs. For instance, the simulations in the design examples of RFICs presented in this book are obtained with the circuit simulators in Agilent ADS [1] and Cadence [2], and the EM simulator IE3D [3]. RFIC designers typically conduct two different kinds of simulations: schematic simulation and post-layout simulation as mentioned in the foregoing general design procedure. Schematic simulation is not based on the circuit layout while post-layout simulation is typically (and desirably) based on the layout of circuit elements that may include active devices (at least the interconnects between the core active devices and surrounding circuit elements.)

Most commercially available CAD software provides an integrated circuit design environment, which includes schematic editor, simulation engines, model library of circuit elements, and layout tool. Usually circuit optimization and synthesis are also contained in commercial software packages. Some software packages also consist of EM simulators.

The schematic editor is used to form schematics of circuits being designed. In the schematic editor, circuit components represented by graphic symbols are picked from the component library and a wire tool is used to draw lines to connect these components with each other and to the circuit's input and output terminals. For large circuits, the circuit schematic can be organized hierarchically; schematics of sub-circuits making up the circuit can be defined as separate function blocks with input/output terminals which are connected to form the entire circuit. The component library provides models for circuit components that can be used for those in the circuit being designed. The accuracy of these built-in models affects the simulation accuracy of the CAD program. CAD programs usually include a set of frequently used ideal components. The models of ideal component are mainly used to verify the circuit topology workability and/or estimate the limit of the circuit performance. Some CAD programs also include models of actual (non-ideal) passive and active components, which are more accurate than the ideal ones. In general, the actual component models used in a certain process is provided by the fabrication foundry in its PDK. The models of the active and passive devices in a PDK are typically obtained based on the actual measurement of fabricated devices, so they are strongly

[2]Parasitic extraction may not be needed when EM simulations are used for the entire layout since EM simulations should already include parasitic effects. Certain parasitic extractions may not be available in some simulators.
[3]Post-layout simulation is also loosely defined – some designers may not include all the EM effects in a post-layout simulation.
[4]For off-chip characterization, extra elements used in packaging such as bonding wires, package, and interconnects and connectors for PCB need to be taken into account in the design and simulation.

related to the fabrication process and more accurate than other models based on calculations available in CAD programs. Some CAD programs also include the PDKs of certain processes and these PDKs appear in the CAD programs under certain symbols with necessary parameters. The actual structures and parameters of these PDKs' device models are often hidden from the program users. Once a schematic is formed, it is sent to the simulation engine in the CAD program to evaluate the performance of the circuit. Various simulation methods are available for checking different parameters of circuits as briefly described below.

15.1.1 DC Simulation

DC simulation is a necessary step to compute the bias point of a circuit. During a DC simulation, all ideal capacitors and inductors are considered open and short, respectively. Besides calculating the static node voltages and currents, a DC simulation also gives the small-signal model parameters of transistors such as the transconductance and gate-source capacitance. Therefore, the simulator engine in a CAD program typically runs a DC simulation first before running other simulations. It is worth noticing that a calculated DC current is different from an actual current consumption of a circuit. The discrepancy between them depends on the AC current swing in the circuit.

15.1.2 Small-Signal AC Simulation

Small-signal AC simulation examines the small-signal performance of circuits. It is performed in the frequency domain over interested frequency ranges. The AC signal amplitude is assumed to be sufficiently small so that the DC operating point is not affected. The small-signal AC simulation is used to calculate small-signal transfer functions such as voltage gain, current gain, transimpedance, and noise. Since this kind of simulation omits all the nonlinear effects of devices, the impact of inter-modulation cannot be observed from the simulation. Small-signal scattering (S)-parameter simulation can be considered a type of small-signal AC simulations. Noise figures of components, which are based on linear noise models, can also be calculated during S-parameter simulations. S-parameters are typically needed for most RFIC components. In RFIC design, particularly in the high RF range, the interconnections between different circuits or between different elements/components in a circuit need to be taken in account and this is usually done using their S-parameters.

15.1.3 Transient Simulation

Transient simulation is used to obtain the responses of circuits to arbitrary excitations in the time domain. Transient simulation is especially useful when simulating circuits such as mixer, sampler, and VCO, and for others whose signal integrity needs to be examined. Nonlinear models of active devices, typically described in the time domain, are used in transient simulations. As such, transient simulations can produce the nonlinear behavior of circuits. Most simulators do not include noise in the transient simulation since the noise models are typically in the frequency domain. For passive components, whose responses are expressed in terms of the S, Z, or Y parameters, transient simulations will treat them as linear components. Theoretically, the frequency-domain response of a circuit can be obtained by simply taking the Fourier transform of the time-domain waveform which includes all nonlinear effects. This, however, may cause some computation problems. As the time-step used in a transient simulation is determined by the maximum frequency stimulus in a circuit, if the circuit contains both low and high frequency stimulus, then the time-step needs to be small enough to capture the fast transition caused by high frequencies. At the same time, the total time-duration of the simulation needs to be sufficiently long so that low-frequency changes can show up in the response. As a result, an enormous amount of data point is required, which might significantly slow down the simulation process or even make the simulation not feasible.

15.1.4 Periodic Steady State Simulation

Periodic Steady State (PSS) analysis is a large-signal analysis which can produce the frequency-domain performance of circuits. Since it can include the nonlinear distortion of circuits, PSS simulation is one of the most

useful tools in the frequency-domain analysis. Spectre in Cadence [1] is one of the simulation engines that use PSS simulation for calculating frequency responses. As the name indicates, PSS simulation only deals with periodic signals in steady state after the transient response diminishes. It is noted that periodic signals include single-frequency signals and multi-frequency signals such as a train of pulses. In PSS simulation, a time-domain simulation is performed over a designated period and a frequency response is obtained by conducting a Fourier transformation for the calculated time-domain results. PSS simulation is also a necessary step before running periodic AC analysis (PAC), periodic transfer function analysis (PXF), and periodic noise analysis (PNOISE) simulations. During a PSS simulation of a nonlinear circuit, if a single tone is used as the stimulus for the nonlinear circuit, then the single tone's frequency will be the base frequency and the calculated frequency response will demonstrate all the harmonics generated by the nonlinear circuit. Since the response is assumed to be periodic, the spectrum only includes discrete spikes with the interval of the single-tone frequency. If the stimulus of the circuits contains two different frequencies, a common factor of these two frequencies needs to be found as the base frequency for the PSS simulation. If the PSS simulation runs into nonconvergent situation, adjusting the error tolerance and iteration method can mostly likely solve the problem. It is obvious that the PSS simulation has its limitation due to the fact that it only deals with periodic signals. For circuits containing arbitrary frequency components, PAC or quasi-periodic steady state (QPSS) simulation needs to be used.

15.1.5 Harmonic-Balance Simulation

Harmonic-balance simulation is another algorithm to analyze the steady-state responses of circuits and considered one of the most widely used nonlinear simulation methods for RFICs. Harmonic-balance technique is a nonlinear tool for the analysis and design of nonlinear RF circuits operating in the time-periodic steady-state regime. Basically, the harmonic-balance method divides a nonlinear RF circuit into a linear subnetwork and a nonlinear subnetwork. The former contains all linear elements of the RF circuit, such as resistors, inductors, capacitors, distributed elements, and associated discontinuities, represented in the frequency domain. The latter consists of only nonlinear elements, which are typically those of semiconductor devices, described in the time domain by a set of relationships. The method takes advantages of both the frequency- and time-domain analysis through implementation of the most natural and straightforward approach to the simulation of nonlinear RF circuits, in which the linear components and nonlinear components are evaluated in the frequency domain and time domain, respectively. It avoids several inherent problems existing in the time-domain method. First, the use of the frequency-domain description for distributed elements, the most efficient and accurate way in describing passive structures, results in simple models which can reduce significantly computational complexity. In addition, a majority of models for distributed elements and their associated effects such as dispersion and discontinuities are available in the frequency domain and thus can be included without any further significant development. Second, because the linear subcircuit is partitioned separately and analyzed in the frequency domain, elements with large time constants do not slow down the convergence to a solution. Finally, the harmonic-balance method analyzes the nonlinear problems directly in the steady state and therefore avoids the computationally expensive process of transient evaluation through numerical integration of the dynamic network equations. The harmonic balance is applicable to strongly as well as weakly nonlinear RF circuits.

The harmonic balance method solves a nonlinear problem by iteration or a repetitive analysis process. That is, a certain set of voltages is taken as the starting point, and the linear part of the circuit is solved in the frequency domain. The response of the nonlinear portion of the circuit is then computed in the time domain. Comparisons of the currents at the interface between the linear and nonlinear subnetworks of the circuit at the fundamental and harmonics are made using a Fourier transform to shuttle between the time domain solution of the nonlinear part and the frequency domain solution of the linear part. In a series of iterations, differences are computed and adjustments made so that the next iteration provides a closer agreement. This process continues until the currents are closely matched or balanced.

15.1.6 Periodic Distortion Analysis

Periodic distortion analysis, also known as QPSS analysis, allows simulations of nonlinear circuits' frequency response even when the response is not periodic. In QPSS simulations, input signals with arbitrary frequencies can be used as stimulus. In QPSS simulations, one base signal has to be assigned as large signal.

15.1.7 Envelope Simulation

Envelope simulation is especially useful to simulate based-band modulated RF signals. The circuit response may contain two frequencies separated far apart. The PSS or QPSS simulation might not be capable to handle this situation. A typical use of envelope simulation is for digital RF modulators.

15.1.8 Periodic Small Signal Analysis

Periodic small-signal analysis includes a series of simulations: periodic AC analysis (PAC), periodic transfer function analysis (PXF), periodic noise analysis (PNOISE), and periodic *S*-parameter analysis (PSP). It conducts small-signal analysis based on a solved steady-state-of-circuit response. A PSS simulation is required beforehand. Periodic small-signal analysis is especially useful for circuits having frequency translation such as mixers. Periodic small signal is integrated in Spectre, the simulation engine of Cadence, along with PSS simulation.

15.1.9 EM Simulation

EM simulations, which are based on the actual layouts of circuit elements such as spiral inductors, are important and, in fact, inevitable for RFICs, particularly in the high RF range. In EM simulations, *S*-parameters are typically obtained for passive circuit elements and used in the overall circuit simulations. It is important to note that not only actual individual elements such as inductors or interconnects in a circuit layout need to be analyzed using an EM simulator, but also the combination of all these passive elements integrated as appeared in the circuit layout needs to be simulated, considering all effects such as discontinuities and mutual coupling, in order to obtain more accurate results. Moreover, it is essential that grounds that are subjected to RF signals (i.e., AC or RF ground) need to be completely taken into account in EM simulations. These grounds may affect the RFIC performance substantially and, under certain conditions and frequencies, can even disrupt the circuit operation totally, and hence cannot be treated as an ideal ground or a simple ground node having zero voltage. Instead, they need to be considered a distributed ground having a finite area with varying RF voltage distribution and parasitics. This is especially crucial for RFICs operating in high RF ranges such as millimeter-wave frequencies and/or having large areas. Typical PDKs do not include EM simulations for passive elements such as inductors. Some advanced PDKs, however, may include passive elements with EM simulations. All PDKs for RFICs should, or are expected to, include EM simulations for all their circuit elements, making RFIC design more convenient and accurate. Besides the *S*-parameters, EM simulators can also generate equivalent-circuit models of circuit elements from the built-in or user-defined models using the calculated *S*-parameters through an optimization process – for instance, a Pi-equivalent circuit for spiral inductors. In order to obtain accurate results for simulations, accurate data need to be provided to the EM simulator for the electrical and physical parameters of the circuit elements and CMOS (BiCMOS) metal/dielectric structures. These can be obtained from the foundry's PDK. The simulated *S*-parameters results can be saved in an *S*-parameter file, which can be used to describe a single- or multi-port network. The single- or multi-port network characterized by the *S*-parameter file or the equivalent-circuit model over interested frequencies can be used in a circuit simulator along with other elements of the circuit to simulate the overall circuit performance. For frequencies that are not included in the EM simulation, the simulator can calculate the performance at these frequencies by interpolation. A transient simulation can also be performed for the EM-simulated circuit element by means of the inverse fast Fourier transform (IFFT).

EM simulators typically include a layout drawing tool and a simulation engine. Although layout capability is available in EM simulators, it is strongly recommended that the layouts of elements be created in a layout editor capable of design rule check (DRC). Otherwise, designers may need to switch between the employed EM simulator and the IC layout editor to create an acceptable layout. The layouts created throughout this book are done in Cadence Virtuoso layout editor [2]. The layout is then exported as GDS file, which can be imported by the layout drawing tool of the employed EM simulator.

As an example showing the significance of EM simulations, we consider a lumped-element band-pass filter as shown in Figure 15.1. Figure 15.2 shows the insertion loss (S_{21}) and return loss (S_{11}) simulated using the circuit simulator of ADS and the EM simulator IE3D for four cases:

Case 1: Simulation is based on the circuit simulator with ideal inductors and capacitors.

Case 2: Simulation is done using the EM simulator with non-ideal or actual inductors and capacitors. The *S*-parameters of each individual filter element are first calculated based on its physical structure. These *S*-parameters are then combined in the circuit simulator to calculate the *S*-parameters of the band-pass filter.

Case 3: All non-ideal or actual inductors and capacitors are integrated, and the *S*-parameters of the entire bans-pass filter are calculated using the EM simulator considering the mutual interactions between the elements. In this case, the filter elements are kept the same as in Case 2. The input and output RF pads are included in the simulation; however, the ground area is excluded.

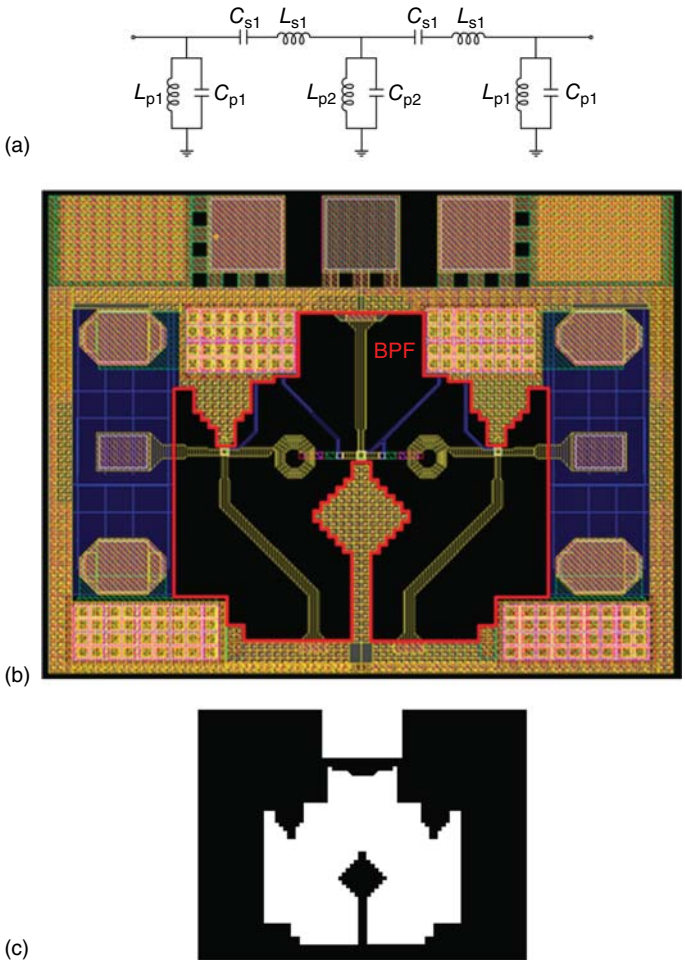


Figure 15.1. Schematic (a), layout (b), and (black) ground area (c) of the band-pass filter (BPF). The BPF is inside the box. Outside the box is the RF pads and ground area.

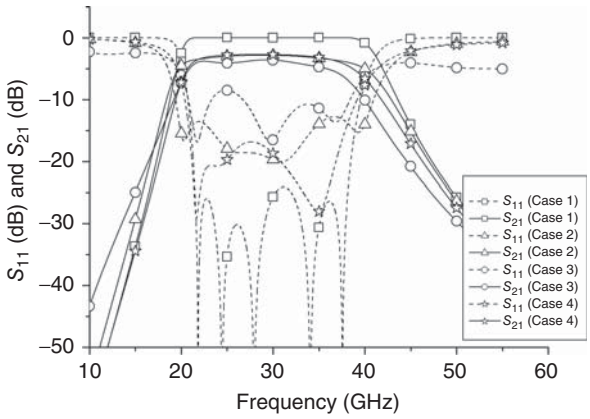


Figure 15.2. Simulated performance of the band-pass filter.

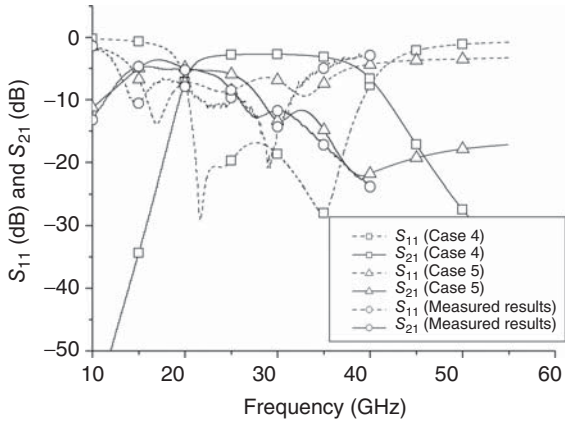


Figure 15.3. Measured and simulated performance of the band-pas filter. Simulation results are for Case 4 (no ground area included) and Case 5 (ground area included). Measurement results are available only up to 40 GHz.

Case 4: All nonideal or actual inductors and capacitors are integrated and optimized together, and the *S*-parameters of the entire bans-pass filter are calculated using the EM simulator taking into account the mutual interactions between the elements. The input and output RF pads are also included in the simulation, but the ground area is excluded.

The results in Figure 15.2 demonstrate the need of using EM simulators for RFIC design. Particularly, as the difference between Case 2 and Case 3 shows, the results signify the importance of simulating the entire circuit with all elements integrated as a whole so that the interactions between elements are taken into account. Finally, from the comparison between Case 3 and Case 4, it is recognized that a final optimization of a circuit should be carried out considering all elements integrated. Although Case-4 simulation takes into account the interactions between the actual inductors and capacitors, it omits the effects of the ground area [Figure 15.1(c)] seen in the circuit layout [Figure 15.1(b)]. This omission causes problems in the circuit performance. Figure 15.3 shows that the simulated results for Case 5 that considers the complete circuit layout – everything including circuit elements and ground – match closer to the measured results as compared to the simulation for Case 4 which neglects the ground area, demonstrating the need for such simulation in RFIC design.

As a general rule for the design of RFICs, EM simulations should be conducted for all passive circuit elements, such as inductors, capacitors, vias, transmission lines, interconnects, metal grounds, RF pads, and passive components such as baluns or transformers, filters, and power dividers, and for the entire circuit layout taking into account all elements appearing in the layout. EM simulators should also be used for active devices,

if available, in addition to using circuit simulators. At the very least for active devices, EM simulations need to be made for metal interconnections used at the terminals of a core active device from a PDK, as the metal interconnects effectively modify the device's characteristics such as increasing the loss and hence reducing and increasing the device's gain and noise figure, respectively. It is noted, however, that using EM simulators for all circuit elements may not be a good practice, especially in the low RF range, as EM simulations are always very time-consuming. Care and judgment, therefore, need to be exercised carefully in using proper EM and/or circuit simulators in order to arrive at accurate simulation results within tolerable specifications without incurring too much computational time.

15.1.10 Statistical and Mismatch Simulation

It is a known fact that measured performance of fabricated circuits always has some discrepancy from calculated results. One of the reasons for discrepancy is due to variations in fabrication processes causing performance changes in circuits and transistors fabricated on different wafers and at different times or even on the same wafer. These variations are inevitable and may vary from process to process. Even on the same wafer, two identical transistors located at different positions can have different performance due to factors such as doping variation. Therefore, it is important that process variations are taken into account in the circuit design process. To facilitate the evaluation of variation effects, most PDKs also include statistical models and mismatch models, which represent die-to-die variation and device-to-device variation (within the same die), respectively. These models should be used on the finally designed circuits to assess possible variations of their performance after fabricated.

15.2 RFIC LAYOUT

The layout of RFIC follows the design rules specified in the process of a fabrication foundry. These rules can be classified as general rules and specific rules. Each particular process has its own specific rules and, although specific rules might vary from process to process, general layout rules are approximately similar for different processes. These general rules govern essential requirements for parameters such as spacing, width, orientation, enclosure, and extension of metals. In general, both general and specific rules for a process need to be satisfied for a circuit to be allowed for fabrication by a foundry. In addition to these minimum set of rules, there are various layout techniques that can be used to improve the performance of RFIC. This section discusses the general layout issues of fabrication and some performance-enhancement layout techniques for active and passive RFICs.

15.2.1 General Layout Issues

The PDK of any fabrication technology governs issues related to metal–metal interaction, metal shapes, active region, and so on. The verification of appropriate layout techniques is usually accomplished by a process known as design rule check (DRC). This process involves subjecting a circuit layout to a series of checks through a layout verification tool in software platforms such as Cadence. Several of these issues are mentioned below.

Minimal and Maximum Metal Widths: The width of any metal strip including interconnect between elements (or components) in a circuit or between circuits cannot fall below a certain number, which is classified as the minimum width of metal strips in a metal layer. Usually this number varies from metal layer to metal layer, but the minimum width of a metal strip in a process is normally proportional to the thickness of the metal layer. The top metal, which is the thickest, has a larger minimum width than the rest of the metals. Using a thick metal can therefore result in a larger minimal width, which can lead to desirable effects such as larger current handling and lower conductor loss. Besides the minimum width, a maximum width is also imposed for metal strips. Hence, wide metal strips cannot be used in the design unless slots are created within the width of the metal strip to keep each width-segment below the maximum allowable metal width.

Minimal Metal Spacing: The spacing between two adjacent metal strips cannot be smaller than a certain minimal value set by the fabrication foundry. The main reason is that there may be a metal extension of a few tenths of micron in the lithographic fabrication process that may result in unintentional connection of different metal strips. Similar to the minimum metal strip width, the minimum spacing is also typically proportional to the thickness of the metal layer.

Minimal Enclosure: Each component, such as resistor, inductor, or transistor, is surrounded by some empty space and not allowed to be enclosed by another metal layer within that space. This is to ensure that, in case the fabricated components exceed the tolerance level, the component connectivity with other metals would not suffer.

Metal Dummy: Metal dummies are typically placed in all metal layers as needed to allow the density test to pass. For instance, in six metal-layer CMOS processes, when a circuit only uses the top-most metal 6 and the bottom-most metal 1, dummies need to be place in metals 2–5 of the circuit to facilitate the density check.

15.2.2 Passive and Active Component Layout

15.2.2.1 Resistors. In RFICs, resistor layout is usually governed by considerations of linearity, capacitive coupling to the silicon substrate, and matching (or similarity) between the same resistors. Polysilicide resistors prove better in layout as they have lower capacitive coupling and allow greater resistance to be realized from smaller dimensions. For RF applications, care must be taken to ensure that the resistance retains its value at high frequencies. This is accomplished by running EM simulations on the resistor and estimating its *S*-parameters.[5] Resistance tolerance values range within ±20% in most fabrication processes, and RFIC designers must take into account the tolerance levels during the design and simulation phases.

15.2.2.2 Capacitors. Just like for resistors, matching between the same capacitors is a prime concern in capacitor layout. Matching issue is further discussed in a subsequent section. Poly silicon capacitors are the most commonly used capacitors as they allow realization of higher linearity capacitors with better matching. To enhance matching and symmetry, the capacitors are usually laid out in a set of symmetrical blocks rather than a single large block as shown in Figure 15.4. It is generally preferred to have smaller capacitors (a few picofarad at most) for on-chip implementations to reduce their physical size.

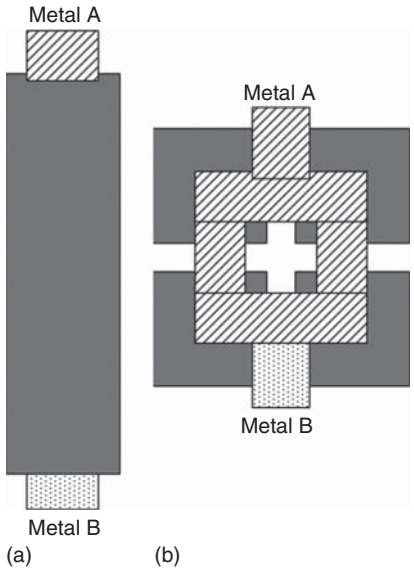


Figure 15.4. Layouts of capacitors with potential mismatch (a) and enhanced matching (b).

[5]This process should be done for all circuit elements.

15.2.2.3 Inductors. Inductors are typically considered the most concerned element among all passive elements for RFIC design and layout. This is due to several reasons. One is the large physical size of typical inductors that take up much space in RFICs. Another is the difficulty in obtaining high quality factor (Q) for inductors, especially for large inductances and at high frequencies. The performance of many RFICs, such as amplifier's noise figure, VCO's phase noise, and filter's loss, depends much on the Q of inductors. Small size and high Q are hence critical for inductors in RFIC design. Layout of inductors plays a crucial role in achieving miniaturization and high Q and thus should be treated as an important part of the inductor design as well. Spiral layout as seen in Chapter 3 is the most commonly used layout configuration for on-chip inductors.

The first issue affecting the layout of inductors is the impact of metal interconnects that connect inductors to other circuit elements. In most cases, the layout, shape, and length of interconnects add additional losses, change inductance, and deteriorates performance of inductors. Inductors should therefore be laid out in such a manner that the lengths of interconnects are minimal. Also, interconnects must be included as a part of the inductors during the design and simulation, particularly when long interconnects are used and/or at high frequencies. Another important issue in the inductor layout is minimizing the fringe capacitance from metal traces to ground so that substantial substrate coupling is avoided. This is usually taken care of by using the top-most metal to increase the distance between inductors and silicon substrate and/or a shield beneath inductors. Using thick metals or (at best) ultra-thick metal in a process also has the additional benefit of lower series metal resistance, which implies lower loss. Care must also be taken not to place the inductor turns too close to one another so that the design rules are not violated. In addition, when an inductor is used as a load or on a current-flowing branch, both the inductor's metal-trace width and the number of vias needed for connecting different metals should be large enough to accommodate the flow of the required DC and AC currents, based on the current limits for metals and vias set forth in the process PDK, with a reasonable margin. Designers should therefore check the current densities of metal layers and vias in the PDK and use sufficiently large trace width (typically about 3–5 times larger than the limit) and as many vias as possible to meet or exceed the expected DC and AC current levels. If this constraint is missed or ignored for inductors, these inductors often burn out under (test) conditions of high temperature and/or high supply voltage.

15.2.2.4 Transistors. Both NMOS and PMOS transistors should follow the rule of inter-digitation for gates or use multiple "fingers" for gate in order to reduce the overall gate resistance due to the fact that the overall resistance is a parallel of smaller gate resistances. A Cadence layout of a typical NMOS transistor with multiple gate fingers is illustrated in Figure 15.5. Gate resistance of any MOS transistor is an important contributor to the noise performance of CMOS RFIC such as the noise figure of amplifiers and the phase noise of VCOs. Inter-digitation is very useful for RFIC. However, it comes with a potential problem of increasing the source-drain perimeter capacitance. This is reflected in the parasitic source-bulk and drain-bulk capacitances that need to be carefully adjusted during simulations. The layouts of all NMOS and PMOS transistors are provided by the PDK of a fabrication foundry. Therefore, the user has only enough control to ensure lower parasitic gate resistance with multi-finger layout. Matching between transistors is also improved using this technique.

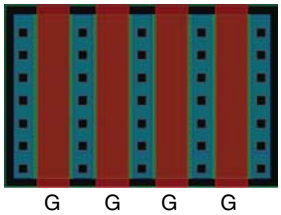


Figure 15.5. A multi-gate NMOS transistor layout. G stands for gate.

15.2.2.5 RFIC Layout Issues

Mismatch and Symmetry

Many RFICs function based on using identical or matched components – for instance, matched transistors in the LO switching pairs of the double-balanced Gilbert mixers discussed in Chapter 13 or matched resonators in band-pass filters. Symmetry is also often required in RFICs – for instance, in differential circuits – where perfect or near-perfect symmetry (both electrically and physically) is desired. Layout plays a significant role in maintaining the physical symmetry, which effectively helps in the electrical symmetry by minimizing the potential difference in the electrical effects of the layouts of the (desired) symmetrical parts. It is noted that "matching" here indicates the similarity between components, not the typical impedance matching in circuit design. Component matching refers not only to the matching between individual elements in a circuit, but also to the matching between subcircuits of a circuit – for example, between the two oscillator halves in a push-push oscillator. For possible optimum performance, RFICs, such as low-frequency analog ICs, require very good (ideally perfect) matching between the same components. Matching is particularly affected by the symmetry between different components or circuits. Among circuit elements, the matching between transistors is most difficult to achieve, due to the fact that the layout of the core of transistors is mostly out of the control of RFIC designers, and affects circuit performance the most. Many circuit design simulators implicitly assume identical (or at least very good matching between) transistors. This is, of course, not true in practice, hence causing circuit performance degradation if layout is not carefully done to ensure matching and symmetry.

As an example of matching and symmetry, we consider, for instance, a simple differential pair having differential input and output as shown in Figure 15.6. The differential pair operates based on two identical halves (identical resistors R and transistors M1, M2) and (theoretically assumed perfect) symmetry[6] at the common plane between the two halves. In typical initial design and simulation using circuit simulators, it is implicitly assumed that transistors M1 and M2 are the same. The differential pair thus produces no output voltage ($\Delta V_{\text{out}} = V_{\text{out}}^{+} - V_{\text{out}}^{-} = 0$) when the input signal is zero ($\Delta V_{\text{in}} = V_{\text{in}}^{+} - V_{\text{in}}^{-} = 0$). However, when M1 and M2 are not matched (e.g., due to improper layout), the two halves are not identical and hence not perfectly symmetrical. As a result, there is a non-zero differential output even when $\Delta V_{\text{in}} = 0$. This phenomenon

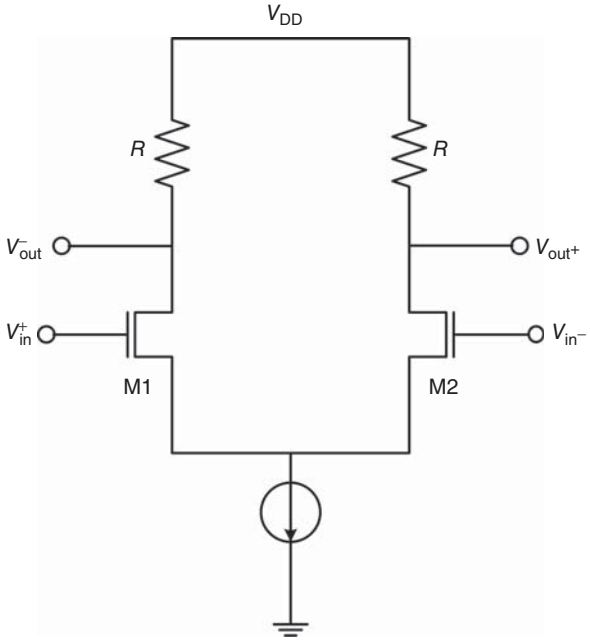


Figure 15.6. A simple differential pair.

[6]It is implicitly assumed that, in order to have perfect symmetry, the two halves must be identical.

is known as "dc offset" as mentioned in Chapter 11. This mismatch in differential pairs is typically modeled as a dc-offset voltage at the input. It also leads to a non-zero even-order harmonic distortion, which deteriorates the linearity of RFICs that employs differential pairs such as the Gilbert mixer cell. The impact of mismatch can be significant in differential-pair based RFICs. As such, mismatch-cancellation circuits may need to be employed to lessen the difference between the transistors in particular or between the two halves of a circuit in general. To improve the matching, both transistors are usually laid out facing each other, so that the gate of one transistor sees two sources on one side of it, rather than being parallel to each other. Another way to cancel or minimize the mismatch is adding dummy transistors on either side of a symmetric structure. This evens out the offset and results in greater symmetry. Dummy transistors, however, are not very suitable for circuits that employ cascoded or advanced differential pairs due to complications in routing.

Parasitic *R*, *L*, and *C*

Parasitic resistances, inductances, and capacitances in layouts may cause significant problems in RFICs, especially those in the high frequency end. These parasitics, unfortunately, are inevitable. Whenever a metal interconnect or passive element is laid out, it inherently couples to the lossy silicon substrate through a parasitic capacitance. The level of coupling depends on the size of the element and its distance to the silicon substrate. This coupling electrically forms a conductive path to the bulk for the RF signal, causing a reduction in the signal strength. This problem is particularly more pronounced in inductors as they occupy the largest die area of any passive elements, hence causing the largest coupling. Passive elements, including inductors, capacitors, vias, transmission lines, interconnects, metal grounds, RF pads, and so on, therefore need to be carefully laid out to minimize the parasitic resistances, inductances, and capacitances. For instance, minimizing the length of the interconnects between circuit elements (e.g., between inductors and others), reducing the size of circuit elements, using a large number of vias between different metals that form an RF pad to lower the contact resistance of the pad, employing multi-finger for the gate of any N/P-MOS transistor to reduce the noisy series gate resistance. Small-value elements such as those in the pH range for inductors (e.g., 30 pH) are very susceptive to their layout parasitics, especially at millimeter-wave frequencies, and the parasitics are thus doubly significant.

Grounding

Grounding in RFIC, particularly at millimeter-wave frequencies, is so critical that it may severely degrade the circuit performance if not properly considered. A ground plane should have width and length as wide and short as possible, respectively, to minimize its inductance and resistance, especially grounds associated with critical elements having small values in a circuit – for instance, a grounded degeneration inductor in LNA. Instead of using only the top-most metal layer for a ground, metal layers stacked together through vias should be used (e.g., stacked metals 1–6 in a 6-metal-layer process) to reduce the ground's inductance and resistance, make the ground more robust, and help meet the metal density rules for a process. The use of multiple vias for stacked metal grounds needs to be carefully considered. The grounds for on-wafer RF pads and other elements in the circuit should be connected with minimum phase difference between them (ideally zero phase). For large circuits such as complete receivers, transmitters, or systems, which require many DC bias sources, a bias distribution circuitry should be integrated on-chip with the circuit to provide DC signals for different sub-circuits from a single power supply to minimize the grounding problems. Using multiple external DC power supplies for large circuits causes severe grounding problems such as preventing proper setting of bias point for each sub-circuit, and hence the entire circuit, and causing fluctuations in measured performance versus time.

15.3 RFIC MEASUREMENT

RFICs can be measured directly on-wafer (or on-chip) or indirectly off-chip (or on-package) depending on requirements and/or circuit usage. For instance, in an integrated-circuit implementation or environment, such as a single-chip system, on-wafer testing is preferred for individual circuits as they will be integrated

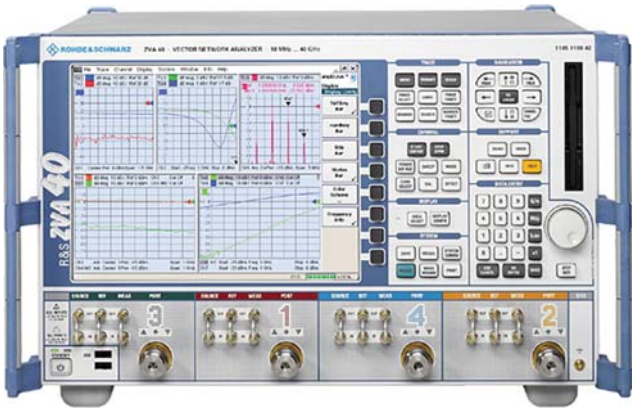


Figure 15.7. A vector network analyzer for RF circuit measurement. (Courtesy of Rohde & Schwarz, Inc.)

in the system. On the other hand, if a circuit is intended to be used as a standalone product, it will need to be tested off-chip via external connectors, bonding wires, package housing the chip, and PCB holding the package. On-wafer measurement provides the most accurate results and performance for circuits due to the elimination of bonding wires, pins, package, interconnects between pins and connectors, and PCB. On-wafer measurement is essential for RFICs, particularly in the high RF range and/or over a wide bandwidth. Its use is inevitable at millimeter-wave frequencies since off-chip testing may not be feasible due to large parasitics arisen from the off-chip testing structures. The equipment used for the on- and off-chip measurement is essentially the same. The main differences are the interface between the instruments and the circuit, and the calibration standards. On-chip measurement requires on-wafer probes and on-chip or on-board calibration standards, while off-chip measurement uses connectors and (off-chip) discrete calibration standards (just like regular testing of discrete RF components.) Perhaps the most widely used instrument for RF circuit measurement is vector network analyzers (VNAs). Figure 15.7 shows a VNA. References [4–6] provide useful information for measurement using VNAs.

15.3.1 On-Wafer Measurement

On-wafer measurement provides measured performance of an RFIC directly on its bare die and, with proper calibrations, it produces the most accurate results. On-wafer testing requires additional measurement accessories including on-wafer probe station, on-wafer RF and DC probes, and on-wafer calibration standards or calibration standards on a separate substrate (off-chip calibration standards). On-wafer measurement, just like off-chip or conventional circuit measurement, involves different kinds of measurements such as *S*-parameter, signal (waveform and spectrum), noise, power, and linearity measurements. On-wafer measurement requires on-wafer calibrations. Our discussion of calibration is based primarily on the *S*-parameter measurement using VNAs. However, the principle and use of the calibration apply to the measurement of other parameters such as spectrum measurement as well. The main difference between different measurements is the use of different instruments such as a spectrum analyzer for spectrum measurement. Figure 15.8 shows an on-wafer measurement setup being used for testing an RFIC.

15.3.1.1 Measurement Interface. On-wafer probes provide the interface between the chip and the equipment for measurement purposes. Typical probe stations can accommodate 2–4 on-wafer RF probes and 2–4 DC probes with multiple pins, making it convenient to measure multiport RFICs, such as mixers or T/R switches, with many bias points on-wafer. When a circuit contains more ports than those provided by a probe station, ports operating at lower frequencies can be bonded on to the pins of an open-cavity package mounted on a PCB and accessed via transmission lines and connectors. If more DC bias ports are needed, the package's leads can also be used to provide the DC bias signals without DC probes.

Figure 15.9 shows a probe station, two single-ended RF probes landing on a wafer in a measurement setup, and a differential RF probe manufactured by Cascade Microtech, Inc. [7]. There are mainly two kinds

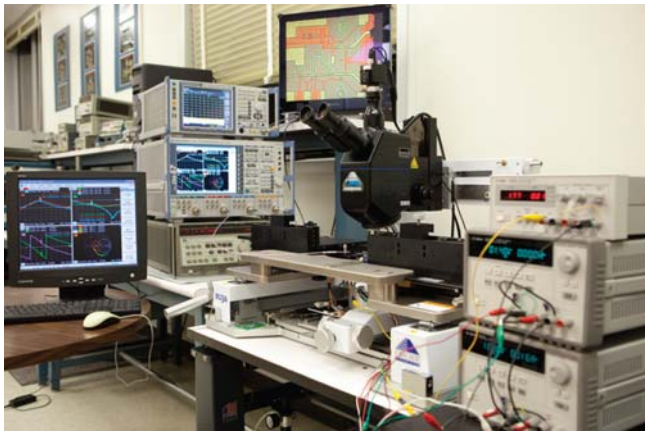


Figure 15.8. On-wafer measurement setup for an RFIC.

of on-wafer RF probes, which are based on coplanar waveguides (CPWs): one is single-ended probes using ground-signal-ground (GSG) CPW structure [Figure 15.9(b)] and another is differential probes implemented using ground-signal-ground-signal-ground (GSGSG) structure [Figure 15.9(c)]. The differential probes are useful for RFICS employing differential ports such as baluns or double-balanced mixers. RF probes are also made for 50 Ω and higher impedance. As 50-Ω is the standard impedance, 50-Ω probes are usually employed for measurement. High-impedance probes are used to sample waveforms at certain nodes inside a circuit while avoiding the (impedance) effect on the circuit being measured. Besides RF probe, on-wafer DC probes are also used to provide DC bias voltages or digital signals.

15.3.1.2 On-Wafer Pads. In order to perform on-wafer measurement, on-wafer pads where the probes are landed on must be included in RFICs. There are two kinds of on-wafer pads: RF pads and DC pads. These on-wafer pads allow the interface of a circuit with the measurement set up and/or connections with other external components. Due to different design constraints and requirements, as well as different interconnect-routing needs for each circuit, having a customized user defined on-wafer pad is better than relying on a PDK-based on-wafer pad.

RF pads can be implemented in (single-ended) GSG or (differential) GSGSG structure, depending on whether single-ended or differential RF probes are used, respectively. Many circuits are single-ended and, hence, GSG RF pads are mostly employed. The pitch[7] of RF pads is typically determined by the pitch of the RF probes used, for instance, 150 μm. The size of RF pads should be small to minimize the parasitic capacitance to the substrate or ground shield, yet sufficiently large to accommodate the probe's foot print. Typical RF pads have size between 50 μm × 50 μm and 100 μm × 100 μm – for instance, several RFICs presented in this book use RF pads having size of 100 μm × 100 μm, gap of 50 μm, and pitch of 150 μm. To reduce the parasitic capacitance, only the top-most metal layer should be used to construct RF pads. However, in order to increase the metal bonding strength, an on-wafer RF pad is typically formed by a stack of metal layers consisting of the top-most and lower metal layers, typically three layers, interconnected through via-holes. The RF pad resistance is lowered by using dozens of vias between these different metal layers which result in reduced contact resistance between the metals. For all RF signal pads, having an ESD protection to prevent accidental penetration of high voltages is desirable. This is accomplished by ESD protection circuits that are shunted to the bonding pad, which effectively absorb and ground these accidentally high voltages.

The bottom-most metal layer (M1), besides used as a metal ground, is also typically used as a metal shield for the signal pad of the RF pads to prevent or minimize RF signal leakage into the lossy Si substrate, just like it is often used as a shield for the same purpose in other circuit elements such as transmission lines. The ground shield helps isolate the signal pad from the substrate, thereby minimizing the coupling to the substrate

[7]Pitch is defined as the center-to-center distance between the signal pad and one of the ground pads of an RF pad.

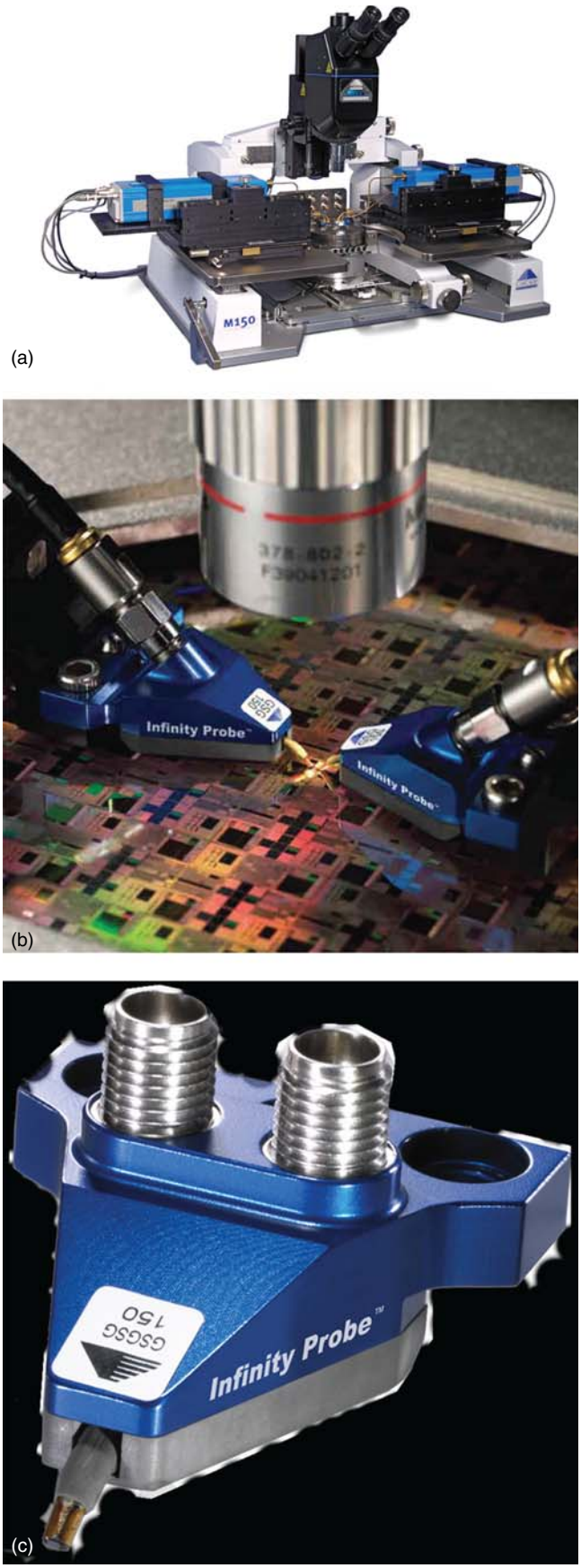


Figure 15.9. Probe station (a), on-wafer probe setup with single-ended probes (b), and a differential probe (c). (Courtesy of Cascade Microtech, Inc.)

from the signal pad as well as the coupling (via substrate) between different signal pads and hence between circuit ports. The ground shield should be larger than the total area of the RF pad that consists of two ground pads and signal pad to lessen the fringing fields between the edges of the signal pad and the substrate which causes the RF signal leakage. Signal leakage from an RF pad to substrate not only results in increased signal loss and noise (due to substrate), which affects the noise characterization for circuits connected between RF pads, and greater coupling between ports, but also makes the de-embedding model more complicated due to additional capacitances, resistances, and inductances in the substrate (from each pad to the bottom of the substrate and between input and output pads), causing error in the de-embedding equation (15.2) given in section "Example" and hence inaccurate de-embedding results. It is apparent that a larger shield leads to a better de-embedding result, but inadvertently with the expense of a larger die size. Use of a ground shield that extends about two times of the length and width of the RF pad from the pad edge along the pad's length and width, respectively, should be sufficient to contain the signal pad from the substrate. Nevertheless, an EM simulator should be used to approximately determine the size of the ground shield that produces maximum isolation between the designed RF pad and the substrate. It should be noted, however, that the use of a large ground shield increases the parasitic capacitance for the signal pad which are normally undesirable. A large ground shield increases the possibility of propagation of unwanted higher-order modes within the operating frequency range of interest, especially when the frequency is high.

On-wafer DC pads are used for DC probes to supply DC bias voltages to transistors or diodes or to feed digital signals to circuits. Typical size of DC pads is 100 μm × 100 μm, and DC pads also normally include multiple metal layers connected by many vias to enhance the strength and reduce the contact resistance.

Figure 15.10 shows an on-wafer RF pad, based on the GSG configuration, used in some of the RFICs presented in this book, its equivalent-circuit model, and the calculated *S*-parameters from the equivalent-circuit model and EM simulator IE3D [3].

To illustrate the significance of the size of a metal shield for on-wafer RF pads, we consider the RF pad shown in Figure 15.10 and repeat the de-embedding example for a microstrip-line device under test (DUT) having width and length of 10.2 and 300 μm, respectively, described in section "Example", for the following three different metal shields on the bottom-most metal layer (M1):

Shield 1 (Zero-Extension): The edges of the shield are aligned with the edges of the RF ground pads. The dimension of Shield 1 is 374 μm × 567 μm.

Shield 2 (L_G/W_G Extension): The edges of the shield are extended by $L_G = 113$ μm and $W_G = 74$ μm from the edges of the RF ground pads along the pads' length and width, respectively, where L_G and W_G are the length and width of the RF ground pad, respectively. The dimension of Shield 2 is 522 μm × 793 μm.

Shield 3 ($2L_G/2W_G$): The edges of the shield are extended by $2L_G = 226$ μm and $2W_G = 148$ μm from the edges of the RF pads along the pads' length and width, respectively. The dimension of Shield 3 is 670 μm × 1019 μm.

Figures 15.11 and 15.12 show the DUT with and without RF pads and the corresponding dummy for Shield 1 and 3, respectively.

Figure 15.13 shows the de-embedding results for three cases up to 60 GHz. As can be seen, the matching between the de-embedded *S*-parameters and actual *S*-parameters of the DUT is improved as the shield coverage for the RF pad is increased, as expected. Particularly, when the shield boundary is aligned with that of the RF ground pad as shown in Figure 15.11, the difference between the *S*-parameters is more pronounced due to more RF signal leakage from the pads to the substrate resulting from increased fringing fields from the pads' edges to the Si substrate. Further increase of the shield coverage will result in better agreement between the *S*-parameters, but at an expense of increased chip area. For most measurements, using a metal shield that extends over the RF ground pads by about twice of the length and width of the RF ground pad along their length and width, respectively, as for Shield 3, is sufficient. For device modeling purposes, however, using a larger shield is preferable for better modeling results.

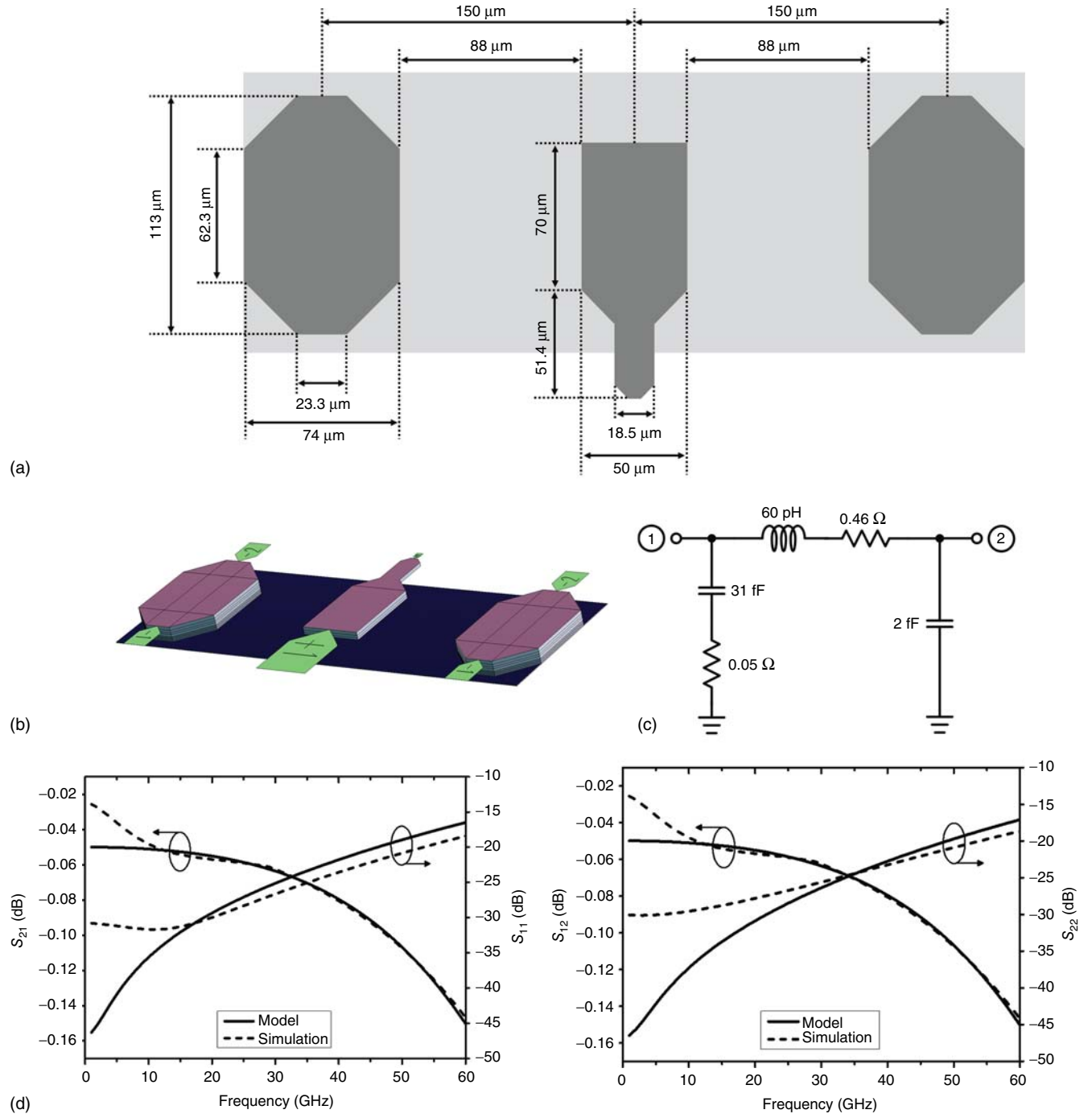


Figure 15.10. (a) Layout with dimensions (all angles are 45°), (b) layout used for EM simulations (1 and 2 are simulated ports), (c) equivalent circuit, and (d) calculated *S*-parameters (Model: equivalent-circuit simulation; Simulation: EM simulation). The signal pad consists of stacked metals 5 and 6 with metal 1 as a shield and the ground pad consists of stacked metals 1 to 6 in a 6-metal process.

15.3.1.3 Calibration Standards. As in the measurement of any RF components using instruments such as a VNA, accurate calibration needs to be performed using an accurately designed and fabricated calibration set prior to on-wafer measurement of RFICs. The calibration is used to remove all known and unknown effects of the measurement system including the instrument and connection accessories due to their imperfections, such as the finite losses of cables and internal reflections inside the VNA, at the measurement reference

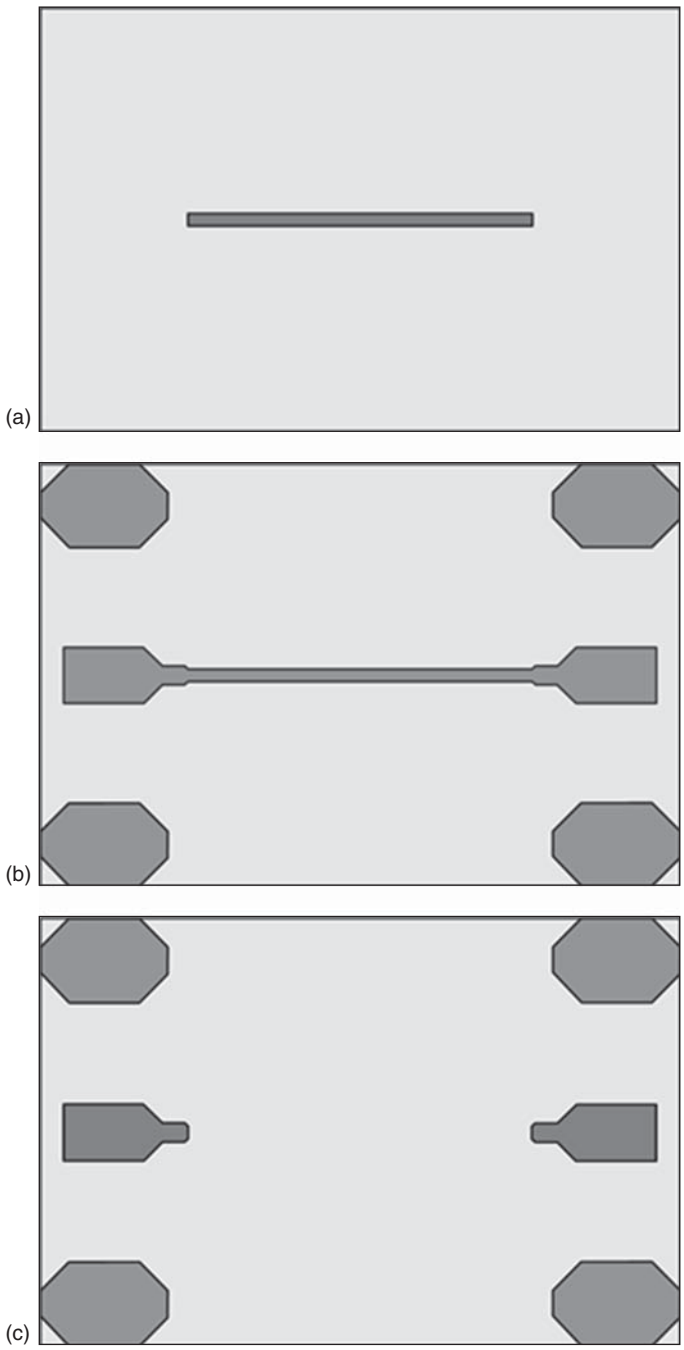


Figure 15.11. Microstrip-line DUT (a), DUT between RF pads (b), and dummy (c), for Shield 1.

planes before and after the DUT, leaving the DUT as the single component whose measured results are obtained. The measurement reference plane is typically defined at the tip of the on-wafer probe, which lands on the circuit.

There are in principle two kinds of calibration processes: one is the short–open–load–through (SOLT) calibration process based on the short, open, load, and through (thru) standards and another is the thru-reflect-line (TRL) calibration process based on the thru, reflect, line and match standards. The SOLT calibration is the most widely used method and is the standard calibration used in VNAs. All VNAs provide calibration kits containing SOLT coaxial standards suitable for off-chip measurements. In any calibration process using a VNA, the calibration standards of the calibration process (e.g., the short, open,

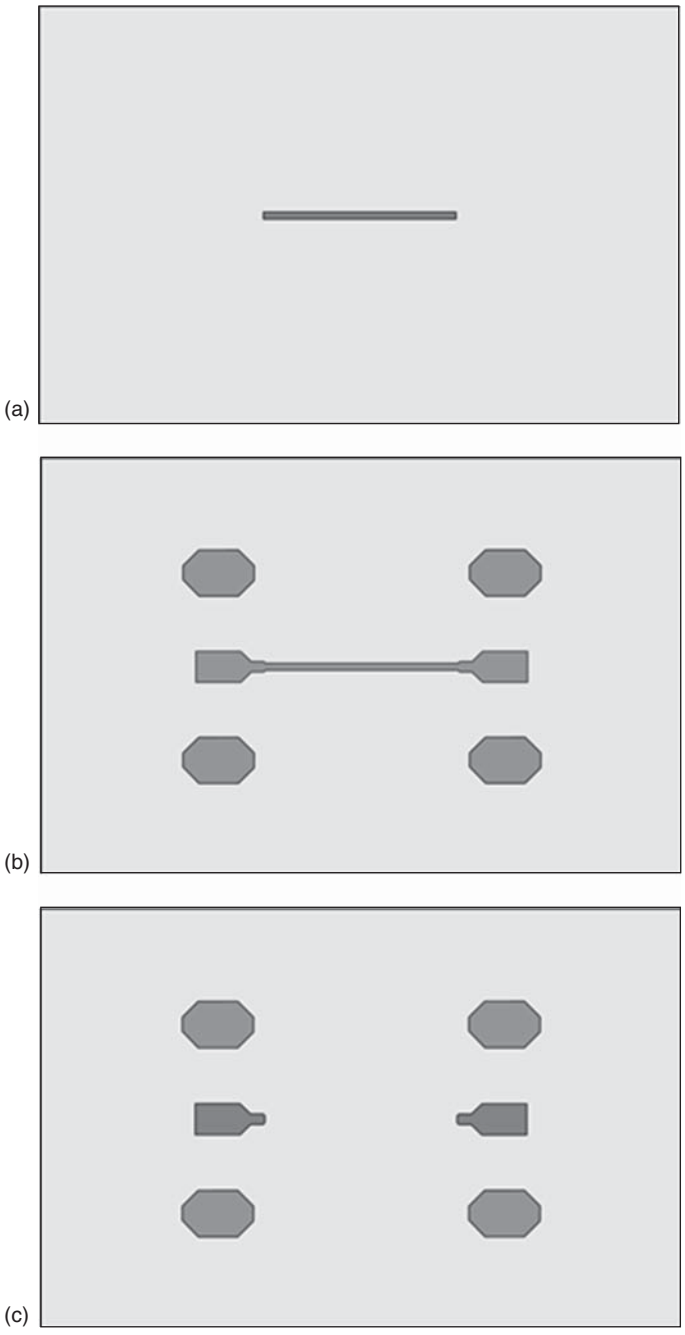


Figure 15.12. Microstrip-line DUT (a), DUT between RF pads (b), and dummy (c), for Shield 3.

load and thru standards of the SOLT calibration process) are measured by the VNA and the results are used by the VNA to remove the known and unknown effects of the measurement setup including the VNA up to the measurement reference planes. The process of removing these effects, which is basically an "error-correction" process, is typically done internally by the VNA based on an error model of the measurement setup that is characterized by the measured data of the calibration standards. The error model can be conveniently formulated using a signal flow graph as discussed in the Signal Flow Graph Appendix of Chapter 11. The TRL calibration technique only needs three calibration standards as compared to four for the SOLT calibration method, making it easier to fabricate and to characterize, hence more reliable. It is noted that characterizing an open circuit needed in the SOLT is prone to error, especially at high frequencies,

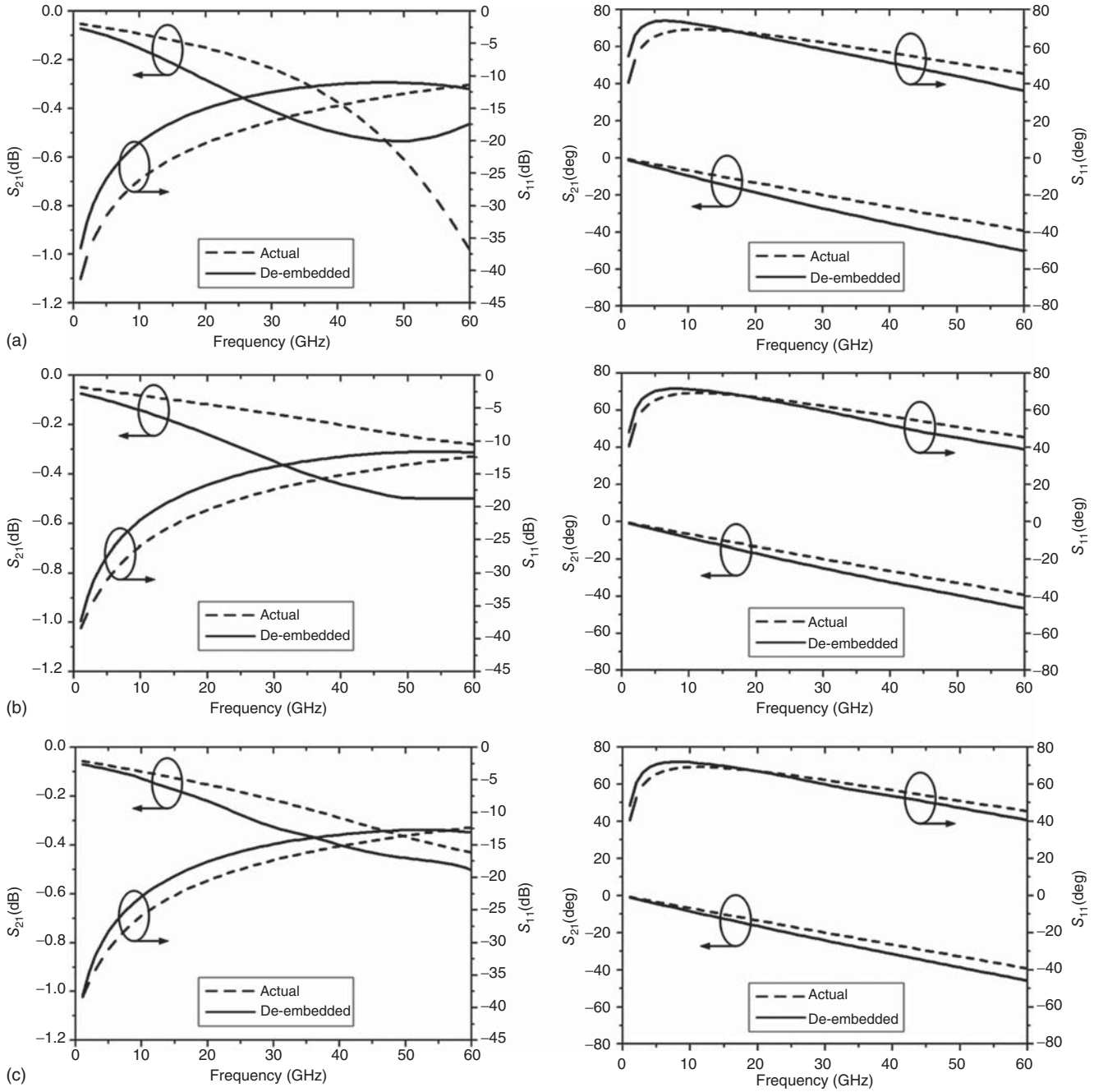


Figure 15.13. De-embedded and actual *S*-parameters of the DUT for (a) Shield 1, (b) Shield 2, and (c) Shield 3.

and fabricating a very accurate resistive load for the SOLT may pose some difficulties. Additionally, the TRL standards have fewer requirements on characterization. In general, the SOLT calibration is preferred over the TRL calibration for low RF frequencies (typically below 20 GHz or so). In this section, both these calibration standards are discussed. The accuracy of the measurement of a DUT depends on various factors, including the employed calibration technique (e.g., TRL or SOLT), the mathematical accuracy of the error model, the completeness of the error model's characterization based on the measured results of the calibration standards (e.g., sufficient calibration measurements for the error model's description), the accuracy of the design and fabrication of the calibration standards, the accuracy of the required input information (e.g., impedances), the repeatability of the measurement reference plane for different calibration standards

(e.g., the on-wafer probe's landing positions for the open, short, load, and thru of the SOLT standards), the physical contact between the probe tips and the on-wafer pads, and the repeatability of the measurement system (e.g., combined VNA, on-wafer probes, cables, and connectors).

SOLT Calibration Standards

One of the important requirements for accurate DUT measurements is the need of having the calibration standards in the same environment of the DUT. For RFIC, the short, open, load, and thru standards should therefore be fabricated on the same wafer/chip of the DUT with identical RF pads as for the DUT. Furthermore, these standards should be fabricated using the same kind of transmission line used at the input/output (feed-line) of the DUT (e.g., microstrip line or CPW). For on-wafer calibrations, the measurement reference plane can be defined on the on-wafer RF pad where the probe touches (e.g., at the center of the pad) or offset from the pad via a transmission line. Using a reference plane on the pad prohibits the use of a zero-length thru since a zero-length on-wafer thru is not physically or electrically possible. A short but sufficient length is needed between the RF pads not only for the convenience of two-port measurement due to the on-wafer probes, but also for avoiding the coupling between the probes and between the pads. Since a finite length is used for the through, a complete characterization of the thru including loss and delay is needed. Using a reference away from the pad on a transmission line enables both zero-length and non-zero length thru standards. For on-wafer standards, however, the use of such offset calibration standards take up more valuable space on expensive wafers.

Figures 15.14 and 15.15 show the on-wafer open, short, load, and thru calibration standards on CPW for 50-Ω systems for measurement reference plane on the RF pad and off from the RF pad, respectively. In practice, the open standard is typically simulated by lifting the RF probe into the air above the chip instead of using the on-wafer open standard. It is noted that different shape such as a diamond or oval shape can be used for the RF pads as desired and/or for minimizing the parasitics of the pad. Figure 15.16 shows photographs of some on-wafer calibration standards consisting of short, open, (50-Ω) load, and thru with measurement reference plane on the RF pad. General design of short, open, load, and thru standards is described in the following.

Short Standard. An on-wafer short standard can be easily realized by connecting the signal line of an on-wafer pad or the transmission line connecting to the pad to ground. For instance, an on-wafer microstrip short is typically obtained by connecting the signal line to ground using a via or multiple of vias as the interconnect. A via resembles electrically a combination of series resistance and inductance, and shunt resistance and capacitance, among which the series inductance being the most concerned one for calibration purposes since it affects the phase shift the most.[8] Therefore, a fabricated short at RF is never a perfect short of zero-impedance. The via's parasitic effects, although small for high-quality vias, cause changes in the characteristic of the short standard, particularly in the high RF range. To minimize these effects, multiple vias in parallel are typically used for the interconnect. In the CPW environment, an on-wafer CPW short can be realized by connecting the (central) signal line to both ground lines (for symmetrical CPW) or ground line (for asymmetrical CPW). Although the effect of the interconnect lines are less than that of the via used in the microstrip short, these interconnects also cause an additional inductance at the short, hence altering the short's characteristic. The interconnects between the signal and ground lines should be as short as possible. Since either the microstrip or CPW short is not exactly at the measurement reference plane of a DUT (due to the interconnect), a certain electrical delay (or electrical length) representing the delay (or electrical length) of the interconnect should be input to the VNA as part of the user-defined calibration process as instructed by the VNA.

Open Standard. An on-wafer open standard can be realized by leaving the signal pad or the transmission line connecting to the pad open (unterminated) or simply lifting the on-wafer probe into the air above the

[8]The series and shunt resistances cause (typically small) loss without phase delay, while the shunt capacitance is typically very small causing only a relatively small effect on phase shift.

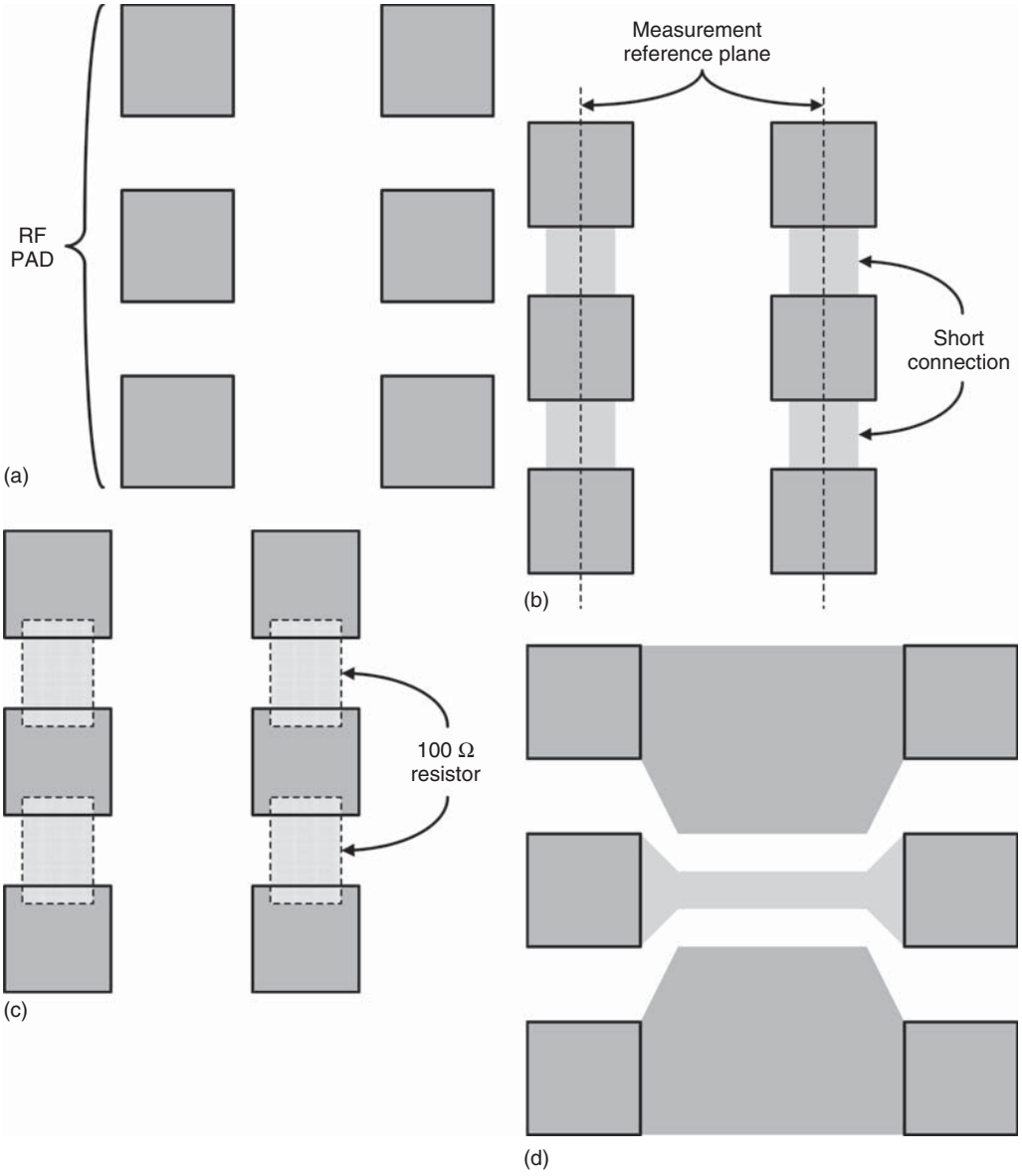


Figure 15.14. On-wafer 50-Ω SOLT calibration standards with measurement reference plane at the center of the RF pad: (a) open, (b) short, (c) load, and (d) thru. A finite length is used for the thru.

wafer. An on-wafer open-circuit, however, has fringing capacitance at the open end causing an electrical delay between the open and the measurement reference plane, which has an effect similar to an interconnect. The fringing capacitance depends on the distance between the open and silicon substrate, which may have a relatively small reactance, particularly at high frequencies, causing the open standard diverting from a perfect open as desired. Raising the probe in air above the wafer simulates better open due to less fringing capacitance and involves no contact between the probes and the on-wafer pads, thus alleviating the repeatability issue. Such an open standard is thus preferred. Since the open-end fringing capacitance causes the open departing from the measurement reference plane, it should be determined and entered into the VNA according to the VNA's instruction to take into account the phase delay (or equivalent electrical length) for accurate calibration.

Load Standard. An on-wafer 50-Ω load can be realized using on-chip resistors and is connected to ground directly or through a transmission line either by vias or interconnects. In practice, however, a fabricated load

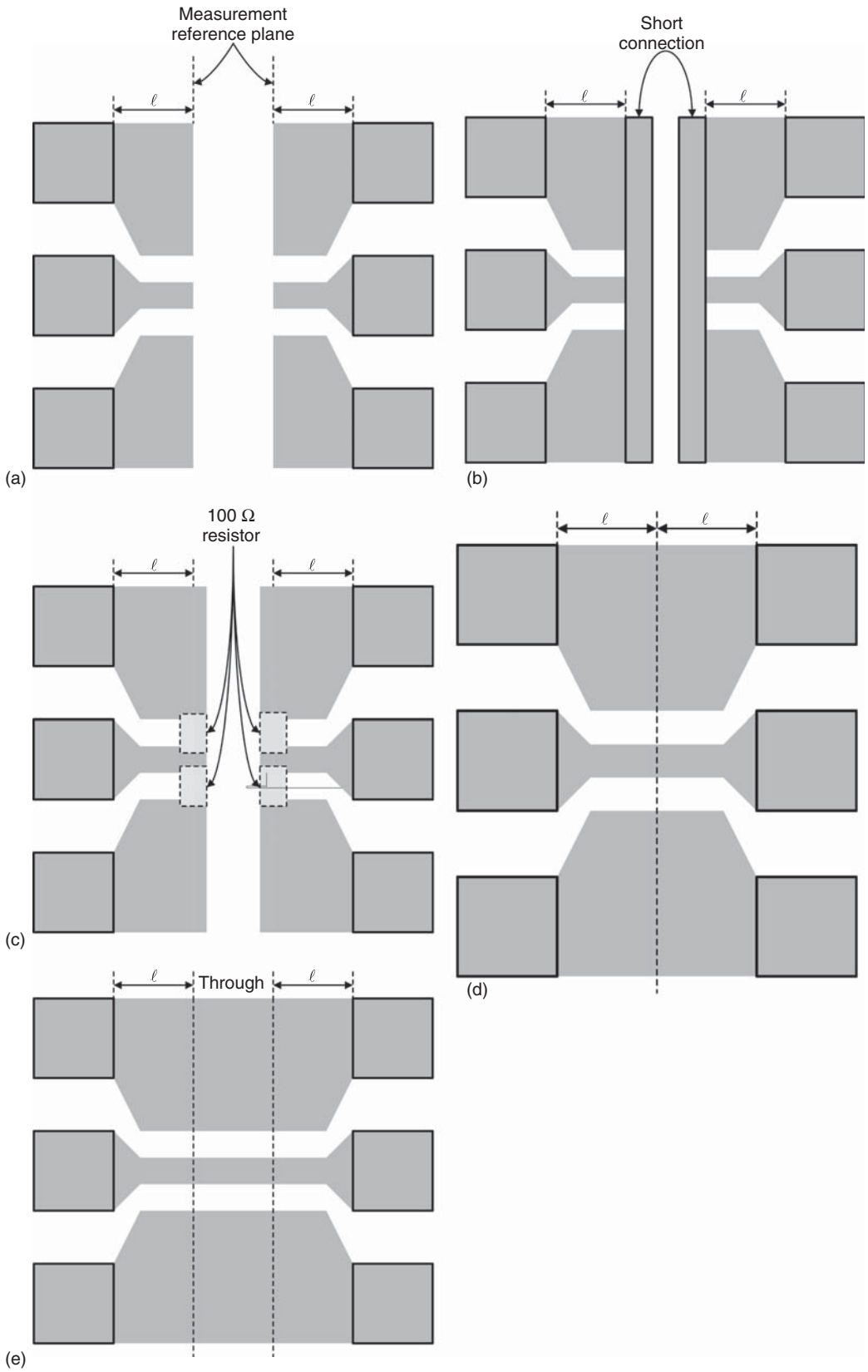


Figure 15.15. On-wafer 50-Ω SOLT calibration standards with measurement reference plane at the end of the 50-Ω transmission line: (a) open, (b) short, (c) load, and (d, e) thru. Both zero and finite length are used for the thru.

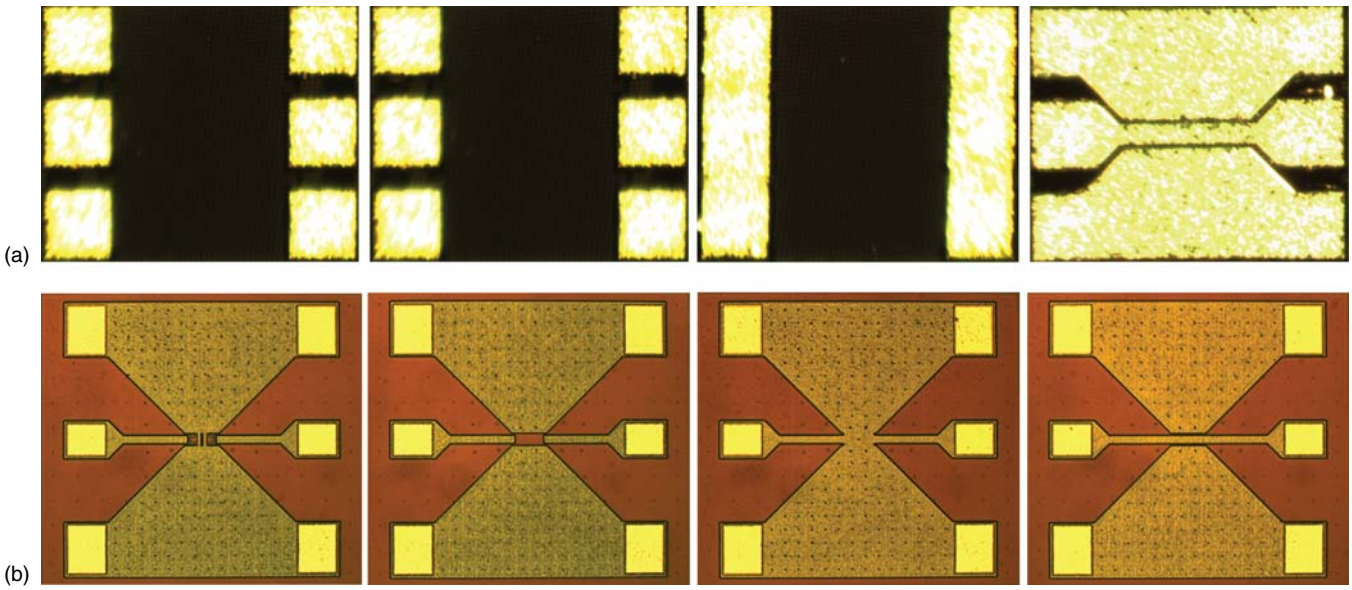


Figure 15.16. Photographs of on-wafer open, short, 50-Ω load, and thru CPW calibration standards for two cases: open, short, and load without (a) and with (b) the same environment for the thru. The layout details with dimensions of the standards in (b) are given in Figure 15.17. The 50-Ω load at each port is formed by two 100-Ω resistors in parallel. In (a), the 50-Ω load is not visible due to its burial underneath metal layers.

does not behave as a pure resistor; parasitic series inductance and shunt capacitance always exist along with the resistance, with series inductance being the most concerned parasitic since it affects the phase more than the shunt capacitance for typically fabricated resistors. An actual on-wafer 50-Ω load always causes reflection and alters the reference plane. To minimize the parasitics, hence improving the return loss, a 50-Ω load is typically formed using two parallel 100-Ω resistors connecting to ground. For instance, in CPW environment, two 100-Ω resistors, each between the signal line and one ground line, are normally used.

Thru Standard. An on-wafer thru is simply a short transmission line of any length including zero between the input and output on-wafer RF pads. The characteristic impedance of the thru must be equal to the characteristic impedance of the transmission lines used in the short, open, and load standards, which is normally 50 Ω. A thru should have input and output return loss as high as possible to minimize signal reflections at the input and output RF pads. A zero-length thru can be simply done by connecting two on-wafer RF pads with connecting transmission lines, whose ends define the reference planes, together. When a finite length is used for the thru, a complete characterization of this on-wafer thru standard including loss and time delay is needed for the VNA calibration, and these parameters can be measured or calculated using an EM simulator at the calibration frequency.

Consideration of Via Interconnect for Short and Load Standards. The total inductance of a lossless transmission line, whose length is very short as compared to a wavelength, is approximately equal to the product of the length (ℓ) and per-unit-length inductance (L) of the transmission line. The time delay of a lossless short transmission line is approximately given by the product of the transmission-line length and the time delay of a lossless transmission-line element, which is also the time delay of a synthetic transmission-line segment given by Eq. (4.233). We also learn from Chapter 8 that a short section of high-characteristic-impedance transmission line is approximately equivalent to a series inductor. These results allow us to approximately determine the time delay of a transmission line (interconnect) representing an inductance L_{ind} as

$$t_d = \ell\sqrt{LC} = \ell L\sqrt{\frac{C}{L}} \simeq \frac{L_{\text{ind}}}{Z_o} \tag{15.1}$$

where $Z_o = \sqrt{L/C}$ is the characteristic impedance of a lossless transmission line given in (4.60).

The parasitic inductance from the via-interconnect (to ground) for a short or 50-Ω load standard can be represented to the VNA by a short section of interconnect having high characteristic impedance to the maximum allowed by the VNA (e.g., 500 Ω) with the time delay calculated based on (15.1). It is noted that (15.1) is only approximate. It, however, gives very good accuracy when the length of the interconnect is kept small as compared to a wavelength and the characteristic impedance is high. This interconnect caused by the parasitic inductance alters the reference plane and the change in the reference plane can be considered offset delay. The characteristic impedance of the interconnect causing the offset delay (e.g., 500 Ω) is referred to as offset characteristic impedance. The loss of the interconnect is referred to as offset loss.

Proper calibration coefficients for the employed standards must be provided to the VNA according to its instruction in order for the VNA to perform accurate calibration along with measurements of the calibration standards.

SOLT Standard Design Example. Figure 15.17 show the on-wafer SOLT calibration standards designed on a 6-metal CMOS process. Figures 15.18–15.21 show the simulation results of the on-wafer calibration standards shown in Figure 15.17 using IE3D from 50 MHz to 70 GHz. The Smith charts are taken directly from the IE3D simulations without change of format. The results for S_{22} and S_{12} are identical to those for S_{11} and S_{21}. Table 15.1 shows the calculated input (output) impedance of the standards at 0.5 and 70 GHz. As can be seen in Figure 15.18, although the short is not perfect due to parasitic as expected, its characteristic shows a good short in the low frequency range. At high frequencies, its short quality degrades but is still acceptable as a short calibration standard, except in the upper region of the frequency range. Figure 15.19 shows that the open behaves almost as an ideal open up to 25 GHz with negligible real-part parasitic. Above 25 GHz, the open quality, however, degrades, yet it is still acceptable as an open calibration standard except at frequencies in the upper end. The results in Figure 15.20 demonstrate an almost constant 50-Ω load across 0.5–70 GHz. The 50-Ω load at each port is formed by two 92-Ω resistors in parallel to compensate for the high-impedance via-interconnect between the resistors and signal lines. The use of two resistors in parallel helps maintain the symmetry as well as reduce effects from possible variations in fabrication. Figure 15.21 shows that the thru standard has a very well match behavior with impedance close to 50 Ω and insertion loss less than 0.22 dB across 0.5–70 GHz.

TRL Calibration Standards

TRL calibration standards are based more on transmission lines (through and line) as compared to those of SOLT which are based more on impedance components (open, short, and load). Since transmission lines are simpler and more accurate to model and fabricate than open, short, and load, TRL standards provide an effective means for calibration. As for the SOLT calibration standards, for RFICs, the through (thru), reflection (reflect), and line standards should be fabricated on the same wafer/chip of the DUT using the same kind of transmission line used at the input and output (feed-line) of the DUT.

An on-wafer thru standard for the TRL calibration method is the same as a thru standard used in the SOTL calibration. An on-wafer reflect standard can be made by using a short or open. An on-wafer line standard is basically the same as a thru standard but with a different length. To cover broad calibration frequency ranges, multiple lines may need to be used. As for the on-wafer SOLT standards, it is not possible to use a zero-length thru for the on-wafer TRL standards. A short but sufficient length is needed for the thru. TRL is not suitable for calibrations at low frequencies as compared to SOTL since at these frequencies, the line standards become too long to be practical. Therefore, in order to cover a broad frequency range from very low frequencies (e.g., 10 MHz) to high frequencies (e.g., 40 GHz), both SOTL and TRL calibration methods may need to be used simultaneously.

As for the SOLT calibration standards, the measurement reference plane can be defined on the on-wafer RF pad (e.g., at the center of the pad) or offset from the pad via a transmission line. Figure 15.22 shows on-wafer TRL calibration standards. The design of these standards requires certain requirements as described in the following.

Thru Standard. The thru standard can be made with zero or non-zero length per the following requirements:

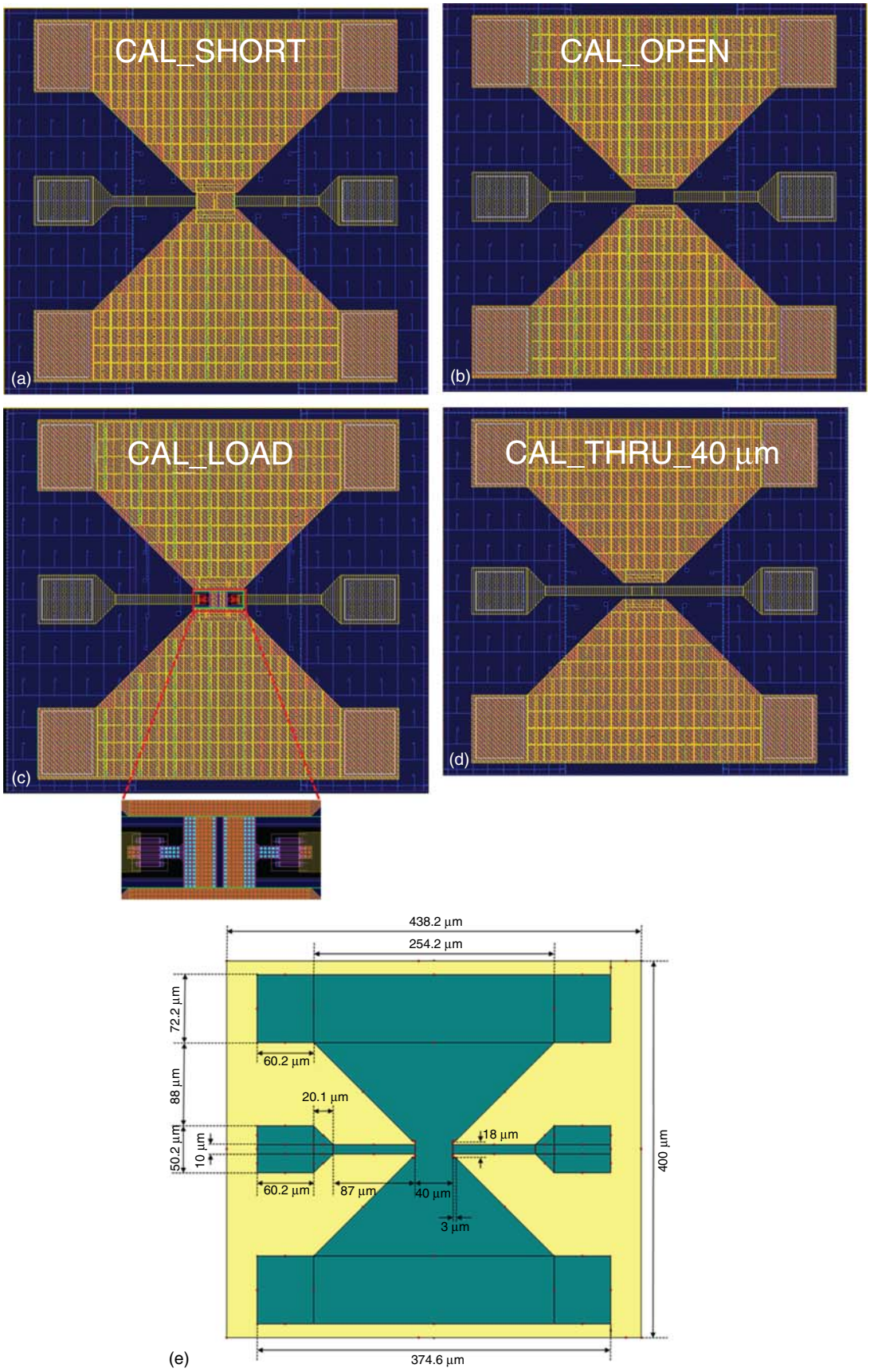


Figure 15.17. On-wafer short (a), open (b), 50-Ω load (c), and thru (d) calibration standards, and dimensions of the short standard (e). Dimensions for the other standards are the same as those for the short.

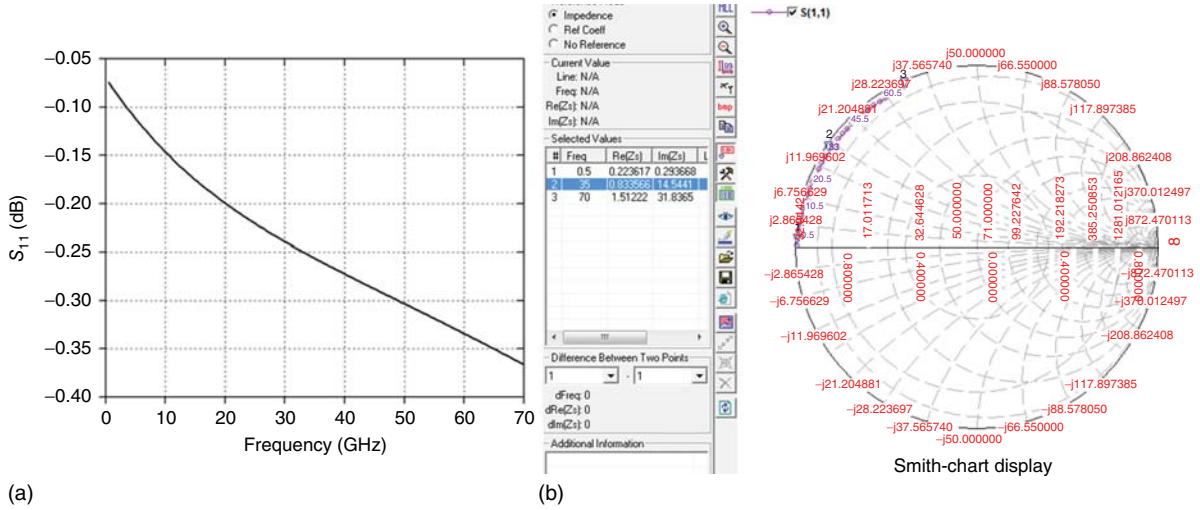


Figure 15.18. Simulated S_{11} in decibel (a) and on Smith chart (b) for the short.

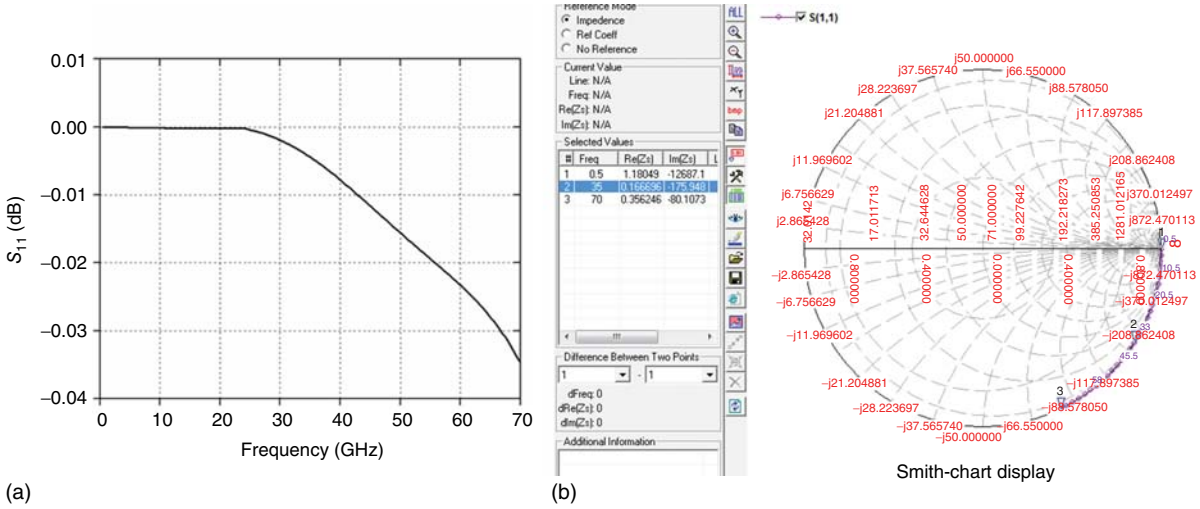


Figure 15.19. Simulated S_{11} in decibel (a) and on Smith chart (b) for the open.

Zero-Length Thru. The zero-length thru standard [Figure 15.22(a)] does not actually contain any transmission line, and so it does not have loss and its characteristic impedance is not required. It can be viewed as a perfect transmission line and hence has $S_{21} = S_{12} = 1 \angle 0°$ and $S_{11} = S_{22} = 0$.

Nonzero-Length Thru. Nonzero-length thru standard can be realized using the same or different characteristic impedance with the line standard but with a different length. Typically, the same characteristic impedance is used. However, when the thru and line have different characteristic impedances, the average of these characteristic impedance is used. The loss of the thru is not needed for the calibration. Zero-length thru [Figure 15.22(a)] is typically used to establish the measurement reference plane. The electrical length (or phase) or the time delay of the thru needs to be specified for the reference plane setting. For the zero-length thru, the electrical length and time delay are zero and the loss is zero. When a nonzero-length thru is used but its time delay is set to zero, the reference plane is set in the middle of the thru.

Reflect Standard. Reflect standard can be realized using a short [Figure 15.22(d)] or an open [Figure 15.22(c)] (or raising the probe in air). For on-wafer calibrations, an open standard realized by raising the on-wafer

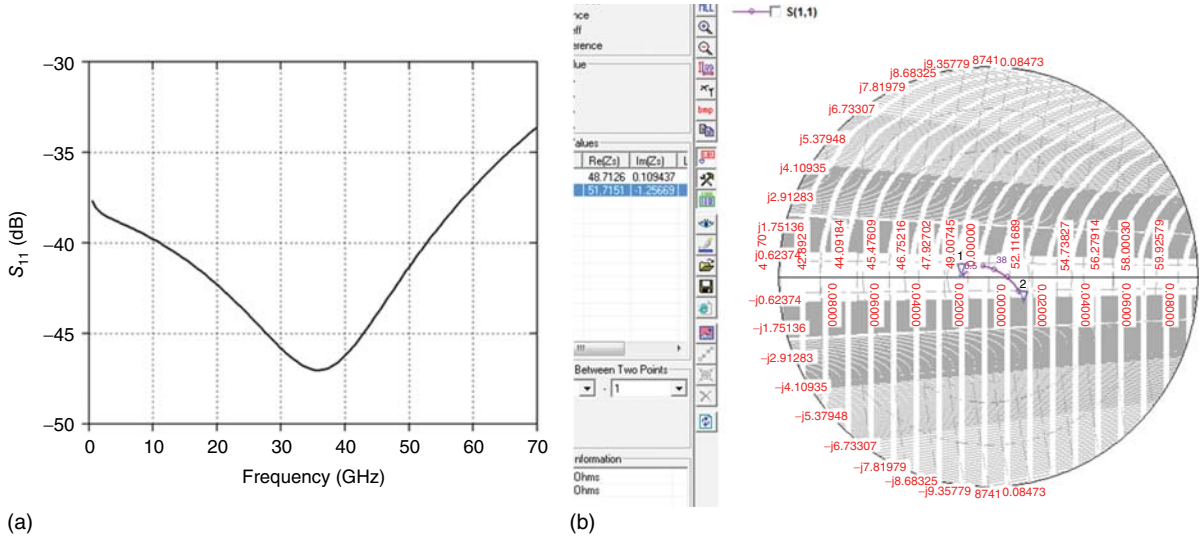


Figure 15.20. Simulated S_{11} in decibel (a) and on Smith chart (b) for the 50-Ω load.

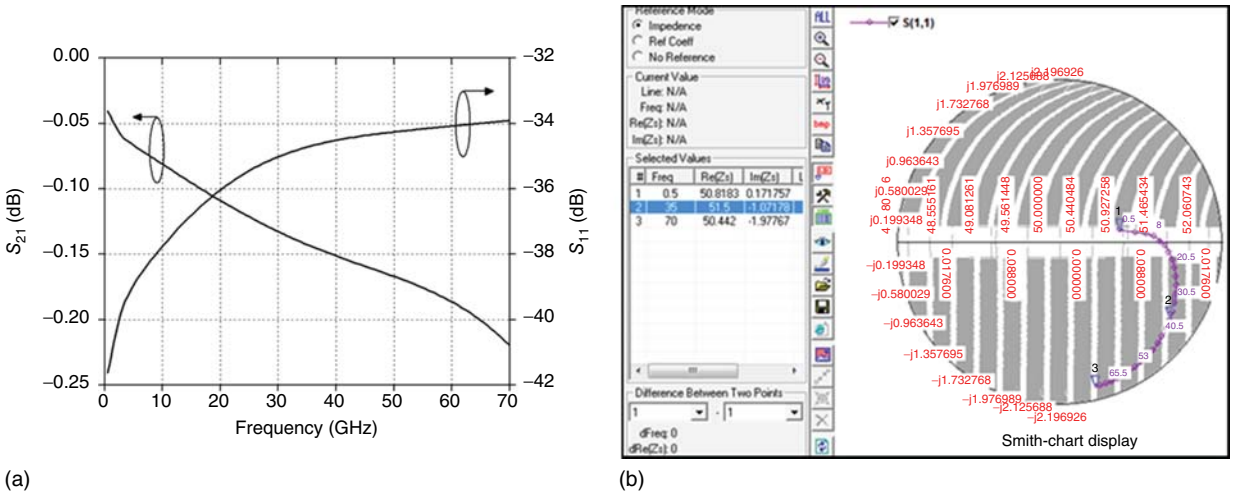


Figure 15.21. Simulated magnitudes of S_{11} and S_{21} (a), and S_{11} on Smith chart (b) for the thru.

TABLE 15.1. Input (Output) Impedance of the Standards

Frequency (GHz)	Short (Ω)	Open (Ω)	Load (Ω)	Thru (Ω)
0.5	$0.2 + j0.3$	$1.2 - j12687$	$48.7 + j0.1$	$51 + j0.2$
35	$0.8 + j14.5$	$0.2 - j176$		$51.5 - j1.1$
70	$1.5 + j31.8$	$0.4 - j80.1$	$51.7 - j1.3$	$50 - j2$

probes in air above the wafer avoids the contact between the probes and the on-wafer pads, thus alleviating the repeatability issue, and simulates better open due to less fringing capacitance. This open standard is thus typically preferred. Optimum reflect has reflection-coefficient magnitude equal to 1; however, the magnitude does not need to be specified. On the other hand, the phase of the reflection coefficient needs to be specified to within ±90°. Moreover, the reflection coefficient must be identical on both ports. Reflect can be used to set the measurement reference plane but its phase response needs to be accurately specified.

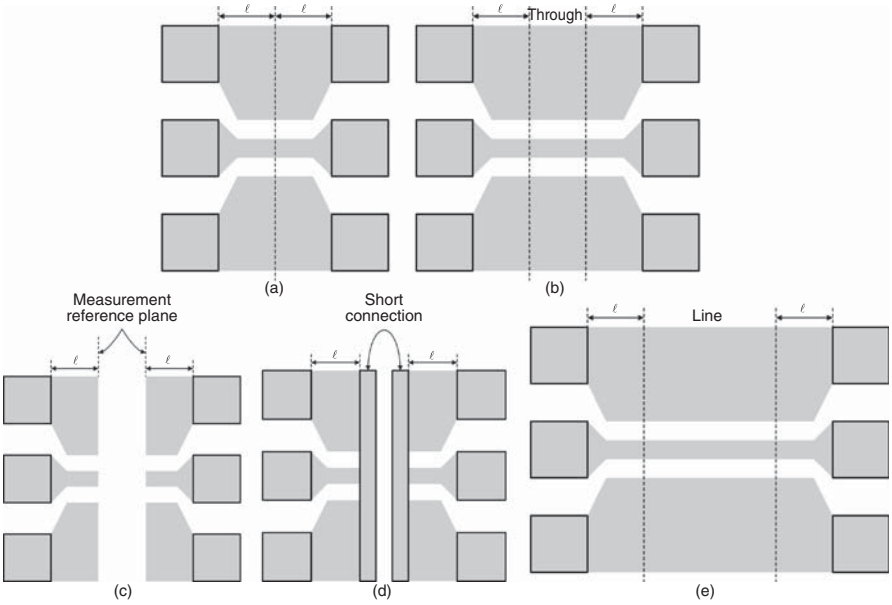


Figure 15.22. On-wafer thru's (a,b), open (c), short (d), and line (e) TRL calibration standards.

Line Standard. Line standard is realized using a transmission line [Figure 15.22(e)] whose characteristic impedance can be the same or different from that of the thru standard. Normally, the same characteristic impedance is used. The line characteristic impedance sets the reference impedance for the calibration and hence measurement, typically 50 Ω. The length of the line, and hence its phase, must be different from that of the zero-length or nonzero-length thru. The electrical lengths of the line and thru standards must be different by an amount between (20° and 160°) $\pm n \times 180°$, where n is an integer, over the frequency range of interest. The difference in the electrical lengths between the line and thru should not be chosen near 0° or an integer multiple of 180° since measurement uncertainty will increase significantly at these values. It is noted that the electrical lengths of 0° and (integer) multiples of 180° give the same response for transmission lines. The optimum length of a line is a quarter-wavelength at the center frequency with consideration of the foregoing mentioned constraint on the electrical length. The usable bandwidth for a pair of thru and line (one line and one thru) is around the frequency span 8:1. When broader bandwidths are required, multiple pairs of line and thru with identical characteristic impedances and different lengths are used, provided that the transmission lines are practically realizable. If the frequency span is less than 64:1, two line/thru pairs should be sufficient. When the electrical length of the line with respect to the thru is out of the constraint 20–160° for a broad frequency range, the frequency range can be divided into multiple ranges and one quarter-wavelength line at appropriate frequency is used for each range. The frequencies for each range can be determined so that the constraint is satisfied. When the range is divided into two, the transition frequency between the two ranges can be chosen as the geometric mean frequency $f_o = \sqrt{f_L f_H}$ where f_L and f_H are the lower and upper frequency of the range, and two quarter-wavelength lines are used: one for the f_L-to-f_o sub-range and another for the f_o-to-f_H sub-range. The loss of the line is not required to be known. However, the phase of the line needs to be specified within ±90°. One particular note is that, the length of a line, being optimally of a quarter-wavelength, becomes relatively too long to be practical at low frequencies, particularly on-wafer. Moreover, the line standard is not very suitable when the frequency range of interest is very wide. Under low frequencies and/or very wide bandwidth, it is more practical to use a "match" instead of a line and correspondingly the TRM calibration, based on thru, reflect, and match standards, is used instead of the TRL calibration. The TRM and TRL are essentially based on the same principle since a match is electrically equivalent to a matched line (i.e., a line terminated by its characteristic impedance) or a very long (ideally infinity) line. Therefore, a match can be used to substitute for a line at low frequencies or over very wide frequency ranges. The TRM implements the same thru and reflect standards as the TRL as expected. The TRM may be more convenient and less expensive than the TRL for on-wafer calibration,

especially if a load can be fabricated with very good accuracy across the interested frequencies, primarily since the TRM standards require less space.

Off-Chip Calibration Standards

Ideally, a calibration for on-wafer measurement should be done on-chip using on-wafer calibration standards as such on-wafer calibration, if performed properly, and accompanying standards, if designed and characterized accurately, would give the most accurate measurement results for RFICs. This, however, may impose inconvenience in the calibration design possibly needed for different chips and increases cost due to larger chip size needed to accommodate the calibration standards. In addition, accurate characterization of these on-wafer standards, namely determining accurately the capacitance of the open, the inductance of the short and load, and the delay time of the thru and line, which in general change with frequency and process variance, is not an easy task; hence leading to errors in measurements. A more convenient and low-cost, and perhaps practically more precise, solution is to design and fabricate calibration standards on a separate RF substrate board and use the same standards for every on-wafer measurement over the same frequency range that the standards are designed for. These off-chip or off-wafer calibration standards are essential the same as the on-wafer standards used in the on-wafer calibration methods such as SOLT or TRL and hence can be designed similarly using any proper RF substrate. It is noted, however, that since the RF pads and interconnects may be different from one design to another, the off-wafer calibration standards do not contain the RF pads and interconnects, and the measurement reference plane is set at the probe tip. The interface between the probe and the off-wafer standards is different from that between the probe and the RF pads of the DUT. As compared to on-wafer calibration, which can remove the effects of the RF pads and interconnections from these pads to a DUT, off-wafer calibration, however, cannot remove these effects. In order to remove these effects, an accurate de-embedding needs to be carried out for the measured data. The use of on-wafer or off-wafer calibration standards is thus dependent on RFIC designers considering factors involving convenience, cost, and accuracy for a particular chip design. One particular note is that, when different fabrication technologies, such as 0.045, 0.090, and 0.13 μm CMOS, are used frequently, different on-wafer calibration standards need to be designed, fabricated, and verified accurately prior to use for on-wafer calibration, which are inconvenient, time-consuming and may not be suitable for certain design environments. For certain measurement environments and requirements, off-wafer calibration standards may be preferred since they only need to be accurately designed, characterized, and verified once and can be used for many different measurements for various processes. Figure 15.23 shows the off-wafer calibration standards, known as Impedance Standard Substrate (ISS) calibration standards, made by Cascade Microtech, Inc. [7].

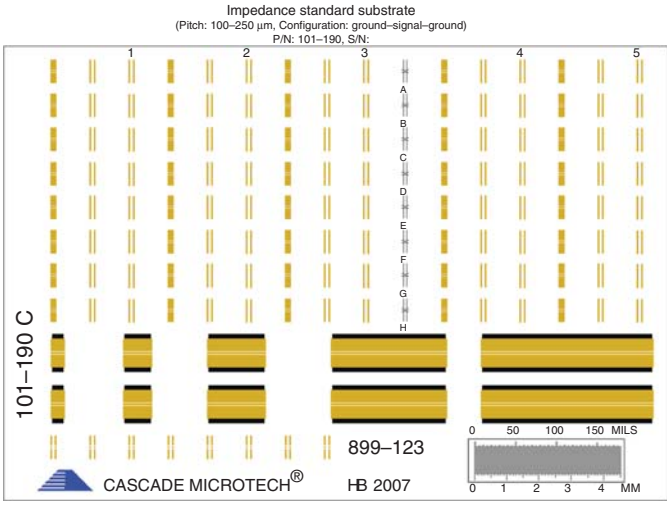


Figure 15.23. Off-wafer calibration standards. (Courtesy of Cascade Microtech, Inc.)

15.3.1.4 Measurement. For illustration purposes, we assume measurements using a VNA. For other non-VNA measurements, similar but simplified procedures can be implemented. For instance, for frequency spectrum measurement with a spectrum analyzer, only loss calibration is needed, and this can be done by measuring the insertion loss of the complete through path including connectors, adaptors, cables, probes, RF pads, and a through line, whose length is equal to the total length of the lines between the input/output RF pad and the DUT, and subtracting this loss from the amplitude of the measured frequency spectrum. The measurement of a DUT, which yields only the performance of the DUT, typically consists of three main steps: calibration, DUT measurement, and DUT de-embedding.

Calibration

In the calibration process, the calibration standards – for instance, open, short, thru, and line for SOTL or thru, reflect, and line for TRL – are measured and saved in the VNA memory in sequence according to the VNA's instructions. Characteristics of these standards are also entered into the VNA per its requirements as instructed by the VNA during the calibration process. The VNA uses these characteristics and actual measurement results of the standards to extract the error factors, caused by the VNA, RF on-wafer probes, adaptors, cables, and possibly RF pads, needed for the calibration. These errors will be removed from the DUT measurement results. For instance, in the TRL 2-port calibration process, there are three basic steps involving *S*-parameter measurements. The first step is the THRU step, in which the test ports are connected directly with a zero-length thru or with a nonzero-length thru. For the REFLECT step, identical one-port open (or short) standards are connected to each test port. For the LINE step, a short transmission line, whose length is different from that of the THRU is inserted between port 1 and port 2. It is noted that, depending on measurements, isolation calibration may be needed. The calibration process using calibration standards is essential for RFIC (on-wafer and off-chip measurements) and other discrete RF components such as amplifiers and filters. The main differences are in the connection between the measurement point and the DUT and the use of different calibration standards. For discrete RF components, the connections are between the connectors and the DUT while, for on-wafer RFIC, the connections are at the locations where the RF probes touch the RF pads of the RFIC. The other difference is in the use of regular connector-based calibration standards included with the VNA for discrete RF components and on-wafer calibration standards described earlier. The calibration procedure for both discrete RF components and RFICs follows the same guidelines by the VNA.

Once the calibration is done, conventional measurements can then be followed for the DUT according to the VNA's instructions – for example, measurement of *S*-parameters of an RFIC.

DUT Measurement

The input and output ports of a DUT need to be located at the measurement reference planes at port 1 and port 2, respectively. For a zero-length thru standard, the DUT thus does not have extra transmission lines at the input and output ports. However, when a nonzero-length thru standard is used, the DUT must have a transmission line at each of its port and these transmission lines are identical (same type, characteristic impedance, length) to each half of the thru. In this arrangement, the measurement will be properly calibrated up to the actual input and output of the DUT. That is, the extra transmission line at each port is electrically removed. One particular note is that, although a non-zero length thru has physical length, it should be specified to the VNA as having zero length in the TRL-standard definition. The actual electrical length of the thru standard must then be subtracted from the actual electrical length of each line standard in the TRL calibration definition.

DUT De-Embedding

De-embedding process involves removal of the electrical effects of elements not belonging to the DUT, which are typically the whole RF pads or parts of the RF pads. It is noted that the RF pads may or may not be removed from the measured results during a calibration process depending on the calibration standards employed. Ideally, the complete effect of the RF pads should be removed from the measurement results.

In practice, however, the RF pads may not or may only partially be removed due to the landing position of the on-wafer probes and/or the calibration standards used – for instance, in the SOLT calibration process, the probe is typically landed at the center of the pad which defines the measurement reference plane. Therefore, afterward a de-embedding process may need to be applied to the DUT measurement results to completely remove the RF pads' effects. The characterization of the RF pad, typically by its *S*-parameters, needed for the de-embedding procedure can be done through measurement or EM-based calculation of the pad.

Various de-embedding techniques have been developed for RF circuits including RFIC, MIC, and MMIC. Two de-embedding methods are described in this section. The first one is the widely used *Y*-parameter (hereafter namely Open-DUT) de-embedding technique based on admittance (*Y*) parameters and open dummy [8], which is one of the simplest methods. The second one is a more complex open-short-DUT de-embedding method based on both *Y* and *Z* (impedance) parameters, and open and short dummies [9].

Open-DUT De-Embedding.

Procedure. The Open-DUT de-embedding is based on two de-embedding structures. One is the DUT de-embedding structure that consists of a DUT and a test structure, which includes the RF pads and interconnects to the DUT, such as that in Figure 15.26(a) to be described later. Another is the open dummy, which is identical to the DUT de-embedding structure but without the DUT, such as that in Figure 15.26(b) to be described later – that is, the DUT is removed and the space is left empty between the input and output on-wafer RF pads. The procedure of the open-DUT de-embedding method is summarized in the following steps:

1. Measure the *S*-parameters of the DUT de-embedding structure using a VNA and convert them into *Y*-parameters.
2. Measure the *S*-parameters of the open dummy. Convert the *S*-parameters into *Y*-parameters.
3. Subtract the open dummy's *Y*-parameters obtained in Step 2 from the *Y*-parameters of the DUT de-embedding structure obtained in Step 1 as

$$[Y_{\text{DUT}}] = [Y_{\text{DEMB-DUT}}] - [Y_{\text{OPEN-DUMMY}}] \tag{15.2}$$

where $[Y_{\text{DUT}}]$, $[Y_{\text{DEMB-DUT}}]$, and $[Y_{\text{OPEN-DUMMY}}]$ are the admittance matrices of the DUT, DUT de-embedding structure, and open dummy, respectively. The obtained *Y*-parameters are then converted into *S*-parameters. These *S*-parameters are the *S*-parameters of the DUT itself.

It should be noted that the VNA needs to be calibrated up to the tips of the probes using calibration standards before the *S*-parameter measurement can take place as normally done.

Example. As an example to illustrate the open-DUT de-embedding process, the *S*-parameters of a DUT, which is a microstrip transmission line of 10.2-μm width and 300-μm length on Metal 6 in a 6-metal CMOS process, as shown in Figure 15.12(a), is extracted from 0.5 to 60 GHz. The on-wafer RF pad shown in Figure 15.10 is used. The *S*-parameters of the DUT de-embedding structure, which consists of the DUT and two RF pads as shown in Figure 15.12(b), mentioned in Step 1 of the foregoing de-embedding procedure, and those of the open dummy shown in Figure 15.12(c), mentioned in Step 2, are obtained by the EM simulator IE3D. For the sake of simplicity without loss of generality, EM simulations are used instead of measurement. The procedure, however, is the same for de-embedding with actual measured data, except that the EM-simulated data are replaced with those measured. It is noted that, for better de-embedding results, the metal ground in the bottom-most layer (M1), which also acts as a shield for the signal pads, are extended pass the RF pads as discussed previously in Section 15.3.1.2. This can be seen in Figure 15.12 which shows that the edge of the ground shield is displaced from the edges of the ground pads of the RF pads. The specific steps for the de-embedding are:

Step 1: Calculate the *S*-parameters of the DUT de-embedding structure in Figure 15.12(b) and convert these *S*-parameters into *Y*-parameters.

Step 2: Calculate the *S*-parameters of the open dummy in Figure 15.12(c) and convert them into *Y*-parameters.

Step 3: Obtain the DUT's *S*-parameters by subtracting the open dummy's *Y*-parameters from the *Y*-parameters of the DUT de-embedding structure and converting the results into *S*-parameters.

To verify the accuracy of the Open-DUT de-embedding, the de-embedded *S*-parameters (obtained in Step 3) and the EM-simulated *S*-parameters of the DUT, herein referred to as the actual *S*-parameters of the DUT, are compared in Figure 15.13(c). It is seen that the de-embedded and actual *S*-parameters of the DUT match well within the de-embedding frequency range of 0.5–60 GHz, validating the employed open-DUT de-embedding method.

Open-Short-DUT De-Embedding.

Procedure. The open-short-DUT de-embedding technique relies on the removal of the (external) parasitics surrounding the DUT. The procedure of the open-short-DUT de-embedding method is described using the DUT (BJT#1), open-dummy, and short-dummy de-embedding structures as shown in Figure 15.24. These structures are slightly modified from the on-wafer structures shown in Figure 15.17 to include additional interconnects to the DUT, which may be needed in practice to accommodate very small DUT such as transistors from existing on-wafer calibration structures such that those shown in Figure 15.17. The employed de-embedding structures thus serve to illustrate a more flexible (and possibly practical) implementation for de-embedding. The procedure is summarized as follows.

1. Measure the *S*-parameters of the DUT de-embedding structure [Figure 15.24(a)] and convert them into *Y*-parameters ($Y_{\text{DEMB-DUT}}$).
2. Measure the *S*-parameters of the open dummy [Figure 15.24(b)]. Convert the *S*-parameters into *Y*-parameters ($Y_{\text{OPEN-DUMMY}}$).
3. Measure the *S*-parameters of the short dummy [Figure 15.24(c)]. Convert the *S*-parameters into *Y*-parameters ($Y_{\text{SHORT-DUMMY}}$).
4. Subtract the open dummy's *Y*-parameters obtained in Step 2 from the *Y*-parameters of the DUT de-embedding structure obtained in Step 1 to obtain the *Y*-parameters of the DUT as

$$[Y'_{\text{DUT}}] = [Y_{\text{DEMB-DUT}}] - [Y_{\text{OPEN-DUMMY}}] \tag{15.3}$$

Convert $[Y'_{\text{DUT}}]$ to impedance matrix $[Z'_{\text{DUT}}]$.

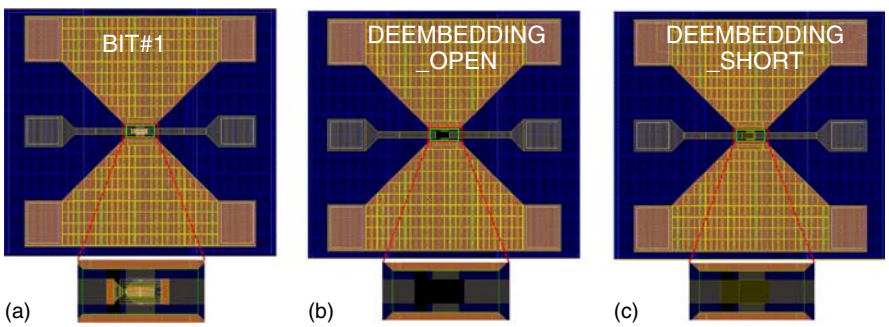


Figure 15.24. DUT de-embedding structure (a), open-dummy de-embedding structure (b), and short-dummy de-embedding structure (c).

5. Subtract the open dummy's *Y*-parameters obtained in Step 2 from the short dummy's *Y*-parameters obtained in Step 3 to obtain the *Y*-parameters of the short as

$$[Y_{\text{SHORT}}] = [Y_{\text{SHORT-DUMMY}}] - [Y_{\text{OPEN-DUMMY}}] \tag{15.4}$$

   Convert $[Y_{\text{SHORT}}]$ to impedance matrix $[Z_{\text{SHORT}}]$.
6. Subtract the short's *Z*-parameters obtained in Step 5 from the DUT's *Z*-parameters obtained in Step 4 to obtain the final *Z*-parameters of the DUT as

$$[Z_{\text{DUT}}] = [Z'_{\text{DUT}}] - [Z_{\text{SHORT}}] \tag{15.5}$$

   Convert this impedance matrix to *S*-parameters which represent the final de-embedded *S*-parameters of the DUT.

As for the Open-DUT de-embedding measurement, the VNA needs to be calibrated up to the tips of the probes using calibration standards before measuring the *S*-parameters.

Example. As an example to illustrate the open-short-DUT de-embedding process, we use a transistor available in Jazz 0.18-μm SiGe BiCMOS process [10] as the DUT (BJT#1). The on-wafer de-embedding structures as shown in Figure 15.24 are used. For demonstration purpose, the IE3D-simulated *S*-parameters of the passive structures and the Cadence-simulated *S*-parameters of the transistor based on its model from the PDK are used instead of actual measurement. The procedure, however, is the same for de-embedding with actual measured data. Figure 15.25 compares the de-embedded *S*-parameters and the simulated *S*-parameters of the transistor (DUT), herein referred to as the actual *S*-parameters of the DUT, to verify the accuracy of

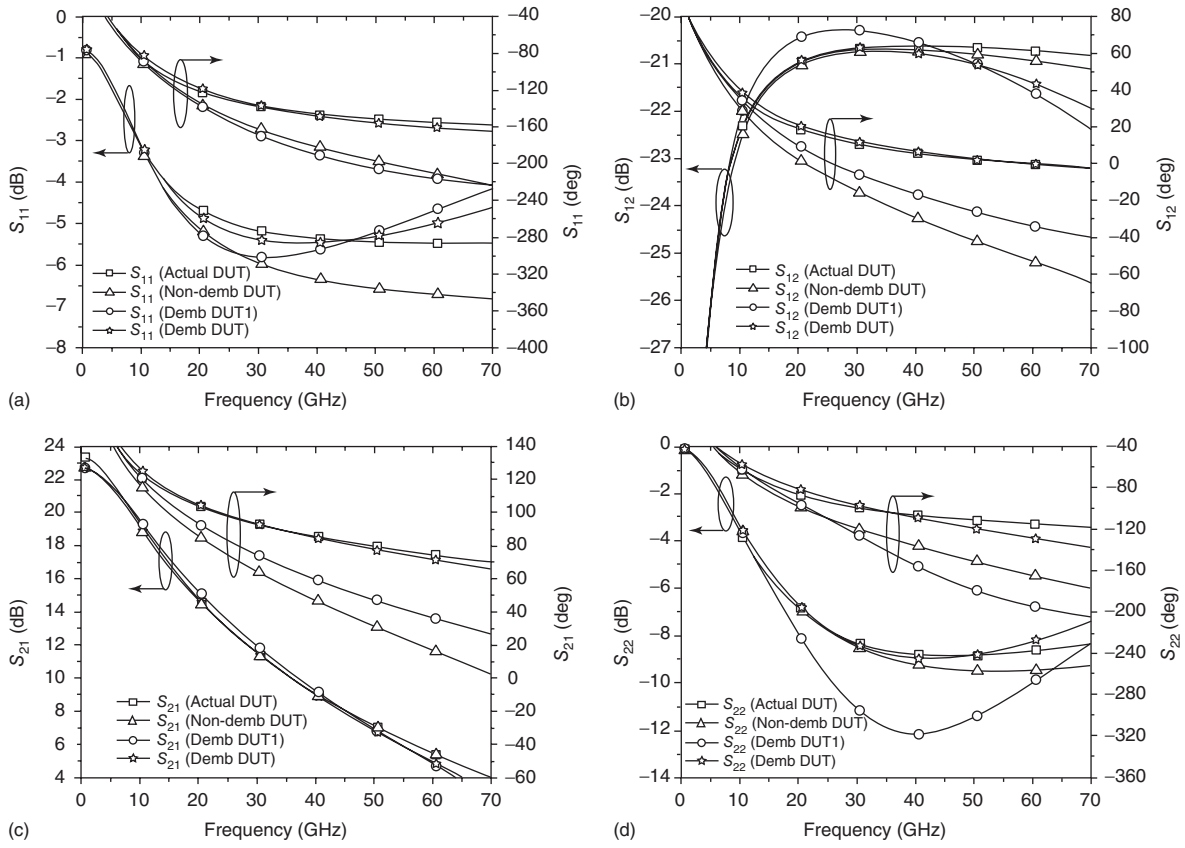


Figure 15.25. (a–d) Comparison between the DUT's de-embedded and actual *S*-parameters.

the de-embedding. The de-embedded *S*-parameters of the DUT based on the open-DUT de-embedding method and the non-de-embedded *S*-parameters of the DUT (raw data without de-embedding) are also included for comparison. As can be seen, the open-short-DUT de-embedding results match reasonably well with the actual data of the DUT close to 70 GHz. It is also observed that the open-short-DUT de-embedding method provides more accurate de-embedding results than the open-DUT method for the considered de-embedding case. While the open-DUT de-embedding method is simple and preferred for de-embedding, the open-short-DUT de-embedding technique may be more suitable for modeling DUT's of more complex structures.

15.3.1.5 Comparison Between SOLT and TRL Calibration. The accuracy of a calibration technique depends significantly on the accuracy of its calibration standards. Virtually, many different calibration methods can be configured and corresponding calibration standards can be designed according to a circuit testing environment such as on-chip or off-chip, transmission lines or interconnects used at the circuit's terminal ports, and operating frequency ranges. The choice of a calibration method is generally based on the measurement environment and the availability or feasible design of accurate standards. The SOLT and TRL discussed earlier are the two most widely used calibration techniques in RFIC testing. Each has its own unique advantages and disadvantages. The following provides important characteristics of the SOLT and TRL, some of which have been discussed previously, and draw a conclusion for a suitable method.

The SOLT calibration method requires four standards (short, open, load, and thru) and all of them must be accurately characterized. The SOLT thus relies heavily on the accurate model for each of the standards. On-wafer open standard exhibits inherent fringing capacitance at the open end, which may be substantial and is not easy to be determined accurately at high frequencies. This fringing capacitance also changes with frequency and process variance. On-wafer short standard exhibits inductance that varies with different frequencies and is affected by process variation. On-wafer load standard not only consists of a resistive load, but also a parasitic inductor that depends on frequency and process variation. Therefore, accurate characterization of on-wafer short, open, and load standards becomes complex at high RF range due to the rising effects of parasitic capacitances and inductances at these frequencies. Since SOLT uses only one transmission line (for the line standard), the on-wafer probes do not need to be moved during a measurement; the probe locations are fixed at the measurement ports, leading to convenience in the probe operation. At low frequencies, however, the parasitic components of the SOLT on-wafer standards are relatively small and vary less, and hence may be ignored. It is noted that high-quality SOLT standards are easier to be obtained in coaxial and waveguide environments than in planar transmission lines, such as CPW or microstrip line, typically used for on-wafer measurement, especially at high frequencies.

TRL calibration process uses only three standards (thru, line, and reflect). TRL relies on the characteristic impedance of transmission lines rather than on a set of discrete impedances of open, short, and load standards as for SOLT. Moreover, TRL does not require all of the standards to be completely modeled, unless a standard is used to determine the reference plane. For instance, the reflect standard needs to be the same on both ports, but its exact value does not need to be known unless it is used to set the reference plane. The line and thru standards are both assumed to have a perfect match and same characteristic impedance, but the exact lengths need not be known. Hence, the TRL calibration process does not rely on the exact models of the standards and is therefore more immune to model errors. Often, the reference plane is set by the thru standard. The TRL standards are easier and more accurately to be fabricated than the SOLT standards. The TRL method requires (thru and line) transmission-line standards of different lengths and therefore cannot be used with fixed-location on-wafer probes. To accommodate the TRL calibration, the probes need to be moved for different lines, causing inconvenience in the probe operation. A limitation of the TRL technique is the limited bandwidth of the line standards. Most line standards can only be used over an 8:1 frequency range. Another problem is, at low frequencies, line standards can become too long to be fabricated on-wafer. However, when accurate on-wafer line standards are available, the TRL method usually offers better accuracy than the SOLT technique. For measurements over broader frequency ranges than 8:1, several line standards may be required for the TRL. The multiline TRL calibration needed for very wideband measurement, however, suffers one important drawback. That is, a set of lines, some of which are quite long

in order to cover the low-end frequencies, is required, resulting in the use of expensive space on the wafer. To overcome the band-limitation problem of TRL, LRM (line-reflect-match) calibration using a broadband load (match) instead of long transmission lines as in TRL can be employed.

The foregoing discussion indicates that, for low frequency on-wafer measurement, SOTL is a better choice than TRL since very accurate SOTL standards can be achieved, while TRL requires unacceptable long transmission lines. At high frequencies, SOTL can be used but the standards have to be redefined at different frequency ranges of the measurement bandwidth. That is, the offset values of the parasitic capacitance and inductance of the open and short standards, respectively, have to be determined for different sub-bands of the frequency band and input into the VNA accordingly. These sub-bands, however, need to be small enough so that the offset values are relatively constant across each sub-band. This implies that that many sub-bands need to be used, thereby slowing down the measurement process. Moreover, it may not be possible to determine the fringing capacitance of the open standard and the inductance of the short standard very accurately over frequencies, particularly at very high frequencies. TRL, which does not use the open and short standards, overcome this problem. At high frequencies, the required lines and thru standards are not very long and can be easily designed with good quality using an EM simulator. TRL thus generally offers a better calibration solution than SOTL for high-frequency on-wafer measurement.

Table 15.2 summarizes the SOTL and TRL comparison.

15.3.2 Off-Chip Measurement

Off-chip measurement does not require on-wafer probes and station, and the measurement is carried out using regular equipment and conventional connecting interface, such as SMA connectors, between measurement instruments and RFICs. To facilitate off-chip measurement, an RFIC chip is normally mounted inside a package which is attached to a PCB using an appropriate RF substrate suitable for the interested frequencies. The DC bias ports and RF terminal ports of the circuits are bonded via their on-chip bonding pads onto the package's pins, which are soldered onto the conductors (for DC signals) and transmission lines (for RF signals) on the PCB. The PCB also accommodates necessary off-chip components such as external bias circuitry. Typical bonding pads have size of 100 μm × 100 μm and use multiple metal layers interconnected through via-holes. To reduce the parasitic effects, the length of a bonding wire from a bonding pad to the inner pin of the package should be as short as possible. Figure 15.26 shows a package housing an RF CMOS chip and a PCB holding the package. Sometimes, an RF chip can be directly attached on to a PCB board, hence

TABLE 15.2. Comparison of On-Wafer SOLT and TRL Calibration

Calibration method	Advantage	Disadvantage
SOLT	– Impedance reference set by Load standard – Simple, easy design for standards – Use of small on-wafer area – Working well at low frequencies – Used with fixed probes – Broad band	– Requiring very well-defined standards – Open, Short, and Load standards difficult to model on-wafer – Low accuracy at high frequency – Limited use at high frequencies due to parasitics (e.g., > 20 GHz)
TRL	– High accuracy, especially on wafer and at high frequencies – Minimal standard definition. Electrical characteristics not strictly required	– Requiring very good transmission lines – Large on-chip standards – Band limited – Standards too long at low frequency – Requiring movable probes – Line standards too long at low frequencies (e.g., 2 GHz) – Not broad band

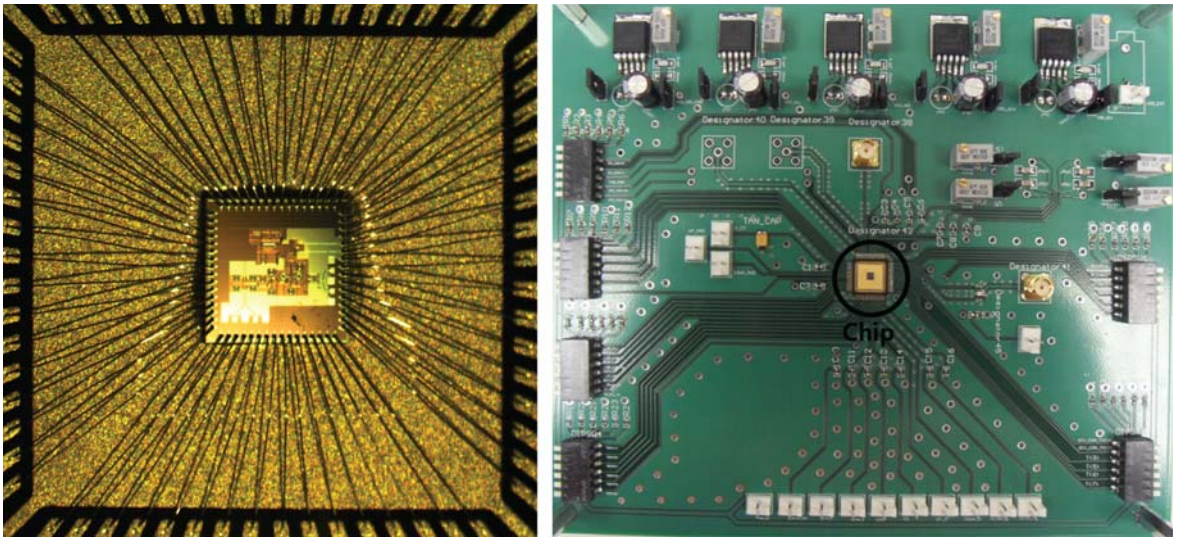


Figure 15.26. (a) (Open-lid) package with an RF CMOS chip inside and (b) PCB holding an (open-lid) packaged RF CMOS chip.

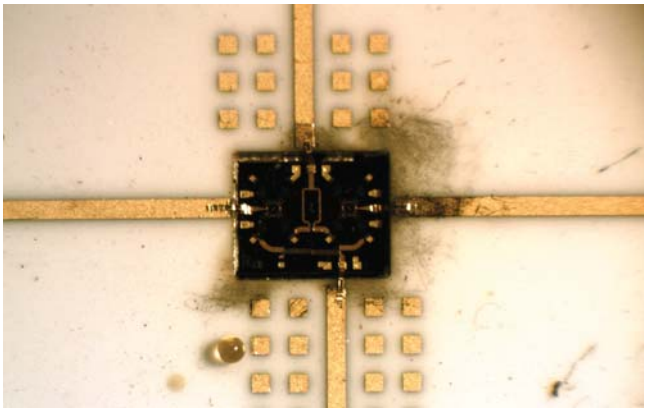


Figure 15.27. Photograph of an RF chip mounted directly onto a PCB. Bonding wires are used for interconnection between the chip and PCB.

no package is needed, as can be seen in Figure 15.27. This procedure, however, requires special mounting process and special substrate for the PCB that is suitable for chip attachment and bonding of wires between the chip and PCB.

A RFIC chip can be packaged either for (off-chip) measurement purpose of the chip itself or for use as a packaged standalone circuit or an integral part of another circuit, subsystem, or system. In the first case, the package is not part of the circuit and, therefore, the package effects need to be removed from the measured data. To that end, an accurate RF model for the package must be known and used for the de-embedding purpose. In the second case when the package is included as part of an RFIC chip (packaged chip), an accurate equivalent-circuit model for the package is needed to facilitate accurate design of the packaged chip. Such model can be obtained through measurement and/or EM simulation. In the circuit design, the package model's elements such as inductors, capacitors and resistors are typically absorbed into the matching networks of the chip.

One important procedure that may help determine reason why an RFIC fails to perform as calculated, apart from incorrect calculations due to incorrect models and/or negligence of circuit elements, is "de-process." De-process refers to a technique that removes certain dielectric layers in a silicon chip

to expose an embedded metal buried inside the chip so that a probe can be used to capture a signal on that metal. The de-processing technique, although is expensive and not available in typical measurement facilities, provides lower cost and quicker solution for circuit trouble-shooting than re-taping out the circuit. This technique is particularly useful for RFIC design as it is very difficult, if not possible, to trouble-shoot a non-working chip due to the fact that typically only the input and output signals can be measured via visible ports. In simulations, this is of course not a problem since we can always capture signals at any node in a circuit and detect possible problems. For example, in an RFIC that has subcircuits A and B connected using metal 3 (which lies between lower and upper metal layers), an on-wafer probe cannot be used to detect signals between A and B since metal 3 is buried inside the chip. To solve this problem, de-process can be used to remove the dielectrics above metal 3 (e.g., within a 50 μm × 50 μm window) without affecting the functionality of A and B. After that, metal 3 is exposed to air and a probe can then be placed on it to detect a signal.

REFERENCES

1. Advanced Design System (ADS), Agilent Technologies, Inc., Santa Clara, CA, USA.
2. Cadence, Cadence Design Systems, Inc., San Jose, CA, USA.
3. IE3D Program, Zeland Software, Inc., Fremont, CA, USA.
4. Agilent Technologies' Product Note 8510-8A, "Agilent Network Analysis Applying the 8510 TRL Calibration for Non-Coaxial Measurements," pp. 1–24, May 2001.
5. Agilent Technologies' Application Note 1287-9, "In-Fixture Measurements Using Vector Network Analyzers," pp. 1–32, Aug. 2000.
6. Cascade Microtech's Application Note, "On-Wafer Vector Network Analyzer Calibration and Measurements," pp. 1–13.
7. Cascade Microtech, Inc., Beaverton, Oregon, USA.
8. P. J. van Wijnen, H. R. Claessen, and E. A. Wolsheimer, "A New Straightforward Calibration and Correction Procedure for "On-Wafer" High Frequency S-Parameter Measurements (45 MHz–18 GHz)," in Proc. IEEE Bipolar Circuits and Technology Meeting, 1987, pp. 70–73.
9. M. C. A. M. Koolen, J. A. M. Geelen, and M. P. J. G. Versleijen, "An Improved De-Embedding Technique for On-Wafer High-frequency Characterization," in Proc. IEEE Bipolar Circuits and Technology Meeting, 1991, pp. 188–191.
10. Jazz Semiconductor, 4321 Jamboree Road, Newport Beach, CA 92660.

PROBLEMS

15.1 Describe in details the simulation procedure that you have used in one of your RFIC designs which may be one of the circuits described in the problem sections of other chapters.

15.2 Grounding in RFIC, especially in the high RF range, needs to be done correctly since ground affects the circuit performance substantially if not properly done. Use any available submicron process or the 6-metal process shown in Figure P15.2. Consider two ground configurations: one is a stacked metal ground consisting of all metals from the bottom to the top interconnected through vias in any available submicron process or a ground consisting of stacked metals 1–6 in the 6-metal process shown in Figure P15.2. Another is a ground using the top-most metal or metal 6 in Figure P15.2. Each metal used for the ground is a rectangular having length L and width W. Perform the following using an EM simulator:

a) Assume the length L of the single-metal ground is 50 and 100-μm. Calculate and plot the inductance, resistance, and quality factor versus frequency from 1 to 60 GHz for the ground

width W of 150, 300 and 600 μm. Compare these results and draw conclusions between different width and length.

b) Repeat Part (a) for the stacked metal ground. Compared the results between the single- and stacked-metal grounds and draw a conclusion.

15.3 Consider a LNA designed in one of the problems in Chapter 11 or a simple single-stage cascode LNA with (grounded) source degeneration inductor designed at 35 GHz using any available submicron process. Simulate the gain and noise figure for three cases: one with ideal ground and others with single-metal ground and stacked metal ground described in Problem 15.2. Compared the results and draw a conclusion.

15.4 Consider an on-wafer "short" standard fabricated in a 0.18-μm CMOS process to be used for the SOLT calibration in a VNA. The short is realized by connecting an on-wafer RF pad on the top-most metal layer 6 to the bottom-most metal layer 1 (used for ground) by an interconnect consisting of a series of vias connecting all the six metal layers. The distance from metal 1 to metal 6 is 10 μm. The interconnect is modeled by an equivalent circuit as shown in Figure P15.1. It is electrically represented by a short transmission line having 400-Ω characteristic impedance. Neglect the series resistance, shunt conductance and capacitance, determine the time delay of the interconnect, which can be used for the VNA calibration.

15.5 Consider the short standard described in Problem 15.4 and assume the calibration frequency is 35 GHz.

a) Calculate the approximate inductance, capacitance, resistance, and conductance per unit length of the transmission line representing the interconnect.

b) Derive an expression for the difference between the time delays of the interconnects with and without the resistance, conductance and capacitance (shown in Figure P15.1) in terms of the interconnect length, phase constant, and frequency.

c) Calculate the time delay of the interconnect considering its resistance, conductance, and capacitance. Compare and discuss the results between this case and the case in Problem 15.4 for calibration purposes.

15.6 Design an on-wafer RF pad based on the GSG configuration on any submicron CMOS/BiCMOS process or the CMOS profile shown in Figure P15.2. Describe the designed RF pad with dimensions and metals used, and provide its equivalent-circuit model and the calculated S-parameters from the equivalent-circuit model and from an EM simulator. Comment on the results.

15.7 On-wafer RF pad is used for on-wafer measurement and it may or may not be removed through calibrations. For instance, the SOLT calibration may not remove the on-wafer RF pad completely due to the landing positions of the on-wafer probe, or the calibrations using off-chip on-board calibration standards fabricated on a separate substrate cannot remove the on-wafer RF pad. Describe a possible de-embedding procedure different from those described in section "DUT De-embedding" to remove the RF pads at the input and output of a two-port DUT that cannot be removed through calibration, from the measured S-parameters of the combined DUT and two RF pads.

15.8 Consider a CMOS process profile as shown in Figure P15.2. You can also use any available submicron CMOS process having at least six metal layers. Design and layout on-wafer TRL calibration standards for 30–40 GHz measurement frequency range. The on-wafer RF pad is 100 μm × 100 μm with 100-μm

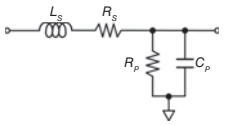


Figure P15.1.

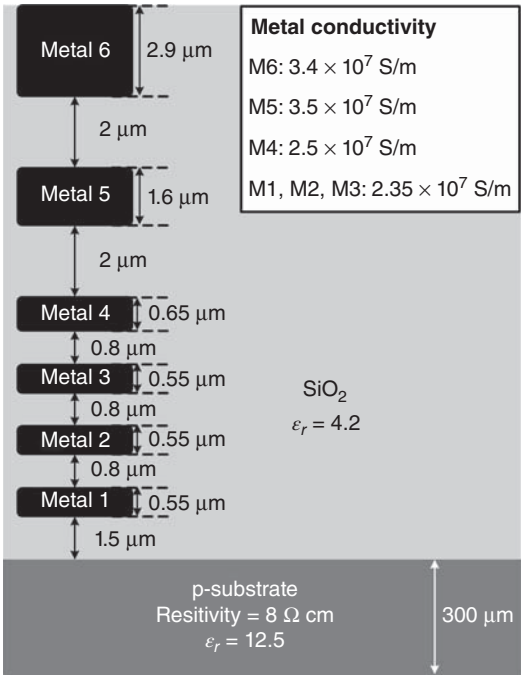


Figure P15.2.

pitch. Provide a detailed description of the design including simulations and required parameters for VNA calibration such as the electrical length of the line. Use an EM simulator in the design and simulations.

15.9 Describe an on-wafer SOLT calibration procedure using any available VNA. The description should be in sufficient details for any RFIC designers without prior knowledge of calibration to understand and perform the calibration by themselves.

15.10 Describe an on-wafer TRL calibration procedure using any available VNA. The description should be in sufficient details for any RFIC designers without prior knowledge of calibration to understand and perform the calibration by themselves.

15.11 Assume we want to characterize on-wafer the S-parameters of an RFIC LNA from 100 MHz to 110 GHz. Describe in details a calibration technique or a combination of calibration techniques that allows an accurate on-wafer measurement to be conducted across such an extremely wide frequency range.

15.12 Derive Eq. (15.2). Describe assumptions you make if any in deriving this equation.

15.13 Consider the on-wafer RF pad as shown in Figures 15.10(a) or 15.17(e), or designed in Problem 15.6, or any available on-wafer RF pad on a submicron CMOS/BiCMOS process. (If you use an on-wafer RF pad different from that in Figure 15.10(a), 15.17(e), or Problem 15.6, then provide a complete description of that RF pad including dimensions and metal layers used for the signal and ground pads, and the process profile.) Consider a 400-μm long 50-Ω CPW with signal and ground lines on the top-most metal layer as a DUT. Any CMOS/BiCMOS process profile, or the one shown in Figure P15.2, can be used.

1) Perform the Open-DUT de-embedding procedure described in section "Open-DUT De-Embedding", using an EM simulator, to obtain the *S*-parameters of the CPW (DUT) from 10 MHz to 60 GHz as following:
   (a) Calculate the *S*-parameters of the CPW with the input and output RF pads. (Provide complete physical dimensions of the CPW.)

(b) Calculate the *S*-parameters of the open dummy. Provide a complete description of the open dummy.

(c) Obtain the *S*-parameters for the CPW only (no RF pads). This is the de-embedded *S*-parameters.

2) For verification purpose, calculate the *S*-parameters of the CPW without the input and output RF pads (actual *S*-parameters), and compare the de-embedded *S*-parameters with the actual *S*-parameters. Comment on the results. Provide rationale for any discrepancy between the de-embedded and actual data.

15.14 Repeat Problem 15.13 using the open-short-DUT de-embedding procedure described in section "Open-Short-DUT De-Embedding". Describe the short dummy. Compare and comment on the results obtained in this problem and the open-DUT de-embedding in Problem 15.13.

15.15 If you also do Problem 15.7, then repeat Problem 15.13 using the de-embedding procedure described in Problem 15.7, and compare the results with those in Problems 15.13 and/or 15.14.

15.16 Repeat Problem 15.13 for a DUT represented by a 70-Ω microstrip line having length of 500 μm on the top-most metal layer and ground plane on the bottom-most metal layer.

15.17 Repeat Problem 15.16 for a 500-μm-long 70-Ω CPW with the signal and ground lines on the top-most metal layer. If you also do Problem 15.16, then are the de-embedding results better for microstrip or CPW? Comment on the results obtained in these two problems and provide rationale for your observation.

15.18 Repeat Problem 15.17 for a 500-μm-long 70-Ω conductor-backed CPW. The signal and ground lines are on the top-most metal layer and the back-conductor is on the bottom-most metal layer. If you also do Problems 15.16 and/or 15.17 then comment on the de-embedding results obtained in Problems 15.16 and/or 15.17, and 15.18, and provide rationale for your observation.

CHAPTER 16

SYSTEMS

RF engineering and scientific applications, whether in communications or sensing, are accomplished through RF systems – not individual RF circuits or components – and so the ultimate objective of RF engineers is developing RF systems. Knowledge of RF systems, besides passive and active RF circuits and antennas, is therefore critical for RF engineering practice. Typical RF systems consist of antennas, RF, and low-frequency components making up receiver and transmitter subsystems, and signal and data processing subsystem. The type of system depends on the architecture of transmitter and/or receiver, which in turn depends upon applications and desired outcomes. There exist various system architectures – some of which are widely used while others are implemented only for specific applications – and, as for RF circuits and antennas, new system architectures can be conceived. In general, each system architecture has its own advantages and disadvantages for certain applications and selection of a proper system is always a crucial task for RF engineers. In this chapter, the commonly used pulsed and frequency-modulated continuous wave (FMCW) systems along with the widely used receiver architectures of homodyne and super-heterodyne will be addressed. We will look at these architectures from a general perspective without delving much in details to provide an overall concept of systems and sufficient fundamentals for RF engineers. While specific applications will be used for illustration, they should not be viewed as the exclusive applications of these systems. It is important to note that RF systems are based on related fundamentals in general and employ like RF components, making them – from a general hardware perspective – similar, no matter what applications they are used for. Knowledge of RF systems (and their components) for one application (e.g., radar) can hence be applied to other applications (e.g., wireless communications.)

16.1 FUNDAMENTALS OF SYSTEMS

16.1.1 Friis Transmission Equation

Friis transmission equation, providing a very simple estimate of the received power with respect to the transmitted power for a general RF system, is a very basic equation for communications and sensing. Figure 16.1

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

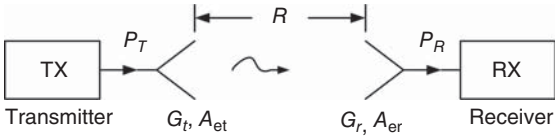


Figure 16.1. A simple RF system's block diagram. P_t is the output power of the transmitter (TX), which is assumed to be equal to the power transmitted by the transmit antenna. P_r is the power arriving at receiver (RX), which is assumed to be equal to the power received by the receive antenna. G_t, A_{et} and G_r, A_{et} are the gain and effective antenna aperture of the transmit and receive antennas, respectively. R is the distance between the transmit and receive antennas.

shows a simple (bi-static) RF system consisting of transmitter, receiver, and antennas. For simplified illustration without loss of generality, we assume that the system and the transmission media are ideal – in which, the system has matched polarization; the antennas, transmitter, and receiver are perfectly matched; the antennas are lossless; there is no scattering in signal's transmission and reception; and the antennas and transmission media are lossless.

The effective antenna area or aperture, which specifies the area of antenna that captures incoming energy, is defined as the ratio between the power received by the antenna and its power density. Mathematically, it can be derived as

$$A_e(\theta, \phi) = \frac{\lambda^2}{4\pi} G(\theta, \phi) \tag{16.1}$$

where θ and ϕ are the (angle) coordinates in a spherical coordinate system, λ is the operating wavelength, and $G(\theta, \phi)$ is the gain of the antenna. Assume the transmit antenna is isotropic, the power density at the receive antenna, produced by the power illuminating from the transmit antenna, is given as

$$P_{\text{dr}} = \frac{P_t}{4\pi R^2} \tag{16.2}$$

The receiving power density corresponding to a transmit antenna having gain G_t, is

$$P_{\text{dr}} = \frac{P_t G_t}{4\pi R^2} \tag{16.3}$$

The power received by the receiver is given as

$$P_r = P_{\text{dr}} A_{\text{er}} = \frac{P_t G_t}{4\pi R^2} A_{\text{er}} \tag{16.4}$$

We can derive, upon using (16.1) and (16.4),

$$\frac{P_r}{P_t} = G_t G_r \left(\frac{\lambda}{4\pi d} \right)^2 \tag{16.5}$$

which is known as the *Friis* transmission equation, which gives the optimum received power from a given transmitted power.

In practice, losses occur in the system itself due to various reasons such as polarization mismatch between the transmit and receive antennas, scattering, mismatch loss at the transmitter and receiver, and loss in antennas, and in the transmission medium. Taking these losses into account yields

$$\frac{P_r}{P_t} = \frac{G_t G_r}{L} \left(\frac{\lambda}{4\pi R} \right)^2 \tag{16.6}$$

where L represents the total loss encountered by the system during operation including the system loss and loss of the propagating medium. The medium loss is accounted for by the loss factor $e^{-2\alpha R}$, where α represents the attenuation constant of the medium.

The maximum range of detection or communication corresponds to the received power equal to the minimum power $P_{r,\min}$, that can detected by the receiver or the receiver sensitivity, and can be determined from (16.6) as

$$R_{\max} = \frac{\lambda}{4\pi}\sqrt{\frac{P_t}{P_{r,\min}}\frac{G_t G_r}{L}} \tag{16.7}$$

As can be seen, in order to double the range, the transmitting power must be increased four times.

16.1.2 System Equation

We consider a monostatic system using two separate antennas colocated or same antenna for transmitting and receiving, as shown in Figure 16.2. The power density at the target is the same as that given in (16.3) with R denoting the distance from the antennas to the target as seen in Figure 16.2. The power transmitted by the transmit antenna is intercepted and reradiated (scattered and reflected) by the target in different directions depending on the target's scattering characteristics. In the direction of the receive antenna, the reradiated power is derived as

$$P_\sigma = \frac{P_t G_t}{4\pi R^2}\sigma \tag{16.8}$$

where σ represents the back-scattering cross section of the target seen by the system, which is commonly known as the "radar cross section" (RCS) of the target.

The RCS mainly depends on the operating wavelength and angle from which the target is viewed by the system. It constitutes an important parameter in the system design and may be calculated or measured. RCS is defined as the effective area of the target that captures the transmitted signal and isotropically radiates the entire power incident upon it as

$$\sigma = \frac{\text{power scattered toward source per unit solid angle}}{\text{incident power density at the target}/4\pi} \tag{16.9}$$

In other words,

$$\sigma = \lim_{R\to\infty} 4\pi R^2 \frac{|\vec{E}_s|^2}{|\vec{E}_i|^2} \tag{16.10}$$

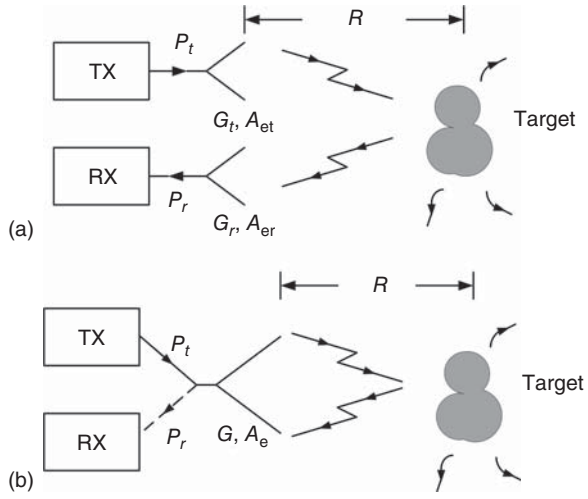


Figure 16.2. A simple monostatic system with two antennas (a) and single antenna (b).

where $\vec{E}_s$ is the scattered field at the receiving antenna, $\vec{E}_i$ is the incident field at the target, and R is the distance to the target or range. When the range R is very large with respect to wavelength, the incident wave is considered a (uniform) plane wave. As can be inferred from its definition, the RCS provides system designers some crucial characteristics of the desired target observed by the system.

The power density at the receive antenna due to the power return from the target can be derived, under ideal conditions without any loss, as

$$P_{d,r} = \frac{P_t G_t \sigma}{(4\pi R^2)^2} \tag{16.11}$$

Using the antenna effective area as given in (16.1), the ratio between the power of the target's reflected signal received at the receiver and the power transmitted by the transmitter can be derived as

$$\frac{P_r}{P_t} = \sigma \frac{G_t G_r \lambda^2}{(4\pi)^3 R^4} \tag{16.12}$$

which is commonly known as the *radar equation*, which we refer herein as the *system equation*.

For practical systems operating under real conditions, the system equation becomes

$$\frac{P_r}{P_t} = \sigma \frac{G_t G_r \lambda^2 L}{(4\pi)^3 R^4} \tag{16.13}$$

where $L < 1$ is the total loss of the system under operation including the system loss and medium loss. It is noted that the propagating loss factor is $e^{-4\alpha R}$ accounting for the propagation over twice of the target's range. The maximum range can then be determined from (16.13) as

$$R_{\max} = \left[\frac{P_t \sigma G_t G_r L \lambda^2}{(4\pi)^3 P_{r,\min}} \right]^{1/4} \tag{16.14}$$

which is proportional to $P_t^{1/4}$. We can now see that, in order to double the maximum range, the transmit power needs to be increased by 16 times, which is significant and may not be achievable for CMOS devices at RF frequencies for high-power applications, particularly in the millimeter-wave regime. Equation (16.13) suggests that larger target's RCS leads to easier detection. As can be recognized by now, the RCS of objects, such as air planes, is a very important parameter to be considered in the design of objects and systems used to detect these objects. When the transmit and receive antennas are the same, the system equation becomes

$$\frac{P_r}{P_t} = \sigma \frac{G^2 \lambda^2 L}{(4\pi)^3 R^4} \tag{16.15}$$

where G is the antenna gain, and the corresponding maximum range is

$$R_{\max} = \left[\frac{P_t \sigma L G^2 \lambda^2}{(4\pi)^3 P_{r,\min}} \right]^{1/4} \tag{16.16}$$

16.1.3 Signal-to-Noise Ratio of System

In practical operations, system's performance is affected by noise. Noise affecting a system operation can be classified into two kinds: external noise and internal noise. External noise represents noise caused by the environment surrounding the system, including noise injected from nearby stationary and moving objects. This noise is typically large at low frequencies but small in the RF range and is, in general, negligible as compared to the internal noise generated by the RF receiver itself. Receiver noise is the dominant noise in a system and is inherent in the receiver. In operation, the output signal of the receiver includes signals

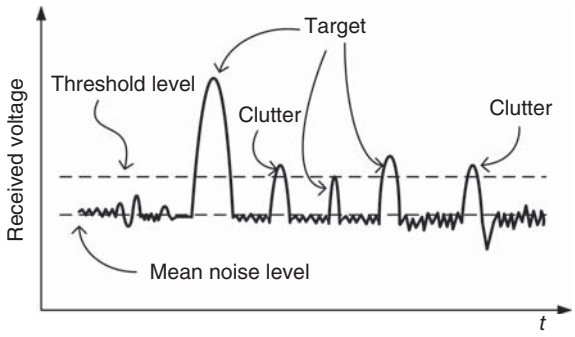


Figure 16.3. Output voltage of a receiver.

produced by the desired targets as well as those from clutters, external noise, and interference. Figure 16.3 shows a sketch of output voltage of a receiver. If the noise level contributed by the receiver itself is high or the received signal is weak, the system cannot perform accurately its intended function such as detecting a target or communication. Typically, a threshold level is used to reduce the noise and clutter effects. However, if the threshold level is set to be sufficiently high, it will reduce the sensing or communication capability of the system. For low threshold levels, on the other hand, inaccurate detection may result. Clutter effects can be reduced and identified by signal processing techniques. To increase the sensing capability or to enhance the communication performance of a system, the receiver's noise needs to be reduced or, equivalently, the receiver's signal-to-noise ratio (S/N) needs to be increased. This noise, although unavoidable, can be controlled to some extent by RF designers.

Receiver noise is, in general, contributed by three different noises. One is conversion noise generated during certain receiver operation – for example, FM-AM conversion noise. The other noise is low-frequency noise generated in the mixing process. The conversion and low-frequency noises depend on the receiver type – for instance, homodyne or FMCW receiver. The third noise contribution is thermal or Johnson noise generated by thermal motion of electrons in receiver's components. This noise always exists in receivers. We consider only thermal noise here.

Following the approach for noisy RF amplifiers described in Section 11.2.1, the maximum thermal noise power available at the receiver's input is given as

$$P_{\text{Ni}} = kTB \tag{16.17}$$

where $k = 1.374 \times 10^{-23}$ J/K is the Boltzmann's constant, T is the temperature in kelvin (K) at the receiver's input, and B is the noise bandwidth in hertz (Hz), which is the absolute RF bandwidth over which the receiver operates. The available noise power P_{Ni} is independent of the receiver's operating frequency. In typically operating room temperature (62°F), the thermal noise is about −174 dBm/Hz. As can be seen, this noise can be sufficiently large over a large bandwidth that degrades the noise performance of receiver and hence system substantially.

We define an ideal or noiseless receiver as a receiver that adds no additional noise as the input thermal noise P_{Ni} passes through it, except increasing the thermal noise level by the gain of the receiver. We now consider an actual (nonideal) receiver that adds extra noise to that produced by an ideal receiver and define the noise figure of such receiver as

$$F = \frac{\text{output noise power of actual receiver}}{\text{output noise power of ideal receiver}} = \frac{P_{\text{No}}}{GP_{\text{Ni}}} = \frac{P_{\text{No}}}{kTBG} \tag{16.18}$$

where G is the available power gain of the receiver defined as

$$G = \frac{P_{\text{So}}}{P_{\text{Si}}} \tag{16.19}$$

with P_{Si} and P_{So} being the (real) signal's available power at the input and output of the receiver, respectively. The total noise power at the output of the receiver can then be obtained as

$$P_{\text{No}} = FkTBG = kT_nBG \tag{16.20}$$

where $T_n \equiv FT$, the thermal noise power per hertz, is defined as the (equivalent) "noise temperature" of the receiver. As for low-noise amplifiers (LNAs) or mixers, the noise figure of receivers can be obtained from (16.18) and (16.19) as the ratio between the input S/N and output S/N of receivers:

$$F = \frac{P_{\text{Si}}/P_{\text{Ni}}}{P_{\text{So}}/P_{\text{No}}} = \frac{\text{signal-to-noise ratio at input}}{\text{signal-to-noise ratio at output}} \tag{16.21}$$

which is more commonly known to RF engineers than (16.18). As can be seen, the noise figure indeed reduces the output S/N level of the receiver used in subsequent processing for sensing or communication purposes. This receiver's figure of merit contributes to the overall receiver's noise and is considered one of the most important parameters of the receiver. In the design of receivers, it is important to minimize the noise figures of individual components, particularly those close to front of the receivers. A typical receiver consists of cascade of components – for example, a superheterodyne receiver front-end primarily comprising band-pass filter (BPF), LNA, mixer, and intermediate-frequency (IF) amplifier – and its noise figure is given by the same equation (11.108) for multistage amplifiers and depends on the mixer's individual components. It is noted that the noise figure of a passive component such as BPF is equal to the reciprocal of the insertion loss of that component. For instance, a BPF with a −3-dB insertion loss would have a noise figure equal to 2 or, in term of decibel, 3 dB. The (output) S/N of the receiver, making use of the system equation (16.13), can be derived as

$$\frac{S}{N} = \sigma \frac{P_t G_t G_r L \lambda^2}{(4\pi)^3 R^4 FkTB} \tag{16.22}$$

which is applicable to various systems. For pulse systems, we can write, taking into consideration the duty cycle τ/T,

$$\frac{\text{S}}{\text{N}} = \sigma \frac{T}{\tau} \frac{P_{\text{avg}} G_t G_r L \lambda^2}{(4\pi)^3 R^4 FkTB} \tag{16.23}$$

which shows that the S/N is proportional to the average transmitted power P_{avg}. It is noted that the bandwidth of a pulse with a 4-dB pulse width of τ is approximately $1/\tau$, leading to the bandwidth-pulsewidth product, $B\tau$, of about 1. The S/N can thus be increased by increasing the average power such as using a long pulse.

It is particularly noted that in actual system operations, the S/N value produced by receivers is much more important than the absolute powers of real and noise signals received by receivers.

16.1.4 Receiver Sensitivity

Receiver sensitivity indicates the minimum detectable input signal level for a receiver and hence measures the ability of a receiver to detect a signal. A system can detect a signal returned from a target or sent by another system if the received power is higher than the receiver sensitivity. The receiver sensitivity (S_R) is determined by the noise temperature (T_n), bandwidth (B), noise figure (F), and S/N ratio (S/R) of the receiver.

Figure 16.4 depicts the sensitivity required for a receiver to detect a returned signal. The noise temperature (kT) entering the receiver is increased over the receiver's bandwidth to kTB, which is then further expanded through the noise figure and S/N of the receiver to reach a value of $kTBF$(S/N). This final noise level is defined as the sensitivity of the receiver or the minimum input signal power that can be detected by the receiver:

$$S_R = kTBF\left(\frac{\text{S}}{\text{N}}\right) \tag{16.24}$$

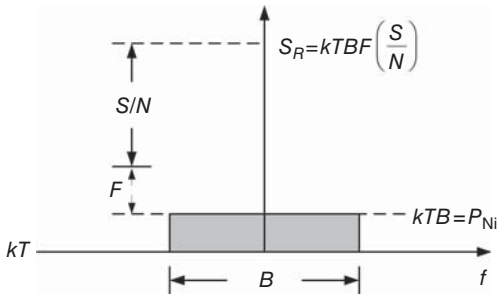


Figure 16.4. Sensitivity of a receiver. $P_{Ni} = kTB$ is the input noise power.

The maximum range of a systems, achieved when the received power is equal to the receiver sensitivity, can be rewritten using (16.13) and (16.24) as

$$R_{\max} = \left[\frac{P_t \sigma G_t G_r L \lambda^2}{(4\pi)^3 kTBF(\mathrm{S/N})}\right]^{1/4} \tag{16.25}$$

The average transmitting power is one of the controllable factors extensively used in designing a system and relates to the type of the waveform used. For instance, the average transmitting power of a pulse system, as discussed later, is equal to the product of the (peak) transmitting power and the signal's duty cycle. The average power is thus better controlled than the peak power. It is therefore useful to consider the average transmitting power in evaluating the system's parameter such as the maximum range given in (16.13).

16.1.5 System Performance Factor

The system performance factor (SF) of a system can be defined as [2]

$$\mathrm{SF} = \frac{P_t}{S_R} \tag{16.26}$$

This factor is the system's figure of merit used to measure the overall performance of a system and is one of the most important parameters in the system equation for estimating the system's range. As the minimum detectable signal corresponds to the maximum range of a system, we can derive the SF, utilizing (16.12), (16.24), and (16.26), by letting the received power equal to the receiver sensitivity as

$$\mathrm{SF} = \frac{(4\pi)^3 R_{\max}^4}{G_t G_r L \sigma \lambda^2} \tag{16.27}$$

Note that the total system loss also includes the loss of the propagating medium which is equal to $e^{-4\alpha R_{\max}}$. The performance factor given in (16.27) neglects the contribution of the receiver. In practical systems, however, the SF is limited by the actual receiver dynamic range. Hence, it is necessary to incorporate a correction into the SF.

The maximum available dynamic range, $\mathrm{DR}_{\max}$, of a receiver is the ratio of the maximum available receiving power, $P_{r,\max}$, that the receiver can tolerate without causing a distortion to the receiver's sensitivity which satisfies a specified S/N at the output of the receiver. The upper limit of the maximum available (compression-free) dynamic range is determined by the 1-dB compression point $P_{1\,\mathrm{dB}}$ of the receiver front-end's LNA in order to avoid its saturation, while the lower limit is determined by the receiver's sensitivity. For safety considerations in practical systems, the maximum available receiver power needs to be below the 1-dB compression point of the LNA.

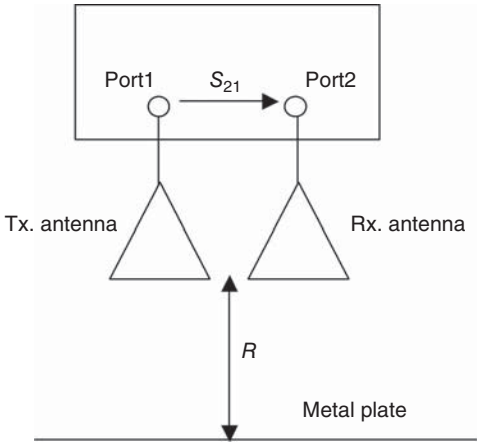


Figure 16.5. Measurement of the transmission loss L_t using a network analyzer. R is the stand-off distance between the antennas and a metal plate.

The maximum available receiving power of a system's receiver occurs when the system is directed toward a metal plate. The maximum available transmitting power $P_{t,\max}$ can be estimated from the maximum available receiving power as [6]

$$P_{t,\max} = P_{r,\max} + L_t \leq (P_{1\text{dB}} + L_t) \quad \text{(dB)} \tag{16.28}$$

where L_t, representing the transmission loss, is considered the difference between the transmitted and received power when their corresponding antennas are directed to a metal plate placed at a stand-off distance, R, as illustrated in Figure 16.5. It should be noted that the above analysis is valid only if the maximum receiving power is less than the saturating power of the receiver.

The transmission loss L_t (as noted by S_{21} in Figure 16.5) is caused by the spreading loss, the antennas' mismatch and efficiency, and other practical losses arising from connectors and cables. The transmission loss can be calculated or measured using a network analyzer if antennas are available, as shown in Figure 16.5.

Figure 16.6 illustrates the SF and dynamic range of a system. The SF (in decibel) can be found using (16.26) and (16.28) as

$$\text{SF} = (P_{t,\max} - S_R) = (P_{r,\max} + L_t - S_R) \quad \text{(dB)} \tag{16.29}$$

The receiver's maximum available dynamic range, $\text{DR}_{R,\max}$, can be defined as the difference between the maximum available receiving power and the receiver's sensitivity [3]:

$$\text{DR}_{R,\max} = (P_{r,\max} - S_R) \quad \text{(dB)} \tag{16.30}$$

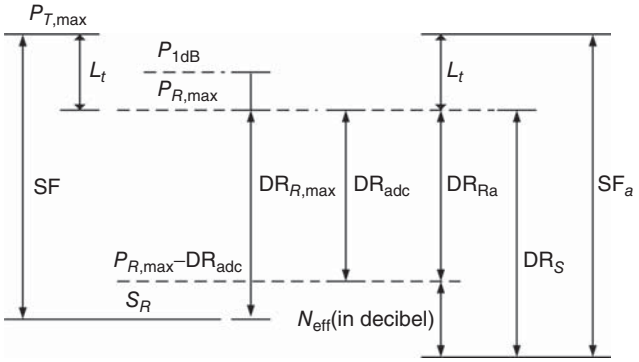


Figure 16.6. Graphical analysis of the system performance factor and dynamic range when $DR_{\text{adc}} \leq DR_{R,\max}$.

This leads to the SF in term of the maximum available dynamic range as

$$\mathrm{SF} = (\mathrm{DR}_{R,\max} + L_t) \quad \text{(dB)} \tag{16.31}$$

The SF represents the maximum performance of the system if the system satisfies the maximum available dynamic range. However, it is important to note that the system also contains analog-to-digital converters (ADC's) for signal processing. Therefore, the ADC's dynamic range, $\mathrm{DR}_{\mathrm{adc}}$, should also be considered in the system evaluation. The ADC's dynamic range can be approximated as [1]

$$\mathrm{DR}_{\mathrm{adc}} \simeq 6N \quad \text{(dB)} \tag{16.32}$$

where N is the number of bits of the ADC. Therefore, the receiver's available dynamic range, $\mathrm{DR}_{\mathrm{Ra}}$, is limited by either the receiver's maximum available dynamic range or the ADC's dynamic range, whichever is narrower. A signal processing gain typically implemented in systems, however, increases the receiver's dynamic range. The system dynamic range, DR_S, can therefore be defined as [3]

$$\mathrm{DR}_S = [\mathrm{DR}_{\mathrm{Ra}} + 10\log(N_{\mathrm{eff}})] \quad \text{(dB)} \tag{16.33}$$

where N_{eff} represents the signal-processing gain achieved in the system. As a result, the actual SF can be obtained as

$$\mathrm{SF}_a = (\mathrm{DR}_S + L_t) \quad \text{(dB)} \tag{16.34}$$

The range of the system can be estimated more accurately using the system equation incorporating the actual SF obtained by (16.34).

16.1.6 Power

16.1.6.1 Peak Power. We first consider continuous-wave (CW) signals and assume a typical CW sinusoidal signal $v(t) = V_m \cos\ \omega t$ for $t \in]-\infty, \infty[$ as shown in Figure 16.7. The (instantaneous) power across the resistor is given as

$$p(t) = v(t)i(t) = V_m I_m cos^2\ \omega t \tag{16.35}$$

The "peak power" for CW signals is equal to the "average power" or the power averaged over a cycle, and hence is obtained as

$$P_{\mathrm{pk}} = \frac{1}{T}\int_0^T p(t)dt = \frac{1}{T}\int_0^T V_m I_m \cos^2\ \omega t\ \ dt = \frac{1}{2}V_m I_m \tag{16.36}$$

where T is the period. This power is the maximum root-mean-square (rms) power with $V_m/\sqrt{2}$ and $I_m/\sqrt{2}$ being the rms voltage and current, respectively. It is noted that the powers (P_t and P_r) used in the system equations (16.12) and (16.13) are peak power.

Now we consider pulse signals such as a rectangular pulse train as shown in Figure 16.8. The duty cycle of a pulse signal is defined as τ/T and represents the relative time that the source (or transmitter) is on during

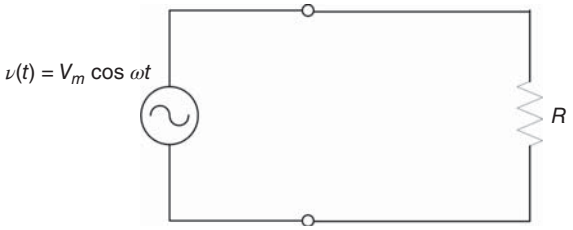


Figure 16.7. A simple circuit with CW sinusoidal signal.

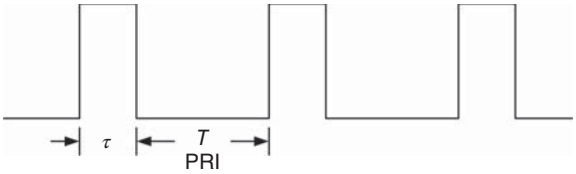


Figure 16.8. Rectangular pulse. τ is the pulse width, length, or duration; T is the pulse period or pulse repetition interval (PRI); and PRF = $1/T$ is known as the pulse repetition frequency representing the number of pulses per second.

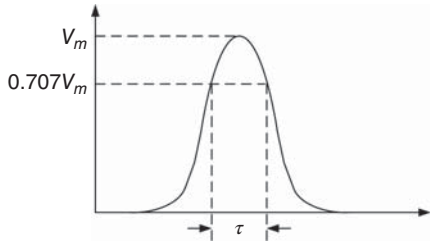


Figure 16.9. A practical pulse with its normally defined pulse width.

one period of the pulse. The peak power of a pulse signal is equal to the average power or the power averaged over the pulse and is given as

$$P_{\text{pk}} = \frac{1}{\tau}\int_0^{\tau} p(t)dt \tag{16.37}$$

This power is normally defined as the "pulse power." This peak power is also the maximum rms power during the time the pulse is on. For (practical) non-perfect rectangular pulses or pulses of different shapes, such as a Gaussian pulse, the pulse width (τ) is typically defined between half-power points as seen in Figure 16.9.

16.1.6.2 Average Power. Average power is the power averaged over a cycle or period. For CW signals, the peak power is transmitted at all time and hence also represents the average power. For pulse signals, however, the peak power is transmitted only during the pulse, hence making the average power less than the peak power. The pulse energy is given as

$$E_p = P_{\text{pk}}\tau \tag{16.38}$$

which is the energy in each transmitting pulse. The average power can be obtained by averaging the pulse energy over the entire period as

$$P_{\text{avg}} = P_{\text{pk}}\frac{\tau}{T} = P_{\text{pk}}(\text{Duty cycle}) \tag{16.39}$$

where

$$\text{Duty cycle} \equiv \frac{\tau}{T} \tag{16.40}$$

The duty cycle of pulse signals is always less than 1 and hence the average power is always less than the peak power. For CW sinusoidal signals, $\tau = T$ which results in 100% duty cycle and hence equal average and peak power.

16.1.7 Angle and Range Resolution

The ability of a system to distinguish targets that are closely located depends on its resolution. There are two kinds of resolution: angle and range resolution depending on the direction of observation from antenna as illustrated in Figure 16.10. Angle and range resolutions are therefore important in a system's operation. The following equations obtained for the angle and range resolutions are general and applicable to different systems.

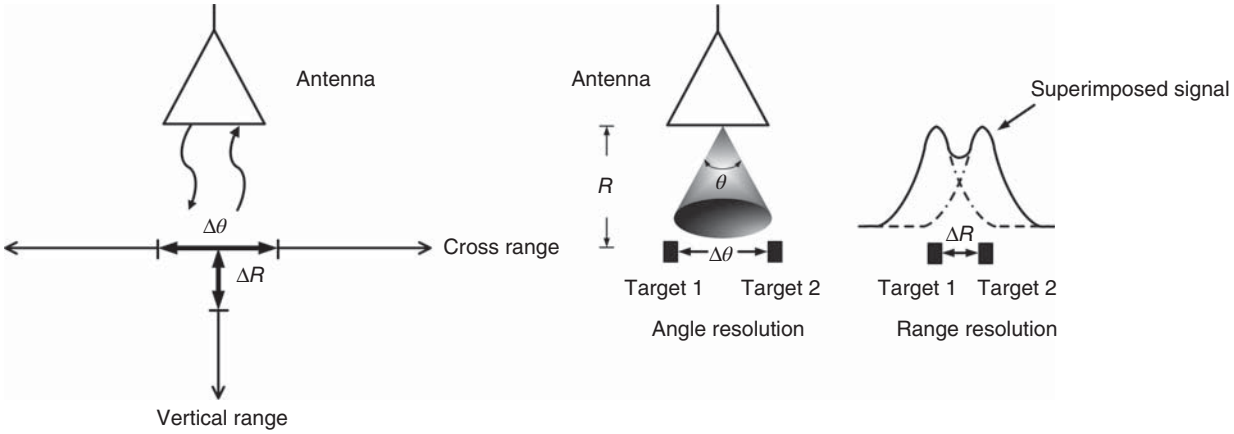


Figure 16.10. Illustration of angle and range resolution of systems. $\Delta\theta$ and ΔR denote the angle and range resolution, respectively.

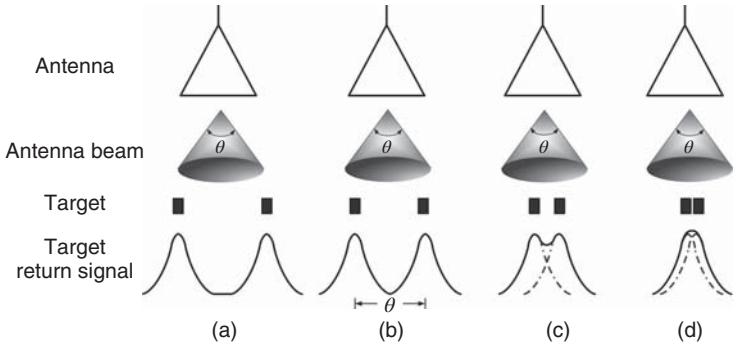


Figure 16.11. Target distinguishing based on angle resolution: (a) targets easily resolved when (angle) separation $> \theta$, (b) targets resolved when (angle) separation equal to θ, (c) targets difficultly resolved when (angle) separation $< \theta$, and (d) targets unresolved when (angle) separation $\ll \theta$. θ is the antenna's beamwidth.

16.1.7.1 Angle Resolution. Angle resolution, also known as cross-range, horizontal, lateral, or azimuth resolution, indicates the minimum angle two targets at the same range must be spaced apart in order to be distinguished. As an example, a system with an angle resolution of 5° can distinguish targets at the same range whose separation is larger than 5°. Angle resolution, as can be expected, is determined by the beamwidth of the antenna and improves as the antenna's beamwidth gets narrower. Figure 16.11 illustrates the angle-resolution phenomenon.

In the antenna's far-field region, occurring when the range between the antenna and targets is sufficiently large with respect to the operating wavelength, we can approximate the angle resolution as

$$\Delta R \simeq R\theta \tag{16.41}$$

where θ is the antenna's beamwidth (in radian) and R is the target's range in the far-field region. For aperture antennas, R can be approximated as

$$R \geq \frac{2D^2}{\lambda} \tag{16.42}$$

where λ is the operating wavelength and D is the maximum dimension of the antenna. The beamwidth of antennas depend on the ratio of the antenna dimensions to the operating wavelength. For a given antenna size, the antenna's beamwidth is proportional to the wavelength or inversely proportional to the frequency. High-frequency system therefore can focus energy into sharp beams, resulting in fine resolution and accurate determination of targets. Angle resolution improves as the antenna gets larger electrically.

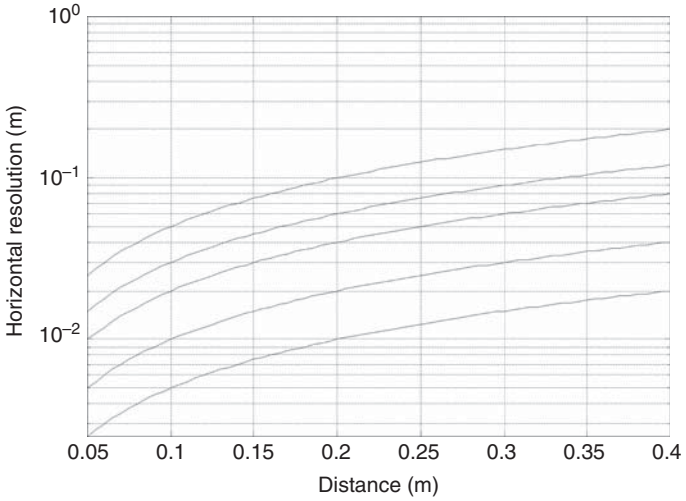


Figure 16.12. Angle (horizontal) resolution versus range.

Figure 16.12 shows the calculated angle resolution versus the (far-field) range for varying beamwidth. In order to achieve angle resolution in centimeter, frequencies in the range of Ka-band (26.5–40 GHz) need to be used. For instance, using a standard Ka-band waveguide horn antenna having about 0.26-radian 3-dB beamwidth, we can estimate the angle resolution as 0.031 m for a target located at 0.12 m from the system.

16.1.7.2 Range Resolution. Range resolution determines how close two targets in different ranges can be spaced in order to be distinguished. For instance, a system having a range resolution of 10 cm can only distinguish targets separated by at least 10 cm in range. Figure 16.13 illustrates the range-resolution phenomenon.

Consider pulse signals returned from targets, the narrow the received pulses, the better targets can be resolved. For the targets to be resolved, they must be separated in time by an amount equal to the pulse duration τ of the received pulse. This is equivalent to a range difference of

$$\Delta R = \frac{v\tau}{2} \tag{16.43}$$

where v is the speed of light in the propagating medium, which dictates the range resolution. In practical pulse systems, the pulse received may not have the waveform as that transmitted. For instance, depending on the transmit and receive antennas, the received pulse can be the first derivative of the transmitted pulse. Furthermore, the pulse width of the received pulse is widened as compared to that of the transmitted pulse. However, for well-designed antennas and receiver's components, the pulse-stretching effect on the received pulse is small and may be neglected, and the width of the received pulse can be assumed approximately equal to that of the transmitted pulse. In pulse-compression systems, such as stepped-frequency radar, the received pulse is processed to produce a pulse having pulse width less than that of the transmitted pulse.

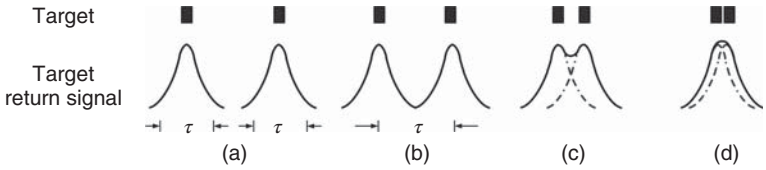


Figure 16.13. Target distinguishing based on range resolution: (a) targets easily resolved when separated by $> \tau$ (pulse width), (b) targets resolvable when separated by τ, (c) targets difficultly resolved when separated by $< \tau$, and (d) targets unresolved when separated by $\ll \tau$. The "τ" indicated in the figure is loosely defined – it is used only for illustration.

The (absolute) bandwidth of a received pulse with a 4-dB pulse width of τ can be approximated as

$$B \cong \frac{1}{\tau} \tag{16.44}$$

It is noted this bandwidth is the same as the system's RF operating bandwidth. Substituting (16.44) into (16.43) gives the range resolution

$$\Delta R \cong \frac{v}{2B} \tag{16.45}$$

For instance, to process a pulse of 1 ns, we need an approximate 3-dB bandwidth of 1 GHz, and the resultant range resolution is 0.15 m assuming air is the transmission medium. As can be seen, the range resolution is inversely proportional to the absolute RF operating bandwidth – the wider the bandwidth or the narrower the pulse, the better the range resolution. A system operating at higher frequencies has better range resolution as compared to that operating at lower frequencies for the same fractional bandwidth. It is noted that a system operating at high frequencies would be relatively easier to be designed than one operating at low frequencies for the same absolute bandwidth due to the reduced fractional bandwidth, which affects the design of system's components including antennas, assuming other considerations, such as circuit design difficulty or frequency effects on circuit design, being equal. Moreover, an important result that can be inferred from the range-resolution formula (16.45) is that the waveform of a signal affects the signal's bandwidth. Proper choice of transmitting waveforms is therefore crucial and, in fact, one of the important design criteria for RF engineers.

Figure 16.14 shows the calculated range resolution as a function of the bandwidth for different relative dielectric constants (ε_r) of the transmission medium.

16.1.8 Range Accuracy

Range accuracy dictates how accurate one can measure the range and is different from range resolution. The rms range error can be approximately derived as

$$\delta R \simeq \frac{\Delta R}{\sqrt{2(\mathrm{S/N})}} = \frac{v}{2B\sqrt{2(\mathrm{S/N})}} \tag{16.46}$$

For pulse systems, the bandwidth is given by (16.44) while, for CW systems, $B = f_H - f_L$, where f_L and f_H are the lower and upper frequencies of the operating band, respectively. As can be seen, the range accuracy also depends on the RF bandwidth of the system; as the bandwidth is increased, the range error is reduced.

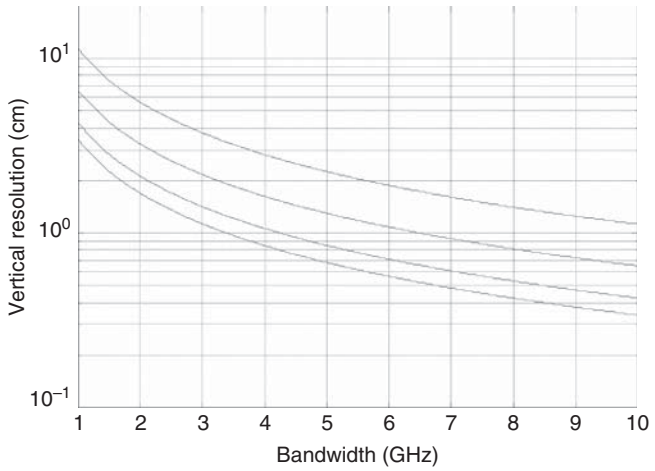


Figure 16.14. Range resolution versus bandwidth.

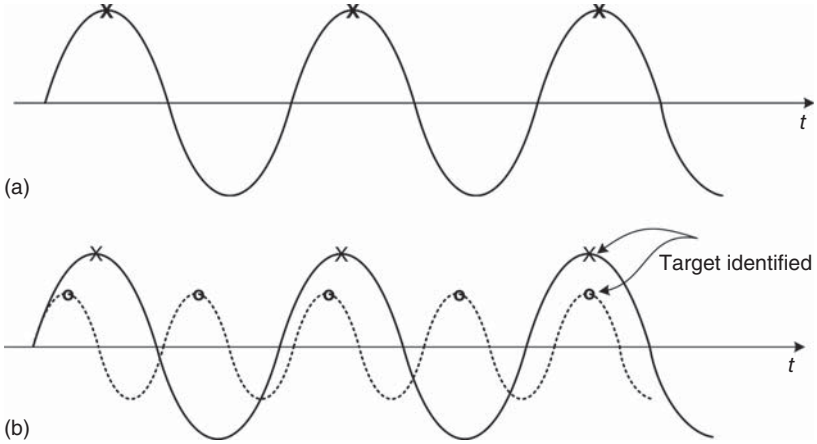


Figure 16.15. (a,b) Illustration of dependence of δR on bandwidth with return signals for a single transmitting frequency (a) and for two transmitting frequencies f_1 and f_2. The "x" and "o" indicate possible target locations corresponding to a phase measurement of 90°. When the "x" and "o" appear close to each other, a corresponding target can be identified.

Furthermore, the range error is inversely proportional to $\sqrt{\mathrm{S/N}}$, indicating that more noise produces less accuracy in range, which is expected, as the higher the noise, the less perfect the pulse or waveform shape is. Figure 16.15 illustrates the dependence of the range error δR on bandwidth. As compared to the return signal corresponding to a single frequency, the double-frequency return signals allow a target to be more accurately determined. As can be induced, when multiple frequencies are transmitted, or the operating bandwidth is large, the accuracy and un-ambiguity of the position of target is improved.

The angle accuracy or cross-range accuracy is different from the angle resolution and can be derived approximately as

$$\delta\theta \simeq \frac{\Delta\theta}{3\sqrt{\mathrm{S/N}}} = \frac{R\theta}{3\sqrt{\mathrm{S/N}}} \tag{16.47}$$

where θ is the 3-dB beamwidth of antenna.

16.2 SYSTEM TYPE

A system type can be classified depending on four general criteria: (i) waveform of transmitting signal – for example, pulse, FMCW, and pulse compression, (ii) location of system such as air-born, ground-based, spaced-based, and ship-based, (iii) function of system, such as communication, detection, search, and tracking, and (iv) application of system – for example, point-to-point communications, traffic control, collision-avoidance, fire control, weapon guidance, air-defense, surveillance, navigation, and detection. In the following section, we will use radar or, in general, a sensing system as an example to discuss systems. Specifically, pulse and FMCW systems, two of the most widely used systems, will be addressed. It is important to note, however, that the system principles and operations to be discussed are not limited to a certain system or application. Communication or other RF systems can be implemented based on similar system and component architectures with proper modifications, such as using different modulation schemes.

16.2.1 Pulse System

In general, pulse and other systems such as FMCW are based on similar architectures, such as that described for FMCW system in Section 16.2.2. The main difference lies in the waveforms produced and transmitted by the transmitters. For instance, a pulse system transmits a train of pulses, whereas an FMCW system transmits a CW signal. In this section, however, we will describe a particular pulse system. This system, also known as impulse system, is distinguished from conventional pulse systems due to its impulse-type signals that have

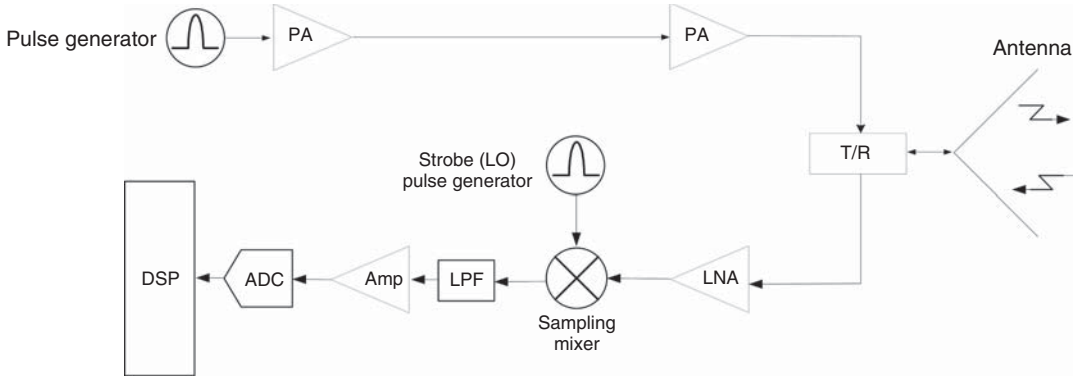


Figure 16.16. Block diagram of a monostatic impulse system.

very wide instantaneous bandwidth approaching a decade. It is relatively simple and low-cost as compared to a conventional pulse system.

Figure 16.16 shows a block diagram of a monostatic impulse system with main components. Block diagram for a bistatic system is similar, except that the receiver and its antenna and the transmitter and its antenna are located at different locations. In operation, the pulse generator generates a pulse train, which is amplified by the power amplifier (PA) and transmitted toward a destination or targets via the antenna and the transmit/receive (T/R) switch, also known as duplexer. The T/R switch switches the antenna back and forth between the transmitter (TX) and receiver (RX). A circulator may be used in place of the duplexer or in addition to the T/R switch to enhance the isolation between the TX and RX. When a pulse is transmitted, the system's clock begins to count time to facilitate the measurement of the time elapsed between the transmitted and received pulses. The return signal from the target or the transmitting signal from other system enters the antenna and goes through the duplexer. This signal is amplified by the LNA and down-converted directly to baseband by the sampling mixer. The down-conversion is accomplished in one stage without any intermediate stage, thus greatly reducing the system complexity. This direct down-conversion is typically done by a repetitive sampling process in which the actual RF signal at different instants is sampled using a "sampling pulse" or "sampling strobe." The baseband signal is filtered, amplified and converted into a digital format by the ADC. Signal processing through the digital signal processor (DSP) is then used to process both desired and undesired signals (such as clutter or interference.)

As for other systems, the waveform of transmitting pulses is important in impulse system design and operation. Different types of pulse waveforms can be used for impulse radar or communication system, depending on requirements for output peak power, bandwidth, and system's components (e.g., antenna) through which the pulse is transmitted – for instance, Gaussian or monopolar impulse, single-cycle or monocycle pulse, and multi-cycle pulse. Step-function and impulse contain DC and a large portion of low-frequency spectral components, which cannot be transmitted through practical antennas. Monocycle pulse, on the other hand, has no DC component and band-limited characteristic, facilitating its transmission using a practical antenna. Furthermore, using monocycle pulse facilitates the design of other components including antenna in the system. As compared to multi-cycle pulse, monocycle pulse has a wider bandwidth, which results in better range resolution. Figure 16.17 shows typical transmitting pulse trains in impulse systems. Figure 16.18 shows examples of the monopolar impulse and monocycle pulse along with their frequency spectrums.

The impulse system has simple system architecture and can produce high or low transmitting power with a relatively simple transmitter circuit. It is suitable for both short- and long-range applications. Its signal-processing cost is also relatively inexpensive due to the fact that no intermediate stage is needed for signal detection. Impulse systems typically use narrow pulses, which occupy very wide bandwidths, and hence are particularly suitable for high-resolution applications such as target identification and imaging. Impulse system can also be used for communications. Particularly, for low-power applications, the total transmitted power is spread over an extremely wide range of frequencies, and hence the power spectral density is extremely low. This very-small power spectral density effectively results in extremely small interference

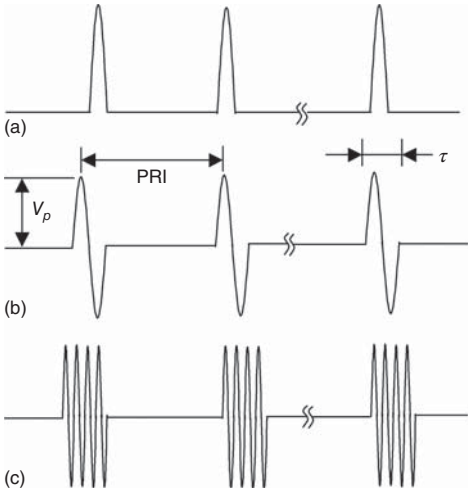


Figure 16.17. Waveforms of typical transmitting signals in impulse systems: impulse (a), monocycle pulse (b), and multicycle pulse (c).

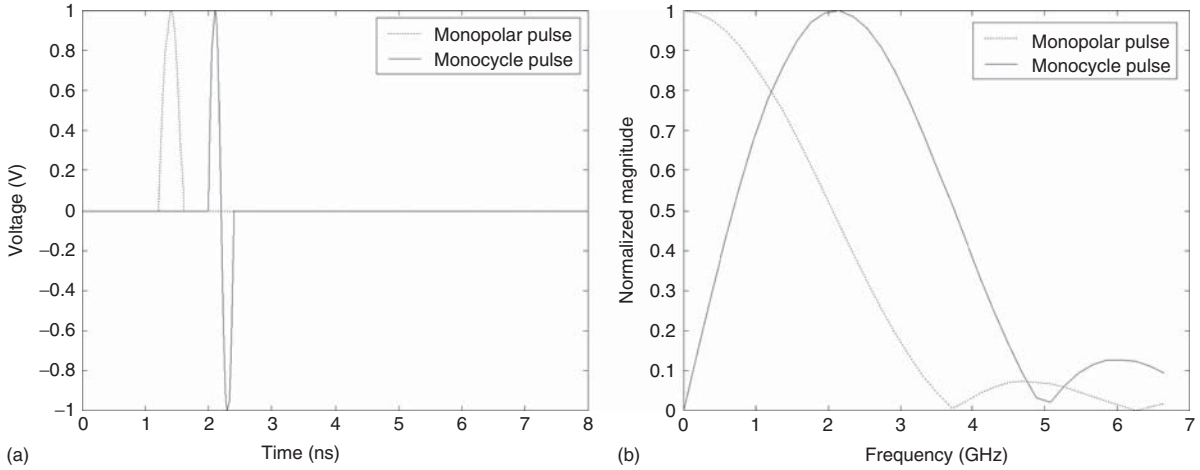


Figure 16.18. Waveforms (a) and spectrums (b) of the monopolar and monocycle pulse with 400-ps pulse duration. They are modeled as one-half and full cycle of a sinusoidal signal, respectively.

to other radio signals while maintains excellent immunity to interference from these signals. Low-power impulse systems can therefore work within frequencies already allocated for other radio services, thus helping to maximize this dwindling resource. Therefore, the low-power impulse technique is attractive for realizing high-data-rate, short-range communications, sensors, and military radar requiring relatively low emission power levels. The extremely wide bandwidth of an impulse system may make it difficult to design components, for example, antennas. For instance, a Gaussian pulse of 0.33 ns would require a bandwidth from DC up to about 10 GHz. A pulse system also needs a high peak power for sufficient range measurement or communications. The inherent wide bandwidth useful for certain applications also comes with a price. Consider the popular monocycle pulse as an example; its center frequency and bandwidth are determined from its pulse width and they are inseparable, and so, for instance, in order to operate at high frequencies for some advantages such as smaller antenna size, a narrow pulse is needed which, in turn, causes a very bandwidth, making the design more difficult.

16.2.2 FMCW System

Figure 16.19 shows a block diagram of a (monostatic) FMCW system that can also be used for other systems such as pulse system by changing the way the transmitting signal is modulated. In a pulse system, for instance,

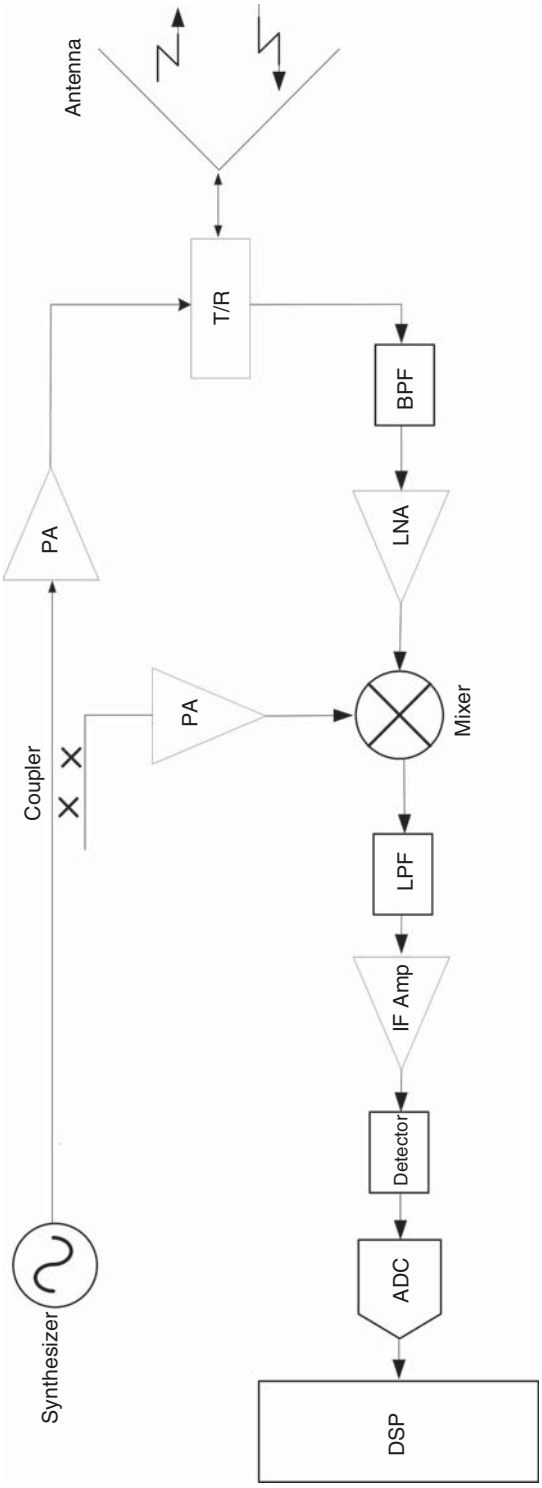


Figure 16.19. Block diagram of FMCW systems.

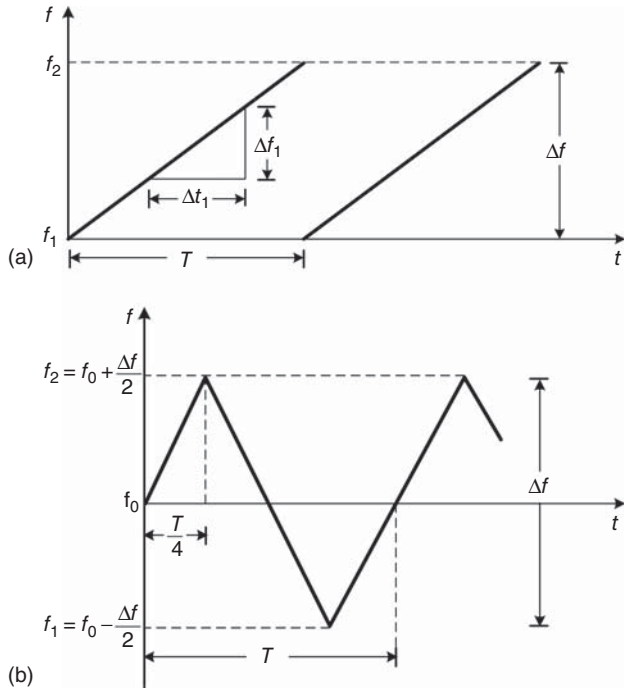


Figure 16.20. Frequency-modulated CW signals: Linear frequency modulation (a) and triangular frequency modulation (b).

the transmitting signal is modulated such that a pulse is transmitted periodically over a certain time. Various frequency-modulation schemes can be used for FMCW systems. However, conventional FMCW systems usually employ a linear and triangular frequency-modulated CW signal as shown in Figure 16.20. These frequency modulations are particularly useful for accounting the Doppler shift due to moving objects. Linear frequency modulation is simpler and can be used for stationary targets. The period T, over which the linear modulation is carried out, is much larger than twice the traveling time between the system and target (i.e., round-trip travel time) to allow for range measurement to be made. The period T also specifies the time the transmitter starts a new cycle of transmission.

The rate of frequency sweeping or change (in hertz per second) for the linear frequency modulation is given as

$$m = \frac{\Delta f_1}{\Delta t_1} = \frac{\Delta f}{T} = f_m \Delta f \tag{16.48}$$

where $T = 1/f_m$. For the triangular frequency modulation, the rate of frequency change can be obtained as

$$m = 2f_m \Delta f \tag{16.49}$$

In operation, the synthesizer generates a frequency-modulated signal. This signal is amplified by the PAs, travels through the T/R switch, and is radiated by the antenna into space. A small portion of the FMCW signal is also injected into the mixer to act as the frequency reference. This signal functions as a local oscillator (LO) signal driving the mixer. The signals returned from targets or transmitted from other systems go through the antenna, T/R switch, filtered by a BPF, amplified in a LNA, mixed with the LO signal in a mixer to down-converted to an IF signal. This IF signal is amplified by an IF amplifier and goes through a detector to produce a baseband or video signal. The video signal is then amplified by a video amplifier and digitized by an ADC. The digital signal is finally processed by a DSP.

If the video bandwidth is approximately equal to or greater than one-half of the IF bandwidth, there is no adverse effect on the signal detectability. The detector may operate as amplitude or phase detector. An amplitude detector can only detect the amplitude of the signal, while a phase detector is used when phase

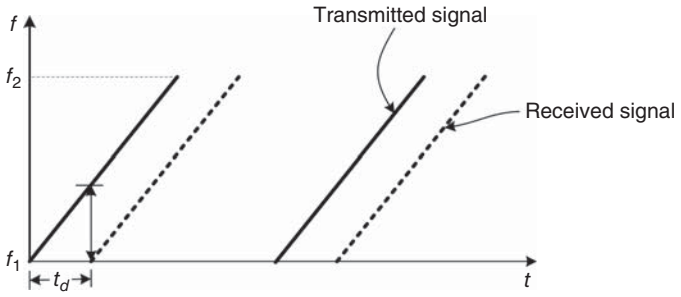


Figure 16.21. Illustration of transmitted and received signals.

information is needed. Phase detection is usually more complicated to implement than amplitude detection. An example of a phase detector is quadrature mixer (which can also be used for amplitude measurement) to be described later. For example, a phase detector is needed in "Doppler systems" to extract the Doppler frequency by comparing the received signal with a reference signal at the transmitted frequency. Proper signal processing increases the desired and suppressing the undesired signals and hence their ratio. A threshold level is typically established to allow the detection decision to be made. If the signal level (desired plus undesired) crosses the threshold, a target is present. Data processing is performed after the detection decision is made. It reveals the target information, such as location, identification, and classification of the target.

The return signal arrives at the system after a "round-trip travel time." This time delay effectively results in a frequency difference, commonly known as the "beat frequency," at a given time as illustrated in Figure 16.21. We can express the time delay between the transmitted and received signals as

$$t_d = \frac{f_d}{m} \tag{16.50}$$

where f_d is the frequency difference accounting for the relative time delay. As the frequency sweeping speed (meter) is known and f_d can be measured, t_d can be determined. Once t_d is determined, the range can be immediately obtained as

$$R = \frac{vt_d}{2} \tag{16.51}$$

where v is the speed of light in the propagating medium. The modulation period (T) sets the maximum range that can be measured:

$$R_{\max} = \frac{vT}{2} \tag{16.52}$$

It is important to note that, for FWCW systems, the rate of sweeping frequency (m) should be carefully observed to obtain a satisfactorily accurate range of the target. However, it is quite difficult to achieve this specification over a wide band, due to the nonlinearity of the synthesizer, particularly when a voltage-controlled oscillator (VCO) is used in its place. Moreover, the wide bandwidth of the source degrades the receiver's sensitivity, which results in reduced range. This drawback limits the FMCW system in some applications that need a greater degree of accuracy. We assume the target is stationary and hence no Doppler frequency shift is produced. The return signal's frequency is different from that of the transmitted signal by f_d, and so the frequency of the down-converted signal produced by the mixer, namely IF frequency, is indeed equal to f_d. This frequency is thus often referred to as the "beat frequency." For multiple targets, each target would produce one frequency difference corresponding to one round-trip travel time and hence one range as illustrated in Figure 16.22.

The foregoing discussion assumes linear or triangular frequency modulation, where the linearity of frequency over a sweeping period is crucial. This linearity, however, is difficult to maintain very well due to the nonlinearity in the VCO producing the signal, particularly over a wide bandwidth and/or at high RF frequencies. When a system is used for detecting single target, for example, an altimeter, it is not necessary to use

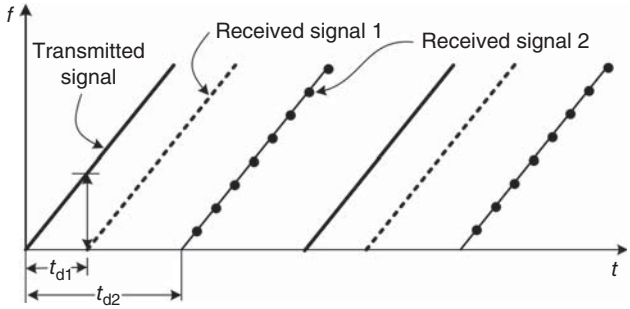


Figure 16.22. Time delay for multiple targets.

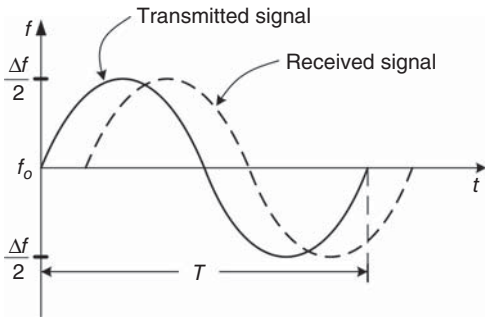


Figure 16.23. Sinusoidal frequency modulation.

linear modulation. It is easier to use frequency modulation as illustrated in Figure 16.23, which is perhaps the simplest modulation. In this modulation scheme, however, the "beat frequency" is not constant over the modulation cycle as compared to the linear modulation. The measured beat frequencies are typically averaged over a modulation cycle. The accuracy of range measurement depends on the rate of the frequency sweeping and the measurement accuracy of the beat frequency. The frequency measurement accuracy is proportional to the time-period T over which the measurement is made. For large frequency-sweeping rate, the beat frequency is large for a given range (and hence time delay), which results in more accurate measurement. For a small range, the time delay is small. To overcome this small time delay and achieve accurate measurement, the frequency needs to be changed sufficiently fast while still producing $T \gg t_d$ as required for ranging.

In order to determine the target's velocity, only a single-frequency CW FMCW is needed, which is basically used to measure the Doppler frequency shift. Single-frequency FMCW, however, is not suitable for range measurement. To determine the range of a stationary or moving target, the system needs to apply a "timing mark" to the CW signal to determine the time of the transmission and the time of the reception, from which a range can be determined. The timing mark can be achieved by modulating the CW signal in frequency, amplitude or phase. Frequency modulation, however, is typically used to obtain the timing mark needed for ranging. The frequency-modulation period (T) is less than or equal to the time duration over which the target is illuminated. This time period depends on the antenna's sweeping rate and beamwidth. It is noted that the speed of the platform carrying the system (e.g., vehicle) is relatively small as compared to the speed of signal propagation, and hence the system can be considered "stationary."

Depending on applications, an FMCW CW system may have some advantages as compared to a pulse system. For instance, it has a much narrower (absolute) bandwidth, which leads to ease in component design. It is possible to use high operating frequencies without increasing the fractional bandwidth, thus increasing the absolute bandwidth resulting in better range resolution while achieving better angle resolution due to reduced beamwidth at high frequencies. In CW systems, the duty cycle is 100% and the peak power is equal to the average power; CW thus requires lower peak power.

16.2.3 Receiver Architectures

Receivers are typically classified into three different types: heterodyne, superheterodyne, and homodyne (or direct conversion.) The heterodyne receiver uses different RF and LO frequencies to obtain IF; it is not an attractive architecture and, in fact, is part of the more widely used superheterodyne architecture. Superheterodyne can be considered a double-stage heterodyne approach, in which it employs one heterodyne stage to convert an RF signal to an IF signal and then another heterodyne stage to further convert an IF signal to a baseband signal. Homodyne uses the same frequency source for RF and LO signals, which essentially result in IF signals in baseband. As the heterodyne is indeed included within the superheterodyne, we will only discuss superheterodyne and homodyne receiver architectures in this section.

16.2.3.1 Superheterodyne Receiver. Figure 16.24 shows a block diagram of a system based on the superheterodyne receiver architecture. The operation of this system is similar to that described for the FMCW system in Section 16.2.2. Particularly, there are three mixers in this superheterodyne configuration. Mixer 1 is used to down-convert part of the transmitted RF signal at frequency f_T, feeding to the mixer via a directional coupler, by mixing with an LO signal at frequency f_{IF} coming from a stable oscillator (STALO). The frequency f_{IF} is relatively small compared to the RF frequency f_{RF}. The signals at the output of Mixer 1 are at frequency f_T and two sideband frequencies $f_T \pm f_{\text{IF}}$. The signals at f_T and the upper sideband $f_T + f_{\text{IF}}$ are suppressed by the low-pass filter. The lower-sideband signal at $f_T - f_{\text{IF}}$ acts as the LO signal for Mixer 2 to down-convert the received signal at frequency f_R coming from the antenna to a signal at $f_{\text{IF}} + f_b$, where $f_b = f_R - f_T$. This signal is then down-converted directly into in-phase (I) and quadrature-phase (Q) signals in base-band via Mixer 3, which is an I–Q or quadrature mixer (to be described later), through mixing with the signal at frequency f_{IF} coming from the STALO. The two I and Q baseband signals are digitized with ADCs and processed via a DSP to produce desired outputs. It is particularly noted that the signal down-converted by Mixer 2 has a narrow bandwidth around the frequency f_{IF},[1] regardless of the frequency range of the received signal; therefore, the quadrature mixer only needs to operate over a narrow bandwidth, which facilitates its design.

16.2.3.2 Homodyne Receiver. Figure 16.25 shows a system employing a homodyne receiver. The homodyne receiver is also known as direct-conversion or zero-IF receiver. It can be considered a superheterodyne receiver with "zero IF." The down-conversion to baseband signals is accomplished directly by mixing the

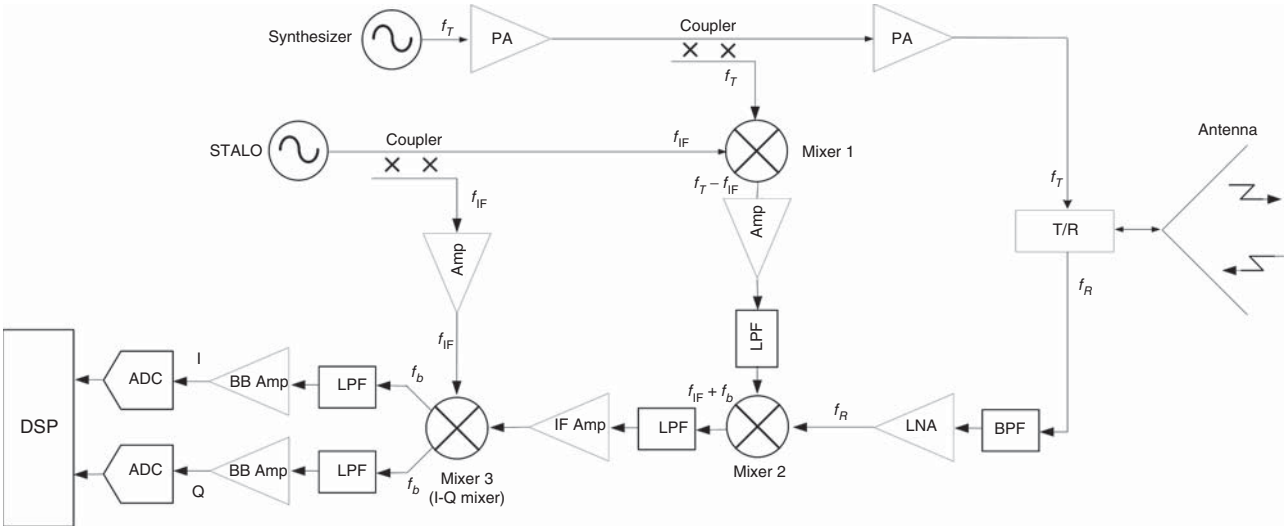


Figure 16.24. Block diagram of a system implementing a superheterodyne architecture. The frequency notations are denoted in the figure. BB, baseband.

[1]In applications where f_b is not generated, such as communications, the down-converted signal has only a single frequency f_{IF}.

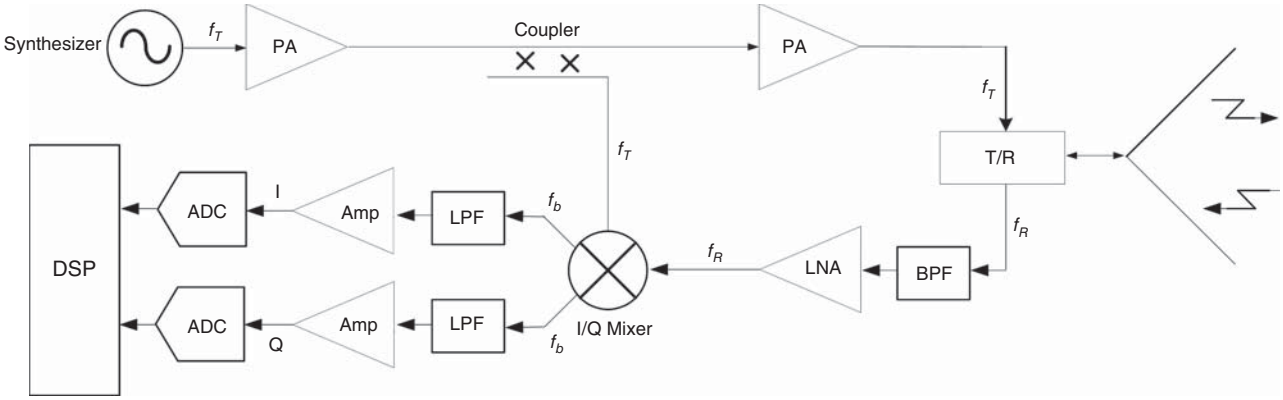


Figure 16.25. Block diagram of a system with a homodyne receiver including frequency indication.

transmitted signal with the receive signal in a single I–Q mixer. The I and Q signals are converted into digital signals through ADCs and processed by the DSP.

As can been seen, the homodyne architecture does not need intermediate stages and hence is much simpler than its superheterodyne counterpart. It, however, is less sensitive than the superheterodyne due to the inherent $1/f$ (or Flicker) noise, which occurs in semiconductor devices at low frequencies and is inversely proportional to frequency (f). At low frequencies, such as those in the audio or video range, the $1/f$ noise contributed by the mixer can be large, which potentially reduces the S/N ratio and hence the receiver's sensitivity. In sensing applications, the frequency of the homodyne receiver's output signal (f_b) is proportional to the distance between the antenna and a target, and so, for short ranges, the receiver's output frequency can be very low and the receiver would suffer from the $1/f$ noise. In short ranges, however, the return signal is enhanced and this may compensate for the increased noise. When a homodyne system is used for moving object detection, the Doppler frequency resulting from movement is usually in low-frequency range, and the $1/f$ noise has more pronounced effect on the measurement. The homodyne configuration is not desirable for optimum sensitivity. On the other hand, the superheterodyne architecture may avoid the $1/f$ noise problem by using a sufficiently high frequency for the receiver's output signal to achieve small $1/f$ noise relative to the other receiver's noises.

It is note that, while the baseband amplifier in the homodyne and superheterodyne amplifies both the $1/f$ noise and the output signal, the IF amplifier in the superheterodyne only amplifies the IF signal, which is at $f_{\text{IF}} + f_b$, not the $1/f$ noise. This increased IF-signal level, and hence the down-converted baseband signal, helps increase the receiver's sensitivity. Depending on specific configurations, in general, some superheterodyne receivers may have sensitivity as high as 30 dB more than a homodyne receiver, which is significant in certain applications. As mentioned earlier, in a superheterodyne system, the I–Q mixer is used after the down-conversion takes place at the other preceding mixers and thus typically works over a narrow band around f_{IF}. This narrow-band operation leads to ease in achieving a nearly constant balance between the I and Q channels, which is desired in the I–Q mixer operation. On the other hand, as the homodyne system down-converts the received RF signal directly into baseband signals by using an I–Q mixer, it requires a wideband I–Q mixer for system operating over a wide bandwidth. This leads to potential problems in the balance between the I and Q channels over the band of interest since the 90° phase difference needed between the I and Q channels would not be constant across the band, particularly for very wide bandwidths. Consequently, despite its complexity, the superheterodyne system is still preferred over the homodyne system. However, in applications requiring simplicity, low cost with less stringent requirements, the homodyne architecture is more preferable than its superheterodyne counterpart, provided that its performance degradation is tolerable.

16.2.3.3 Receiver Bandwidth. Receiver's bandwidth (B) is an important parameter in the receiver design. We want the bandwidth to be wide enough to pass the desired RF signals – for instance, those at frequency ($f_T + f_b$), where f_T and f_b are the respective transmitting and beat frequency, for target's range measurement,

in which f_b varies over a frequency range due to different target ranges. The baseband amplifier in the homodyne and superheterodyne systems or the IF amplifier in the latter must be sufficiently wide to pass the baseband signal over a range of frequencies caused by changes in the target distance. However, if the bandwidth is too wide, it will produce increased noise and hence reduced receiver sensitivity. If the frequency f_b resulting from a target range can be estimated – for instance, that encountered in measuring typical speeds of traveling vehicles – a narrow bandwidth can be used in the amplifier or filter to pass the desired signal and reject noise.

We now consider CW transmitting and receiving signals at frequency f_T and f_R, respectively, operating over an infinite duration. At any time over an infinite duration, there is only a single frequency, regardless whether it is the transmitting or receiving frequency, and hence no bandwidth except that due to different target ranges. Assume the frequency f_b arising due to the target range is sufficiently small, the resultant bandwidth can be considered very small. Although the return signal is continuous and occurs over an infinite duration, its portion corresponding to a target occurs over a finite period. This finite duration is equal to the time the signal illuminating or staying on the target and is needed for target characterization. It is given as

$$\tau = \frac{\theta}{\theta'} \tag{16.53}$$

where θ (degree) and θ' (degree per second) are the antenna's beamwidth and its scanning rate, respectively. The finite-duration signal is illustrated in Figure 16.26, which is in fact similar to a pulse of duration τ received in a pulse system. The received voltage can be expressed as

$$v_R(t) = V_R \sin(2\pi f_R t) \tag{16.54}$$

where $t \in [-\tau/2, \tau/2]$ and V_R is the voltage amplitude. Taking the Fourier transform of (16.54) gives

$$V_R(f) = V_R \int_{-\tau/2}^{\tau/2} \sin(2\pi f_R t) e^{-j2\pi ft} dt \tag{16.55}$$

which leads to

$$\frac{V_R(f)}{V_R} = \frac{1}{2j}\left[\frac{e^{j\pi\tau(f-f_R)} - e^{-j\pi\tau(f-f_R)}}{2j\pi(f_R - f)} - \frac{e^{j\pi\tau(f+f_R)} - e^{-j\pi\tau(f+f_R)}}{2j\pi(f_R + f)}\right] \tag{16.56}$$

For f near f_T (i.e., over a narrow bandwidth), we can approximate

$$\frac{V_R(f)}{V_R} \simeq \frac{\tau}{2}\left|\frac{\sin\ \pi\tau\,(f_R - f)}{\pi\tau(f_R - f)}\right| \tag{16.57}$$

whose spectrum is sketched in Figure 16.27. At 3-dB points corresponding to $|V_R(f)/V_R| = 1/2$, the bandwidth is determined as

$$\Delta f_{3\text{dB}} = \frac{0.886}{\tau} \tag{16.58}$$

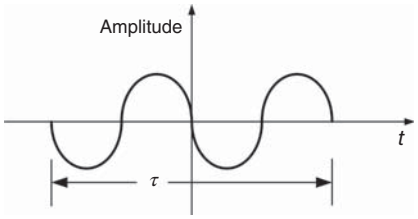


Figure 16.26. Sinusoidal signal of finite duration τ and frequency $f_R = f_T - f_b$.

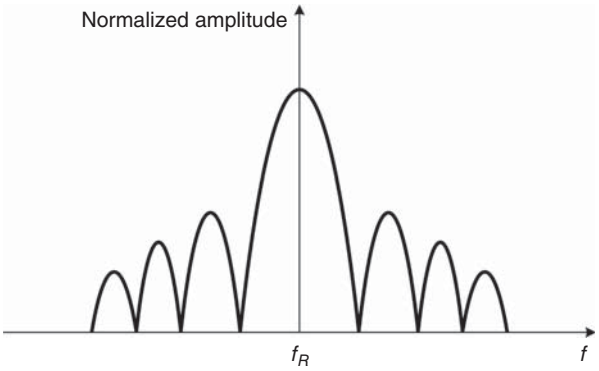


Figure 16.27. Spectrum of a finite-duration sinusoidal signal.

The 4-dB bandwidth can also be obtained as

$$\Delta f_{4\text{dB}} = \frac{1}{\tau} \tag{16.59}$$

The first nulls in the spectrum correspond to

$$\sin\left[\pi\tau(f_R - f)\right] = \sin\ \pi \tag{16.60}$$

from which the null-to-null bandwidth is obtained as

$$\Delta f_{n-n} = 2(f_R - f) = \frac{2}{\tau} \tag{16.61}$$

It is noted the foregoing specific bandwidths of received signals, for example, $\Delta f_{3\text{dB}}$ in (16.58), are caused by the finite illumination time on the target, which effectively broaden the overall receiver's bandwidth. This illumination-induced bandwidth is typically very small and independent of the transmit frequency, and thus has only a slight effect on the receiver noise for a given receiver bandwidth according to (16.17).

The null-to-null bandwidth and the 3- and 4-dB bandwidths are the bandwidth of the receiver in addition to any other bandwidth imposed by the system's frequency plan itself. These bandwidths are also the bandwidth of the baseband amplifier and/or baseband low-pass filter. If the IF range is large due to large target ranges, the bandwidth is also large, and a bank of narrow-band filters may be needed to measure the responses and improve the S/N of the receiver. In this filter bank, each filter operates over a narrow bandwidth and only passes the desired signal and reject noise. All the IF signals are then added before going to a DSP. Assuming the filters' cross-over frequencies are at 3-dB points, the S/N at these frequencies would be 3-dB lower than that at the center frequency.

16.2.3.4 Quadrature Detection. The most critical function of a receiver, whether superheterodyne or homodyne, is to fully recover a (RF) signal. The recovery may be done through quadrature detection accomplished by an I–Q or quadrature mixer, also known as quadrature detector or quadrature demodulator. Quadrature mixer is basically used to measure both the amplitude and phase of a received signal relative to a transmitted signal by determining the two orthogonal components, namely the I and Q components, of the down-converted signal.

Principle of Quadrature Mixers

Block diagrams of quadrature mixers are shown in Figure 16.28, with each consisting of two identical mixers (Mixer I and Mixer Q) of any type (e.g., active double-balanced mixer) in the I and Q channels. A quadrature mixer requires a 90° phase shift in either the LO or RF signal path. However, practical quadrature mixer structures usually employ a 90° phase shift in the LO path (Figure 16.28(a)) due to a narrow LO bandwidth

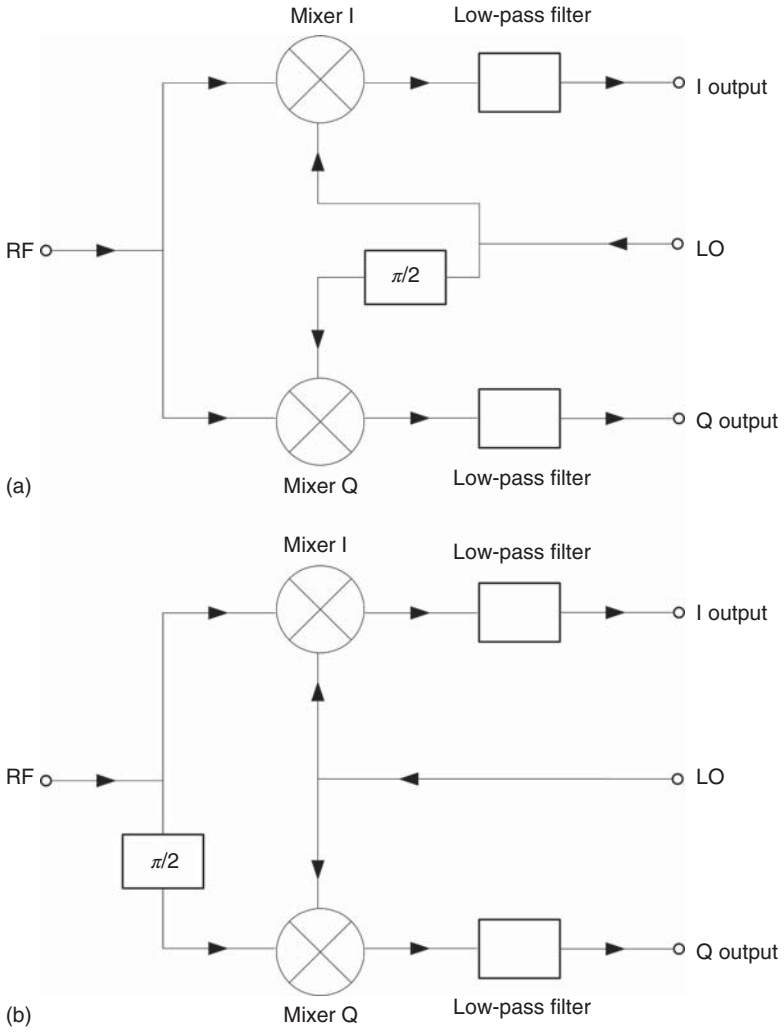


Figure 16.28. Block diagrams of quadrature mixers with 90° phase shift in the LO path and equal phase in the RF path (a) and 90° phase shift in the RF path and equal phase in the LO (b).

typically encountered in mixers which facilitates the design of the phase-shifting network. To illustrate the mixer's operation, we consider the mixer shown in Figure 16.28(a) and assume it is an ideal component with all ideal constituents, hence perfect balance between the I and Q channels, and no leakage between the LO and RF ports of the constituent mixers. The LO signal (for instance, the reference signal from the STALO or the transmitter as seen in the superheterodyne (Figure 16.24) or homodyne (Figure 16.25), respectively) splits into two signals of equal amplitude and 90° phase difference. The 90° out-of-phase is typically achieved using a 90° 3-dB hybrid. The LO signal driving Mixer I can be expressed as

$$v_{\mathrm{LO}}^{I}(t) = V_{\mathrm{LO}} \cos(2\pi ft) \tag{16.62}$$

where f is the frequency of the signal and V_{LO} is $1/\sqrt{2}$ of the amplitude of the LO signal pumping the quadrature mixer (i.e., at the input of the 90° phase shifter), while that arriving at Mixer Q is

$$v_{\mathrm{LO}}^{Q}(t) = V_{\mathrm{LO}} \cos\left(2\pi ft - \frac{\pi}{2}\right) = V_{\mathrm{LO}} \sin(2\pi ft) \tag{16.63}$$

which is delayed by 90° with respect to the LO signal at Mixer I. The RF signal (such as the signal return from target) splits into two signals of equal amplitude and phase using an in-phase splitter such as a power

divider. The RF signals arriving at Mixer I and Mixer Q can be described as

$$v_{\mathrm{RF}}^{I}(t) = v_{\mathrm{RF}}^{Q}(t) = V_{\mathrm{RF}} \cos\left[2\pi ft - \phi(t)\right] \tag{16.64}$$

where V_{RF} is $1/\sqrt{2}$ of the amplitude of the RF signal reaching the mixer (i.e., before splitting to the constituent mixers) and $\phi(t) = 2\pi f t_R - \phi_o$ is the phase of the return signal at the mixer, excluding the initial phase produced by the difference in the electrical lengths between the LO and RF paths leading to the quadrature mixer. Specifically, $\phi(t) = 2\pi f t_R - \phi_o$, assuming the received signal is from a target, where t_R represents the roundtrip time delay in signal propagation to and from the target, and ϕ_o is the initial phase. The contribution from the initial phase to the detected phase of the received signal can be eliminated by inserting a variable phase shifter in the LO or RF path to the mixer to nullify the initial phase.

The RF and LO signals mix at Mixer I to produce the I signal in the I channel:

$$\begin{aligned} v_I(t) &= V_{\mathrm{RF}} \cos[2\pi ft - \phi(t)] \cdot V_{\mathrm{LO}} \cos(2\pi ft) \\ &= \frac{1}{2} V_{\mathrm{RF}} V_{\mathrm{LO}} [\cos(4\pi ft - \phi) + \cos(\phi)] \end{aligned} \tag{16.65}$$

The second term in (16.65) represents the average of DC value and is proportional to the magnitudes of the LO and RF signals and *cosine* of the return signal's phase angle. The first term represents the second harmonic of the signal which can be removed by a low-pass filter. The I signal can be rewritten, after passing through a low-pass filter, as

$$v_I(t) = \frac{1}{2} V_{\mathrm{RF}} V_{\mathrm{LO}} \cos\left[\phi(t)\right] \tag{16.66}$$

Similarly, from the Q-channel, we get the Q signal after its second harmonic is filtered out:

$$v_Q(t) = \frac{1}{2} V_{\mathrm{RF}} V_{\mathrm{LO}} \sin\left[\phi(t)\right] \tag{16.67}$$

It can be seen that the I and Q signals are orthogonal to each other with an exact 90° phase difference between them. It is also recognized that both the final I and Q output signals are baseband signals. The I and Q signals completely represent the received RF signal $v_{\mathrm{RF}}(t)$ and hence can be used to reconstruct it. We consider a (complex) exponential represented by

$$v(t) = v_I(t) + j v_Q(t) = \frac{1}{2} V_{\mathrm{RF}} V_{\mathrm{LO}} \{\cos\left[\phi(t)\right] + j \sin\left[\phi(t)\right]\} \tag{16.68}$$

It is recognized, from (16.66) and (16.67), that

$$V = \frac{1}{2} V_{\mathrm{RF}} V_{\mathrm{LO}} \tag{16.69}$$

and

$$\phi(t) = \tan^{-1}\left[\frac{v_Q(t)}{v_I(t)}\right] \tag{16.70}$$

are the amplitude and phase of the received signal, respectively. Hence, $v(t)$ in (16.68) indeed represents the (composite) received signal. We can rewrite this signal as

$$v(t) = v_I(t) + j v_Q(t) = V e^{j\phi(t)} \tag{16.71}$$

from which V can be determined as

$$V = \sqrt{v_I^2(t) + v_Q^2(t)} \tag{16.72}$$

that, together with (16.70), shows that the received signal can be reconstructed rom the measured I and Q signals. The response of the output signal $v(t)$ is rotated in a circle on a complex plane. When the phase of the received signal is constant, that is, having only a single value (e.g., signal return from a single target in radar), $v(t)$ has only one value. In general, the phase of the received signal varies (e.g., signal return from multiple targets in radar), causing the output signal to rotate in the counter-clockwise (CCW) or clockwise (CW) direction depending on whether the phase $\phi(t)$ is positive or negative, respectively.

Practical Quadrature Mixers

In practice, the in-phase splitter, such as a power divider, used for the RF signal does not provide equal magnitude and phase for the splitting signals, and the 90° phase shifter, such as a 90° hybrid, does not divide the LO signal exactly equal in amplitude and 90° out of phase. These imperfect components along with the difference between other (same) components in the I and Q channels, such as the constituent mixers, low-pass filters, transmission lines, lumped elements, cause mismatch between these channels. This mismatch results in amplitude and phase imbalances between the channels, which vary with frequency due to the frequency-dependence of all the components. This is the most fundamental and severe problem in quadrature mixers, widely known as the I/Q error, which limits the accuracy of measurement, particularly over a wide frequency range. The (frequency-dependent) I/Q error causes actual quadrature mixers to deviate from the ideal behavior and results in nonlinear response. The nonlinear phase response of quadrature mixers is a critical problem in systems as it affects significantly the measurement accuracy. The effect is less in a superheterodyne system as compared to a homodyne system since a single constant intermediate frequency (IF) is typically used for the quadrature mixer, leading to a constant I/Q error over the operating frequency range. The instability of the frequency source also affects the measurement. This, however, should produce a negligible effect provided that the time delay between the transmit and receive signals of the system is short. Additionally, due to finite isolation between the ports of the constituent mixers, signal leakages between the LO and RF ports occur. The leakage of the RF signal to the LO port is typically negligible due to the small RF power as compared to that of the LO signal. However, the leakage of the LO signal onto the RF port can be significant. This LO-leaking signal mixes with the original LO signal to produce an additional signal, which results in a DC component in each of the output signals, known as the DC offset, worsening the nonlinearity caused by the amplitude and phase imbalances. These DC offset voltages, however, can be filtered out by using a BPF. Inevitable errors are thus generated in practical (nonideal) quadrature mixers, which can be severe at high RF frequencies such as those in the millimeter-wave range.

For simplicity without loss of generality, we assume the in-phase splitter used for the RF signal is perfect and all components in the I and Q channels are identical, so the mismatch is caused only by the 90° phase shifter. We also assume that effect of the phase noise of the LO signal source pumping the mixer is negligible; for short delay time between the transmit and receive signals, this assumption is actually valid. We can describe the LO signal arriving at Mixer I as

$$v_{\mathrm{LO}}^{I}(t) = V_{\mathrm{LO}}(1 + \delta V_{\mathrm{LO}}) \cos(2\pi f t) \tag{16.73}$$

or

$$v_{\mathrm{LO}}^{I}(t) = (V_{\mathrm{LO}} + \Delta V_{\mathrm{LO}}) \cos(2\pi f t) \tag{16.74}$$

where δV_{LO} represents the relative amplitude imbalance or relative loss (gain) imbalance between the I and Q channels and ΔV_{LO} denotes the absolute amplitude imbalance between the two channels at frequency f. They are related by

$$\delta V_{\mathrm{LO}} = \frac{\Delta V_{\mathrm{LO}}}{V_{\mathrm{LO}}} \tag{16.75}$$

The LO signal at Mixer Q is given as

$$v_{\mathrm{LO}}^{Q}(t) = V_{\mathrm{LO}} \sin(2\pi f t + \Delta\phi) \tag{16.76}$$

where $\Delta\phi$ represents the (absolute) phase imbalance between the I and Q channels. We now consider two cases: with and without DC offsets in the channels.

No DC Offsets. The RF signals arriving at the constituent mixers are given in (16.64). The I signal produced by the I channel is obtained as

$$v_I(t) = \frac{1}{2}V_{\text{RF}}(V_{\text{LO}} + \Delta V_{\text{LO}})[\cos(4\pi ft - \phi) + \cos(\phi)] \tag{16.77}$$

which becomes, after the second harmonic is suppressed,

$$v_I(t) = \frac{1}{2}V_{\text{RF}}(V_{\text{LO}} + \Delta V_{\text{LO}})\cos[\phi(t)] \tag{16.78}$$

Similarly, we can write the Q signal emerging from the Q channel as

$$v_Q(t) = \frac{1}{2}V_{\text{RF}}V_{\text{LO}}[\sin(4\pi ft - \phi + \Delta\phi) + \sin(\phi + \Delta\phi)] \tag{16.79}$$

or, after the low-pass filter,

$$v_Q(t) = \frac{1}{2}V_{\text{RF}}V_{\text{LO}}\sin[\phi(t) + \Delta\phi] \tag{16.80}$$

Equations (16.78) and (16.80) show that, with the amplitude and phase imbalances, the I and Q signals are no longer orthogonal and balanced; that is, they have phase difference deviating from 90° and unequal amplitude.

With DC Offsets. As mentioned earlier, the LO leakage from the LO port to the RF port of each constituent mixer causes DC offset. Under this condition, the RF signals appearing at the RF ports of Mixer I and Mixer Q in the I and Q channels, respectively, would consist of the original RF signal and the LO leakage and can hence be expressed as

$$v_{\text{RF}}^I(t) = V_{\text{RF}}\cos[2\pi ft - \phi(t)] + \alpha(V_{\text{LO}} + \Delta V_{\text{LO}})\cos(2\pi f_o t) \tag{16.81}$$

and

$$v_{\text{RF}}^Q(t) = V_{\text{RF}}\cos[2\pi ft - \phi(t)] + \beta V_{\text{LO}}\sin(2\pi f_o t + \Delta\phi) \tag{16.82}$$

where $0 < \alpha < 1$ and $0 < \beta < 1$.

Taking a product of v_{LO}^I and v_{RF}^I from (16.74) and (16.81), respectively, gives the I output signal:

$$v_I(t) = \frac{(V_{\text{LO}} + \Delta V_{\text{LO}})}{2}\{V_{\text{RF}}[\cos(4\pi ft - \phi) + \cos\phi] + \alpha(V_{\text{LO}} + \Delta V_{\text{LO}})[\cos(4\pi ft) + 1]\} \tag{16.83}$$

which becomes, after its second harmonic is filtered out,

$$v_I(t) = \frac{(V_{\text{LO}} + \Delta V_{\text{LO}})}{2}[V_{\text{RF}}\cos\phi + \alpha(V_{\text{LO}} + \Delta V_{\text{LO}})] \tag{16.84}$$

The second term

$$V_{\text{OSI}} = \frac{\alpha(V_{\text{LO}} + \Delta V_{\text{LO}})^2}{2} \tag{16.85}$$

is the DC offset voltage of the I signal caused by the LO-to-RF leakage in Mixer I.

Similarly, the Q output signal can be determined from (16.76) and (16.82) as

$$v_Q(t) = \frac{V_{\text{LO}}}{2}\{V_{\text{RF}}[\sin(4\pi ft - \phi + \Delta\phi) + \sin(\Delta\phi + \phi)] + \beta V_{\text{LO}}[1 - \cos(4\pi ft + 2\Delta\phi)]\} \tag{16.86}$$

or, after filtering out the second harmonic,

$$v_Q(t) = \frac{V_{\text{LO}}}{2}\{V_{\text{RF}} \sin(\phi + \Delta\phi) + \beta V_{\text{LO}}\} \tag{16.87}$$

where the second term

$$V_{\text{OSQ}}(t) = \frac{\beta V_{\text{LO}}^2}{2} \tag{16.88}$$

represents the DC offset of the Q signal caused by the LO-to-RF leakage in Mixer Q.

The I and Q output signals of a practical quadrature mixer can now be rewritten from (16.84) to (16.85) and (16.87) to (16.88), respectively, as

$$v_I(t) = (V + \Delta V)\cos(\phi) + V_{\text{OSI}} \tag{16.89}$$

and

$$v_Q(t) = V \sin(\phi + \Delta\phi) + V_{\text{OSQ}} \tag{16.90}$$

where $V \equiv V_{\text{RF}}V_{\text{LO}}/2$ and $\Delta V \equiv V_{\text{RF}}\Delta V_{\text{LO}}/2$.

The function of a quadrature mixer, whether it is used in a homodyne or superheterodyne system, is to directly down-convert a signal; that is, it functions as a homodyne system itself. In addition to the amplitude- and phase-imbalance and DC offset issues seen in (16.89) and (16.90), the $1/f$ noise contribution is also a critical problem in the direct down conversion. A simple way to overcome this problem is to slightly shift the LO frequency so that the frequency of the mixer's output signals is located sufficiently far away from the $1/f$ noise spectrum.

The phase of the received signal for practical quadrature mixers, considering the nonlinear phase response due to their phase and amplitude imbalances as well as the DC offset voltages, can be obtained by solving equations (16.89), (16.90) for $\phi(t)$ as [2]

$$\phi(t) = \tan^{-1}\left(\frac{1}{\cos\ \Delta\phi}\frac{V}{(V + \Delta V)}\frac{v_I(t) - V_{\text{OSI}}}{v_Q(t) - V_{\text{OSQ}}} - \tan\Delta\phi\right) \tag{16.91}$$

The measured I and Q signals are typically obtained through an average of many measured values to cancel out the noise components, which are composed of the phase noise of the LO source and the white noise generated by components in the system. As an example, Figure 16.29(a) shows the measured I and Q signal voltages, excluding DC-offset voltages, for an actual millimeter-wave system used for measuring displacement of objects, and Figure 16.29(b) shows the phase detected and constructed considering errors resulting from the amplitude and phase imbalances of the quadrature mixer [2].

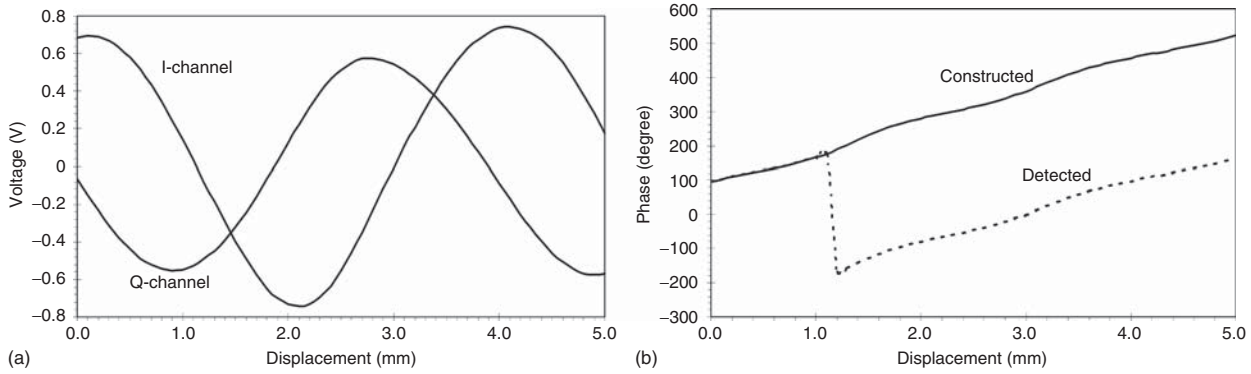


Figure 16.29. Measured I and Q signal voltages (a) and detected and constructed phase (b) versus an object's displacement.

Error Analysis for Quadrature Mixers

***I/Q* Error.** In the foregoing analysis, we consider same RF and LO frequencies. Now we assume that the RF and LO frequencies are slightly different by $f_{IF} = f_{RF} - f_{LO}$. Accordingly, the I and Q output signals of non-ideal quadrature mixers may be expressed, considering the mixer's I/Q error and DC offsets, as

$$\begin{aligned} v_I(t) &= (V \cdot \Delta V)\cos(2\pi f_{IF}t + \phi + \Delta\phi) + V_{OSI} \\ v_Q(t) &= V\sin(2\pi f_{IF}t + \phi) + V_{OSQ} \end{aligned} \tag{16.92}$$

where the amplitude imbalance ΔV is considered the ratio of the signal amplitudes. The actual phase detection process is performed on the detected signals of (16.92) excluding DC offset terms, because these DC offsets can be easily removed by a band-pass filter in the system. Therefore, the phase error produced by the nonideal quadrature signals can be calculated by [2]

$$\phi_e = \tan^{-1}\left[\frac{v_Q(t) - V_{OSQ}}{v_I(t) - V_{OSI}}\right] - \phi \tag{16.93}$$

As an example, Figure 16.30 shows the nonlinear phase response of a quadrature mixer, accompanied by phase error corresponding to an amplitude imbalance of 2 dB, phase imbalance of 10°, and DC offset of 100 mV. As can be seen, the phase response is non-linear for the linear change of the input phase and shows undulating behavior, which implies deterioration in measurement accuracy.

In a homodyne system, the frequency of RF and LO signals is the same, so the output of the quadrature mixer generates only DC terms. This makes the estimation of phase error difficult. To alleviate this problem, the RF signal is mixed with an LO signal whose frequency is slightly different from the RF frequency to generate a signal of intermediate frequency, f_{IF}, at the output of the mixer, which is used as the test signal for measurement purposes. The frequency f_{IF} is usually chosen sufficiently low to be processed by digital signal processing. Based on the test signal, it is possible to estimate the amplitude and phase imbalance as well as the DC offset of the quadrature mixer in frequency domain using Fourier transform, as explained in details in the following.

For mathematical convenience, we consider a complex signal made up of the two signals described in (16.92)

$$v(t) = v_I(t) + jv_Q(t) \tag{16.94}$$

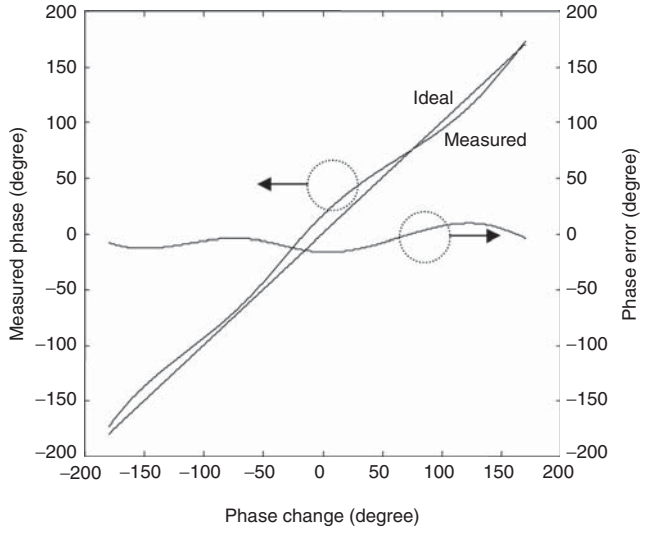


Figure 16.30. Example of nonlinear phase response of a quadrature mixer.

which is the complex output signal produced by real quadrature mixers. The Fourier transform of $v(t)$ in (16.94) is an impulse (delta) function (in frequency domain.) For an ideal quadrature mixer with perfect amplitude and phase balances, the impulse function appears only at the frequency f_{IF}. For a non-ideal quadrature mixer, however, the imbalances cause image response at the negative (or image) frequency of f_{IF}.

Taking Fourier transform (F) of (16.94), making use of (16.92), gives

$$\begin{aligned} F[v(t)] &= F[v_I(t) + jv_Q(t)] \\ &= F(0) && : \text{DC Term} \\ &\quad + \frac{1}{2}Ve^{j\phi}[\Delta V\cos(\Delta\phi) + j\Delta V\sin(\Delta\phi) - 1]\delta(f + f_{IF}) && : \text{Image Signal} \\ &\quad + \frac{1}{2}Ve^{j\phi})[\Delta V\cos(\Delta\phi) + j\Delta V\sin(\Delta\phi) + 1]\delta(f - f_{IF}) && : \text{Primary Signal} \end{aligned} \tag{16.95}$$

which constitutes a DC term coming from the DC offset, upper-side (Primary) and lower-side (Image) signals. The ratio of the lower to upper signal powers, namely the Image-to-Signal Ratio (ISR), measures the amount of deviation of the real quadrature mixer's response with respect to an ideal response, and is defined by

$$\text{ISR} = \frac{[\Delta V\cos(\Delta\phi) - 1]^2 + [\Delta V\sin(\Delta\phi)]^2}{[\Delta V\cos(\Delta\phi) + 1]^2 + [\Delta V\sin(\Delta\phi)]^2} \tag{16.96}$$

The amplitude imbalance can be derived from (16.96) as [4]

$$\Delta V = \frac{(1 + \text{ISR})\cos(\Delta\phi) + \sqrt{[(1 + \text{ISR})\cos(\Delta\phi)]^2 - (1 - \text{ISR})^2}}{1 - \text{ISR}} \tag{16.97}$$

It is relatively easier to measure ISR and the amplitude imbalance ΔV using a spectrum analyzer than to measure the phase imbalance $\Delta\phi$. From the measured ISR and amplitude imbalance, the phase imbalance can be deduced as

$$\Delta\phi = \cos^{-1}\left[\frac{\left(\Delta V^2 + 1\right)(1 - \text{ISR})}{2\Delta V(1 + \text{ISR})}\right] \tag{16.98}$$

Figure 16.31 illustrates the frequency response of a complex output signal of a quadrature mixer for both ideal and nonideal cases. The image signal shown in the figure is generally called *Hermitian* image. It is undesirable as it degrades the system performance. For instance, it produces false target and deteriorates resolution in radar used for most sensing applications. It also causes a nonlinear phase response, as seen in Figure 16.30. It is thereby desirable to suppress or eliminate the image signal.

***I*/*Q* Error Correction Based on Gram–Schmidt Orthogonalization.** As discussed in the previous section, the image signal affects the quadrature phase detection. In most radar sensing and communication applications, it is desirable to correct the I/Q error of quadrature mixers. In this section, the most common method to correct the I/Q error is presented by means of the correction coefficients derived from the test signal [5].

We begin by expressing the quadrature signals at frequency f_{IF}, taking into account the I/Q errors, as

$$\begin{aligned} v_I(t) &= (V + \Delta V)\cos(2\pi f_{\text{IF}}t) \\ v_Q(t) &= V\sin(2\pi f_{\text{IF}}t + \Delta\phi) \end{aligned} \tag{16.99}$$

where the RF and LO signals are assumed to have the same phase for simplicity. The DC offset is excluded because it can be simply determined by the averaged DC level (or zero frequency component in the Fourier transform) of each quadrature signal.

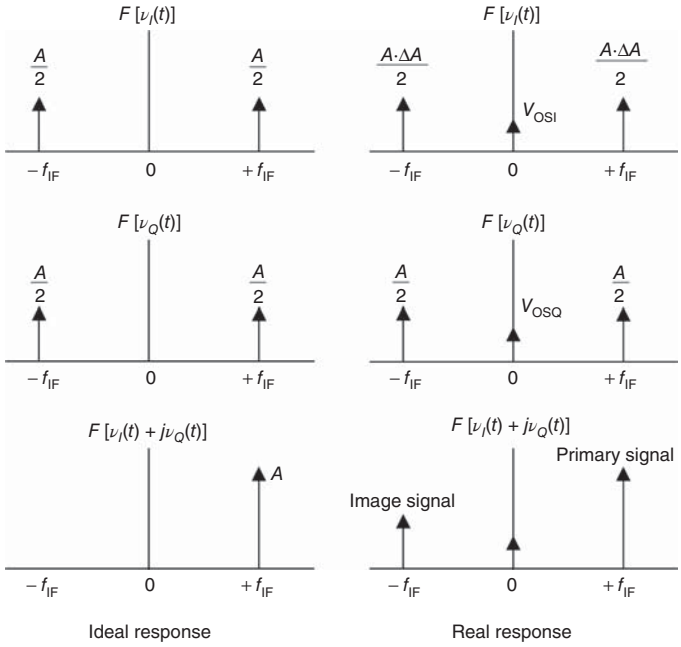


Figure 16.31. Frequency responses of ideal and real quadrature mixers.

The problem of I/Q error correction is analogous to the Gram–Schmidt orthogonalization, which transforms non-orthogonal signals into orthogonal signals. Mathematically, this transformation is described as

$$\begin{bmatrix} v'_I(t) \\ v'_Q(t) \end{bmatrix} = \begin{bmatrix} S & 0 \\ R & 1 \end{bmatrix} \begin{bmatrix} v_I(t) \\ v_Q(t) \end{bmatrix} \tag{16.100}$$

where S and R are the respective rotating and scaling coefficients of the coefficient matrix that transforms the quadrature signals $v_I(t)$ and $v_I(t)$ into new signals $v'_I(t)$ and $v'_Q(t)$, respectively, that are orthogonal (i.e., exactly 90° out of phase) with equal amplitude. In [5], a digital signal processing technique is suggested to obtain estimates of the coefficient matrix using discrete Fourier transform (DFT). The DFT of the complex quadrature signal $v(t) = v_I(t) + jv_Q(t)$ can be expressed as [2]

$$F\left(\frac{k}{NT_s}\right) = \frac{1}{N}\sum_{n=0}^{N-1} v(nT_s)\exp\left(-j\frac{2\pi kn}{N}\right), \quad k = 0, 1, 2, \ldots, N-1 \tag{16.101}$$

where T_s is the sampling time, N is the number of samples, and

$$v(nT_s) = v_I(nT_s) + jv_Q(nT_s) = (V + \Delta V)\cos\left[2\pi f_{\text{IF}}(nT_s)\right] + jV\sin\left[2\pi f_{\text{IF}}(nT_s) + \Delta\phi\right] \tag{16.102}$$

Substituting (16.102) into (16.101) and solving for the DFT components at the frequencies of $1/NT_s$ and $(N-1)/NT_s$ give

$$F\left(\frac{1}{NT_s}\right) = \frac{V}{2}\left[(1 + \Delta V) + \cos(\Delta\phi) + j\sin(\Delta\phi)\right] \tag{16.103}$$

and

$$F\left(\frac{N-1}{NT_s}\right) = \frac{V}{2}\left[(1 + \Delta V) - \cos(\Delta\phi) + j\sin(\Delta\phi)\right] \tag{16.104}$$

which represent the primary and image signal components, respectively, that are related to the amplitude and phase imbalance. The estimates of the coefficients S and R in (16.100) can then be obtained from the components of the DFT in (16.103), (16.104) as

$$\hat{S} = -\mathrm{Re}\left\{ \frac{2F\left[(N-1)/NT_s\right]}{F^*(1/NT_s) + F[(N-1)/NT_s]} \right\} + 1 \tag{16.105}$$

and

$$\hat{R} = -\mathrm{Im}\left\{ \frac{2F\left[(N-1)/NT_s\right]}{F^*(1/NT_s) + F[(N-1)/NT_s]} \right\} \tag{16.106}$$

respectively, where F^* indicates conjugate of F.

Figure 16.32 illustrates an example of I/Q error correction by the coefficients obtained by DFT, where the amplitude imbalance ΔV is 0.8 dB, the phase imbalance $\Delta\phi$ is 10°, and the DC offset voltages V_{OSI} and V_{OSQ} are 100 and 200 mV for the I and Q channels, respectively [2]. Figure 16.32(a) shows the original quadrature signals corresponding to the imbalances, and Figure 16.32(b) demonstrates the geometric transformation resulting from the correction process. The in-phase signal $v_I(nT_s)$ and quadrature signal $v_Q(nT_s)$ of the original signal $v(nT_s)$ plotted in the x- and y-axis, respectively, constitute an ellipse in the xy complex plane with center offset from the origin (0,0), as seen in Figure 16.32(b), which is undesirable due to the amplitude and phase imbalance. From a geometric viewpoint, the procedure for correcting the I/Q error can be interpreted as rotation and scaling of the ellipse in the xy complex plane, so that it finally turns into a perfect circle centered at the origin, as shown in Figure 16.32(b). The coefficients S and R defined in (16.100) can be geometrically interpreted as the scaling and rotation coefficients to convert the original signal represented by the right-most vector $v(t) = v_I(t) + jv_Q(t)$ in (16.100), which is neither orthogonal (90° phase difference) nor equal amplitude, into a new signal represented the left-most vector $v'(t) = v'_I(t) + jv'_Q(t)$, which is orthogonal and equal in amplitude. This transformed signal $v'(t)$ is geometrically represented by the centered circle in Figure 16.32(b).

Worst-Case Error Analysis. Measurement accuracy can be estimated by an analysis of the maximum phase imbalance using the method proposed in [5]. In this method, the maximum phase error resulting from the image-rejection level of quadrature mixers is calculated. From equation (16.96), the ISR described in terms of the amplitude and phase imbalance can be reduced to [2]

$$\mathrm{ISR} = \frac{1 + \Delta V^2 - 2\Delta V \cos(\Delta\phi)}{1 + \Delta V^2 + 2\Delta V \cos(\Delta\phi)} \tag{16.107}$$

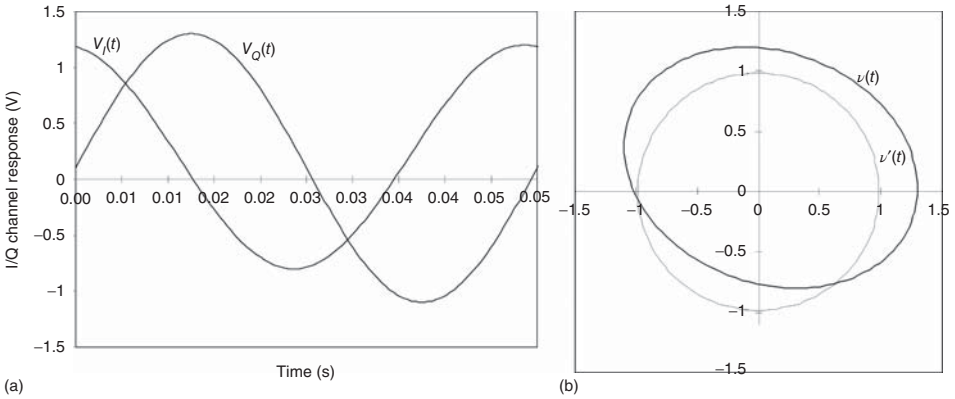


Figure 16.32. I/Q error correction: (a) I/Q channel response and (b) transformation of nonorthogonal to orthogonal signals through I/Q error correction.

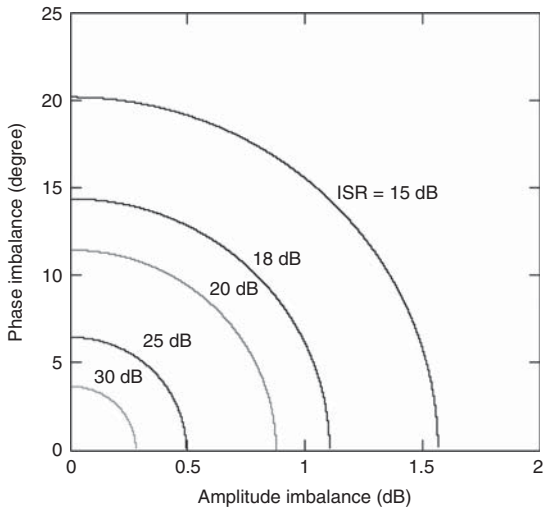


Figure 16.33. Constant ISR contours.

Equation (16.107) can be approximated by an ellipse

$$\left(\frac{\Delta\varphi}{X}\right)^2 + \left(\frac{\Delta V}{Y}\right)^2 = 1 \tag{16.108}$$

where

$$X = \cos^{-1}\left(\frac{1 - \text{ISR}}{1 + \text{ISR}}\right) \tag{16.109}$$

and

$$Y = \frac{1 + \sqrt{\text{ISR}}}{1 - \sqrt{\text{ISR}}} \tag{16.110}$$

Figure 16.33 shows constant ISR contours for several different ISR values of a quadrature mixer. As can be seen, the maximum phase imbalance occurs when the amplitude imbalance is 0 dB. These curves can be used to estimate the maximum phase error from a given ISR value, from which the maximum error of a measured quantity, such as range, can be determined. For instance, for an ISR of 18 dB, the maximum phase error is obtained as 14.4°.

Measurement-Based *I*/*Q* Error Compensation. Specifically, a practical system produces "common" and "differential" amplitude and phase errors in the I and Q channels. The "common error" is the error caused by common circuits in the signal-propagation paths (connecting) to both the I and Q channels, which consist of antennas, amplifiers, other mixers, transmission lines, filters, etc. The "differential error" is caused by a mismatch between the I and Q channels; it is the main error and is the commonly known I/Q error that we discussed previously. For a superheterodyne scheme, the differential amplitude and phase errors in the I and Q channels are normally constant over the RF band of interest due to a (down-converted) single constant IF.

For illustration without lost of generality, we assume the RF signal transmits in N multiple frequencies separated by a constant step; that is, the transmit frequencies are f_i, $i = 0, 1, 2, \ldots, N-1$. In the absence of errors in the I and Q channels, the phase ϕ_i of the base-band I and Q signals, corresponding to f_i, expressed in terms of the target range R is [3]

$$\phi_i(R, f_i) = -\frac{4d\pi f_i}{v} = -2\pi f_i t_d, \quad i = 0, 1, \ldots, N-1 \tag{16.111}$$

where v is the speed of light in the propagating medium and t_d is the time delay equal to a two-way travel time of $2R/v$. The complex vectors corresponding to a fixed target are expressed in terms of the frequency f_i, assuming equal amplitude, as

$$\begin{aligned} I_i(f_i) + jQ_i(f_i) &= A_i \cos[\phi_i(f_i)] + jA_i \sin[\phi_i(f_i)] \\ &= A_i e^{-j2\pi f_i t_d} \end{aligned} \tag{16.112}$$

where A_i is the amplitude.

If the common and differential errors are included, the complex vectors become, assuming $A_i = 1$ for simplicity without loss of generality:

$$I_i(f_i) + jQ_i(f_i) = \left(1 + \frac{cg_i}{2}\right) \cos[2\pi f_i t_d + cp_i] - j\left(1 + dg_i + \frac{cg_i}{2}\right) \sin[2\pi f_i t_d + dp_i + cp_i] \tag{16.113}$$

where cg_i, cp_i and dg_i and dp_i are the common and differential amplitude and phase errors, respectively.

The differential amplitude and phase errors generate a Hermitian image of the output response, degrading system performance as mentioned earlier. In a superheterodyne system, these errors are constant across the RF operating frequency range since a single constant IF is typically used for the quadrature mixer. Consequently, the measurement and compensation of these errors is simple. The differential amplitude and phase errors in the I and Q channels at IF can be measured by using the method presented in [5].

The common phase error can be described as consisting of a "linear" phase error $2\pi f_i \alpha$ and a "nonlinear" phase error β_i as

$$cp_i = 2\pi f_i \alpha + \beta_i \tag{16.114}$$

The common linear phase error results in a constant shift of the output response due to the fact that a frequency-dependent linear phase is transformed into a constant time delay through the inverse Fourier transform. Therefore, it is not necessary to correct for the common linear phase error. On the other hand, the nonlinear phase error causes shifting as well as imbalance in the response. The common amplitude error also affects the response – for instance, the shape of synthetic range profiles in radar as they tend to defocus the response in the profile and increase the magnitudes of side lobes. Therefore, these common nonlinear phase and amplitude errors need to be corrected. The following formulates a simple, yet effective, and accurate, technique for compensation of these errors.

The complex vector given in equation (16.113) for a fixed frequency f_k is rewritten in terms of the range R as

$$I(R) + jQ(R) = \left(1 + \frac{cg_k}{2}\right) \cos[2\pi f_k t_d(R) + cp_k] - j\left(1 + dg_k + \frac{cg_k}{2}\right) \sin[2\pi f_k t_d(R) + dp_k + cp_k] \tag{16.115}$$

where $t_d(R)$ signifies the dependence of t_d on R. It is then seen that these complex vectors will rotate circularly if the I and Q channels are completely balanced when R is increased or decreased at a constant rate. In the process of correction, the complex vector, $I(R) + jQ(R)$, is measured when a metal plate is moved along a track at a fixed frequency. Initially, the complex vector rotates elliptically, either clockwise or counter-clockwise, with respect to the direction of the metal plate, as the I and Q components are not orthogonal due to the differential phase errors. After these differential errors are corrected, equation (16.115) becomes

$$I(R) + jQ(R) = \left(1 + \frac{cg_k}{2}\right) \cos[2\pi f_k t_d(R) + cp_k] - j\left(1 + \frac{cg_k}{2}\right) \sin[2\pi f_k t_d(R) + cp_k] \tag{16.116}$$

from which, it is seen that the I and Q components become orthogonal in phase and balanced in amplitude; hence the complex vector $I(R) + jQ(R)$ starts rotating circularly during the movement of the metal plate at a fixed frequency. The magnitude of the rotating vector is then measured and stored. This procedure is repeated at each frequency of the RF operating frequency range. These measured magnitudes are used as reference data to compensate for the common amplitude errors.

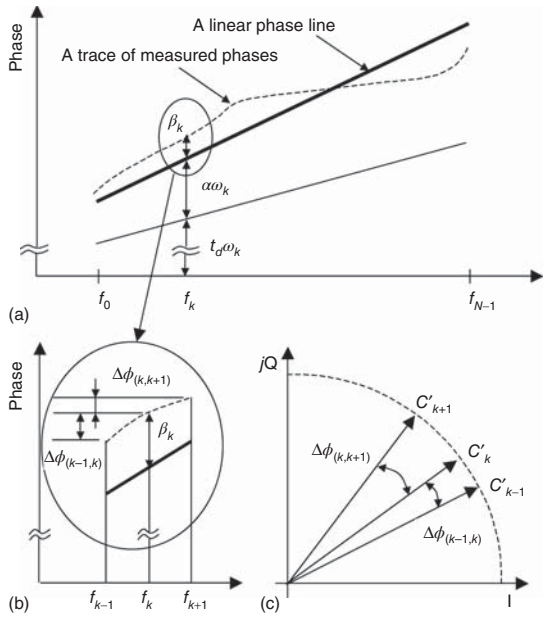


Figure 16.34. Phase of the complex vector $I + jQ$ versus frequency: (a) linear transformation of the trace of calculated phases to a linear phase line, $(\alpha + t_d)\omega_k$, (b) a magnified drawing of (a) showing the trace of the calculated phases obtained by cumulating the phase differences $\Delta\phi_{0,1}, \ldots, \Delta\phi_{k-1,k}, \ldots, \Delta\phi_{N-2,N-1}$, (c) nonlinearity of the calculated phases in polar form, where C'_k is the kth complex vector after compensating for the common amplitude deviation.

After compensating for the common amplitude errors, the normalized complex vectors $I + jQ$ can be expressed, using (16.116), as

$$I(R) + jQ(R) = \cos\left[2\pi f_k t_d(R) + cp_k\right] - j\sin\left[2\pi f_k t_d(R) + cp_k\right] \tag{16.117}$$

From (16.117), the phase of the complex vector $I + jQ$ is obtained as

$$\phi(f_i) = 2\pi f_k t_d + 2\pi f_k \alpha + \beta_k \tag{16.118}$$

with the aid of equations (16.114) and (16.115). As mentioned earlier, the non-linear phase error β_k needs to be corrected. Figure 16.34 illustrates the calculated phases $\phi(f_i)$ versus frequency [3]. Cumulating the phase difference between two consecutive RF frequencies

$$\Delta\phi_{k-1,k} = 2\pi f_k t_d + 2\pi f_k \alpha + \beta_k - (2\pi f_{k-1} t_d + 2\pi f_{k-1}\alpha + \beta_{k-1}) \tag{16.119}$$

unwraps the calculated phases and makes it easy to draw the trace of the calculated phases as shown in Figures 16.34(a) and (b). Figure 16.34(c) shows that the rotation of the vector $I + jQ$ is not constant, which is due to the nonlinear phase error β_k.

After drawing an appropriate linear phase line as shown in Figure 16.34(a), the nonlinear phase error β_k is then determined by subtracting the linear phase line from the trace of the calculated phases. Consequently, the complex vector is obtained, after correcting for the nonlinear phase error, as

$$\begin{aligned} I(R) + jQ(R) &= \cos\left[2\pi f_k t_d(R) + 2\pi f_k \alpha\right] - j\sin\left[2\pi f_k t_d(R) + 2\pi f_k \alpha\right] \\ &= \exp\{-2\pi f_k[t_d(R) - \alpha]\} \end{aligned} \tag{16.120}$$

The nonlinear phase error β_k at all the RF frequencies for the metal plate is stored in memory and used as reference data for compensating for the nonlinear phase error of an actual target. The flow chart in Figure 16.35 [3] shows the procedure for extracting the common amplitude and nonlinear phase errors.

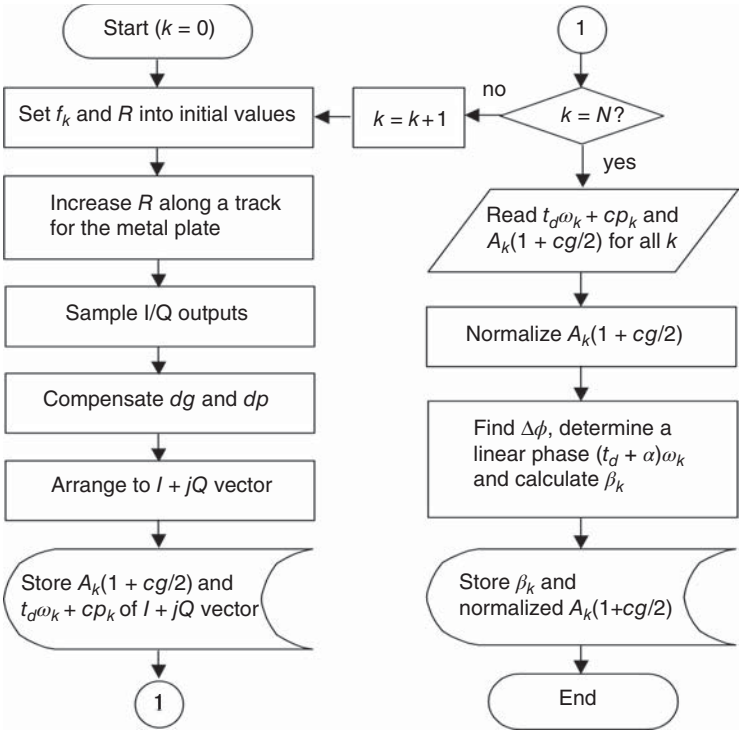


Figure 16.35. Flow chart for calculating the common errors.

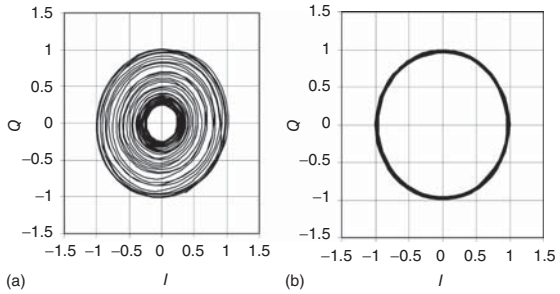


Figure 16.36. Normalized I/Q (a) before and (b) after compensating for the amplitude deviations and non-linear phase errors of a quadrature mixer.

In order to compensate for the measured complex vectors of targets for the common amplitude and nonlinear phase errors, the reference data, extracted from a metal plate, as described earlier, are applied to these vectors. The stored reference data for the common amplitude errors are normalized, inversed, and multiplied to the target's measured complex vectors. The stored reference data for the common nonlinear phase errors are subtracted from the extracted phases of the target's measured complex vectors. As an example, we show in Figure 16.36 the normalized I/Q outputs of a quadrature mixer before and after compensating for the common amplitude and nonlinear phase errors.

Digital Quadrature Mixer

The inherent imbalance problem of (analog) quadrature mixers coupled with the DC offsets, and hence the nonlinear phase response, critically limits measurement accuracy. To overcome this problem, a digital quadrature mixer (DQM) based on [7] can be used in place of the quadrature mixer. In this approach, we assume that the phase information is detected at an intermediate frequency f_{IF}, which is low enough to be handled with a DSP. The DQM is realized by software and functions as a phase detecting processor that

Figure 16.37. Implementation of DQM in systems.

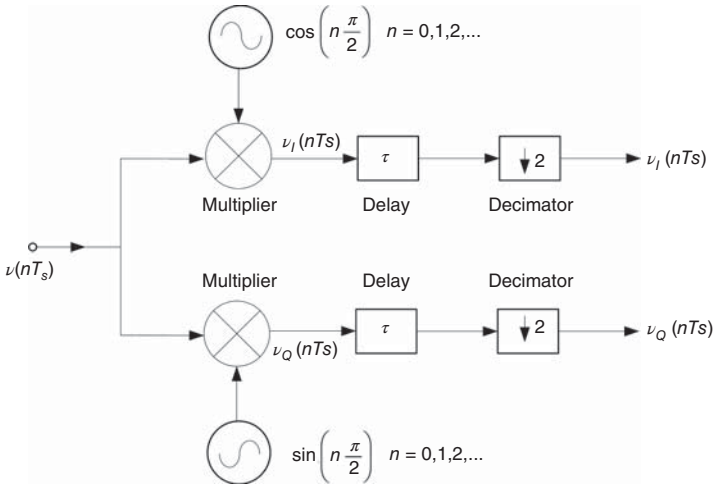


Figure 16.38. Configuration of the digital quadrature mixer.

employs the quadrature sampling signal-processing technique to measure the phase. Figure 16.37 shows an implementation of the DQM in systems. The down-converted received signal at f_{IF} is amplified by an IF amplifier and converted into a digital form by an ADC. This digital signal can be expressed as

$$v(nT_s) = V \sin [2\pi f_{\text{IF}}(nT_s) + \phi(nT_s)], \quad n = 1, 2, 3, \ \ldots \tag{16.121}$$

which, upon processed by the DQM, produces the digital in-phase $v_I(nT_s)$ and quadrature $v_Q(nT_s)$ signal components.

Figure 16.38 shows a configuration of the DQM [2]. Various quadrature sampling schemes have been proposed for coherent detection in radar and communication receivers. The advantage of the quadrature sampling is that it can eliminate or, at least, minimize the nonlinear phase response of a conventional analog quadrature mixer, which is caused by the phase and amplitude imbalances as well as the DC offset voltages of the mixer itself. As the operating frequency is increased, the nonlinearity becomes severe and difficult to control. The DQM processes each digitized channel signal to generate the in-phase and quadrature components of $v_I(nT_s)$ and $v_Q(nT_s)$. The sampling frequency is set as four times of the intermediate frequency, $4f_{\text{IF}}$, so that the digital LOs become a quadrature sequence of only -1, 0, and 1, which implies that LOs feed signals of exact 90° out of phase and equal amplitude into the constituent mixers, because their phases are integer multiples of $\pi/2$. Each constituent mixer designated in Figure 16.38 performs as a multiplier. The multiplication process samples the following in-phase and quadrature components of the down-converted received signal [8]:

$$v_I(nT_s) = \begin{bmatrix} 0 & n = \text{odd} \\ I(nT_s)\cos\,\phi_i + Q(nT_s)\sin\,\phi_i & n = \text{even} \end{bmatrix} \tag{16.122}$$

and

$$v_Q(nT_s) = \begin{bmatrix} 0 & n = \text{even} \\ Q(nT_s)\cos\,\phi_i - I(nT_s)\sin\,\phi_i & n = \text{odd} \end{bmatrix} \tag{16.123}$$

where $I(nT_s) = V\cos\phi(nT_s)$ and $Q(nT_s) = V\sin\phi(nT_s)$. ϕ_i is the initial phase, which is static in nature. As seen in (16.122) and (16.123), the odd samples of the in-phase signal $v_I(nT_s)$ and the even samples of the quadrature signal $v_I(nT_s)$ always produce zero, which is caused by multiplication with zero from the digital oscillators, and they need to be discarded. Decimating by two discards those samples to eliminate zero output in (16.124). In this quadrature sampling approach, a time delay in the quadrature signal occurs because the first sample of $v_I(nT_s)$ produces zero and it is discarded. Therefore, adding a time delay of $\tau = 1/4f_{\text{IF}}$ to the in-phase signal eliminates the time delay between the two orthogonal signals, $v_I(nT_s)$ and $v_Q(nT_s)$. Taking arctangent then produces the phase of each channel signal within 2π radians, $[-\pi, \pi]$, as

$$\phi(nT_s) = \tan^{-1}\left[\frac{v_I(nT_s)}{v_Q(nT_s)}\right] \tag{16.124}$$

In this configuration, a low-pass filter is not needed for the rejection of harmonics as in a typical mixer configuration, thus avoiding the filter's transient response to appear in the quadrature outputs, another advantage of the DQM approach.

REFERENCES

1. J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, New Jersey, 1996.
2. S. T. Kim, "Development of Millimeter Wave Integrated-Circuit Interferometric Sensors for Industrial Sensing Applications," Ph.D Dissertation, Texas A&M University, College Station, Texas, 2004.
3. J. S. Park, "Development of Microwave and Millimeter-Wave Integrated-Circuit Stepped-Frequency Radar Sensors for Surface and Subsurface Profiling," Ph.D Dissertation, Texas A&M University, College Station, Texas, 2003.
4. S. J. Goldman, *Phase Noise Analysis in Radar Systems Using Personal Computers*, John Wiley & Sons, New York, 1989, Chapter 10.
5. F. E. Churchill, G. W. Ogar, and B. J. Thompson, "The Correction of I and Q Errors in a Coherent Processor," *IEEE Trans. Aerosp. Electron. Syst.*, Vol. AES-17, pp. 131–137, Jan. 1981.
6. R. A. Monzingo and S. P. Au, "Evaluation of Image Response Signal Power Resulting from I–Q Channel Imbalance," *IEEE Trans. Aerosp. Electron. Syst.*, Vol. AES-23, pp. 285–287, Mar. 1987.
7. D. W. Rice and K. H. Wu, "Quadrature Sampling with High Dynamic Range," *IEEE Trans. Aerosp. Electron. Syst.*, Vol. AES-18, pp. 736–739, Nov. 1982.
8. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1984.

PROBLEMS

16.1 Derive Eq. (16.20).

16.2 Consider a monostatic system with different transmit and receive antennas as shown in Figure 16.2(a). Derive Eq. (16.12).

16.3 What is the maximum range of a radar operating at 10 GHz using a 2 m dish antenna to detect a target having a RCS of 100 m^2? The minimum detectable signal is 10^{-10} W and the transmitter's peak power is 10 kW. The maximum gain (directivity) of a dish antenna is given as

$$G = 0.55\left(\frac{\pi D}{\lambda}\right)^2$$

where D is the antenna's diameter and λ is the operating wavelength in air.

16.4 Derive the radar equation for a bistatic system shown in Figure P16.1 assuming the antennas are lossless and no loss is involved in the transmission and reception.

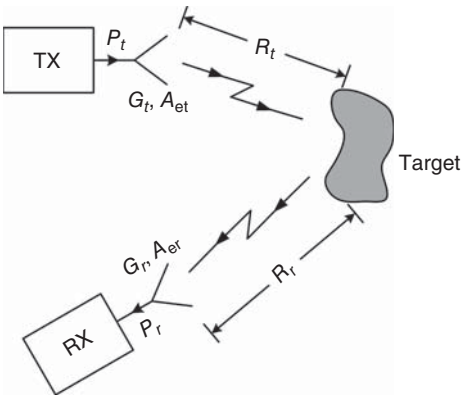


Figure P16.1.

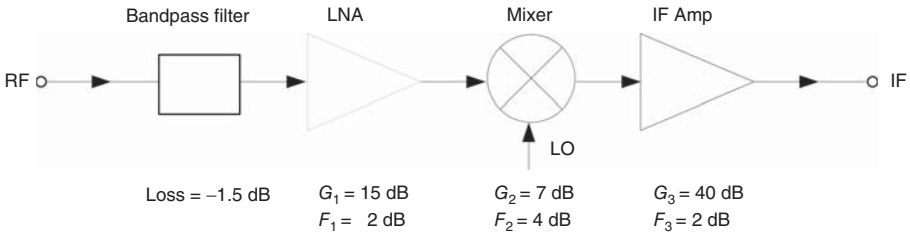


Figure P16.2.

16.5 Derive the radar equation for a bistatic system shown in Figure P16.1 assuming the antennas is lossless and the losses in the transmission and return paths are L_t and L_r, respectively.

16.6 Derive the general noise figure equation for an RF subsystem consisting of N cascaded components.

16.7 Consider a superheterodyne receiver frontend as shown in Figure P16.2. The gain (G's), loss and noise figure (F's) for each component are given in the figure. Compute the receiver's noise figure.

16.8 Consider a pulse system radiating a train of pulses with peak power of 10 W, pulse length of 1 ns and pulse repetition frequency (PRF) of 1 MHz. Find

a) Pulse period.
b) Duty cycle.
c) Average power.

16.9 Consider a rectangular pulse having peak voltage $V_p = 1$ V and pulse length $\tau = 0.25$ ns.

a) Compute and plot the pulse's frequency spectrum.
b) Determine, from the numerical results, the 3- and 4-dB bandwidths of the spectrum. Compare the results with respect to $1/\tau$ and, from which, estimate the 3- and 4-dB bandwidths in terms of $1/\tau$.

16.10 Consider a monocycle pulse, modeled as a single-period sine wave. Compute and plot the pulse's frequency spectrum. Assume the maximum voltage amplitude is $V_p = 1$ V and the pulse duration (one sine-wave period) is $\tau = 0.25$ ns.

a) Compute and plot the pulse's frequency spectrum.
b) Determine, from the numerical results, the 3- and 4-dB bandwidths, and center frequency of the spectrum. Compare the determined bandwidths with respect to $1/\tau$ and, from which, estimate the 3- and 4-dB bandwidths in terms of $1/\tau$. Compare the calculated center frequency to $1/\tau$.

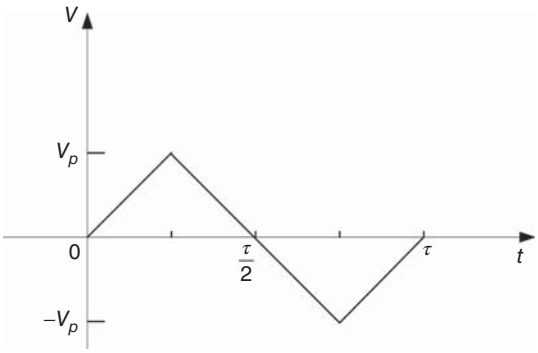


Figure P16.3.

16.11 Repeat Problem 17.10 with a monocycle pulse modeled as two triangular pulses as shown in Figure P16.3.

16.12 Consider a monostatic radar operating at 10 GHz in air. The antenna's gain is 15 dB; the transmitting peak power is 100 W; the target is located at a range of 10 m and its cross section is 0.1 m^2; and the receiver has a bandwidth of 100 kHz, gain of 30 dB and noise temperature of 600 K. Assume the system loss is negligible. Calculate the receiver's S/N ratio.

16.13 Consider a single-antenna monostatic pulse system with the following specifications: peak power = 10 W, PRF = 100 kHz, antenna gain = 20 dB, frequency = 10 GHz, system loss = −3 dB, receiver noise figure = 4 dB, and receiver temperature = 290 K. Consider a target having a cross section of 100 cm^2 located at 25 m from the antenna and assume the transmission medium is air.

a) Determine the transmitting pulse width to achieve a range resolution of 1 cm and calculate the corresponding S/N.

b) Assume 50-dB S/N, determine the pulse width of the transmitting pulse. What are the corresponding bandwidth and range resolution?

16.14 A single-antenna monostatic pulse system has the following specifications: peak power = 10 W, PRF = 1 kHz, pulse length = 1 ns, antenna gain = 15 dB, frequency = 1 GHz, S/N = 20 dB, system loss = −15 dB, receiver noise figure = 5 dB, and receiver temperature = 290 K. Consider a target with a cross-section of 100 cm^2, calculate the detection range of this system.

16.15 In pulse systems, the return echo from a target due to a transmitted pulse may arrive after the transmission of the next pulse, causing "ambiguity" in range measurement. This problem is more severe when multiple targets need to be detected. To avoid range ambiguity, proper PRF must be selected according to the "maximum unambiguous range" of

$$R_u = \frac{v}{2(\text{PRF})}$$

where v is the velocity of the pulse signal. One way to increase the detection range of a pulse system is to increase its PRF. This, however, may cause the problem of range ambiguity. Assume there are three targets located at ranges of 10, 15, and 20 inches from the system and the relative dielectric constant of the propagating medium is 6. Calculate the maximum PRF the system can have in order for these ranges to be measured without any ambiguity.

16.16 For certain applications such as that requiring an altimeter, it is not necessary to use a linear frequency modulation waveform in FMCW systems. In such systems, a sinusoidal frequency modulation can be used and is typically easier to perform than a linear frequency modulation. How do you obtain a sinusoidal modulation and why is it easier than a linear frequency modulation?

16.17 Consider an FMCW system and assume the detected IF frequency is 1 kHz, calculate the needed rate of change of frequency in order to measure ranges of 1, 10, and 100 m.

16.18 The transmitter of an FMCW is swept at a rate of 30 MHz/s. Calculate the frequency difference between the transmit and receive signals for ranges of 10, 10.5, 12, and 15 inches. The modulation is assumed to be linear and the propagating medium is air. Comment on the results: can the frequencies be measured accurately? If not, what can be done to measure them correctly? What causes such frequency values?

16.19 An FMCW system implements a linear frequency modulation with a modulation frequency of 1 kHz and frequency change of 100 MHz. Calculate the average frequency difference for range increments of 10, 15, and 20 inches.

16.20 Repeat Problem 17.19 for a triangular frequency modulation.

16.21 Compare the required average power transmitted by pulse and FMCW systems to achieve the same range.

16.22 An antenna with a half-power beamwidth of 10° radiates a CW signal. Assume the antenna's scan rate is 20°/s; determine the bandwidth of the target's return signal (i.e., received signal.)

16.23 Consider a quadrature mixer and assume the initial phase $\phi_o = 0$ and:

1) $v^I_{\text{RF}}(t)$ and $v^Q_{\text{RF}}(t)$ are related by $|v^I_{\text{RF}}|_{\max} = |v^Q_{\text{RF}}|_{\max} \pm \Delta V_{\text{RF}}$ and $\angle v^I_{\text{RF}} = \angle v^Q_{\text{RF}} \pm \Delta\phi_{\text{RF}}$, and
2) $v^I_{\text{LO}}(t)$ and $v^Q_{\text{LO}}(t)$ are related by $|v^I_{\text{LO}}|_{\max} = |v^Q_{\text{LO}}|_{\max} \pm \Delta V_{\text{LO}}$ and $\angle v^I_{\text{LO}} = \angle v^Q_{\text{LO}} + \frac{\pi}{2} \pm \Delta\phi_{\text{LO}}$.
   (a) Determine the I and Q output signals and $v(t) = v_I(t) + jv_Q(t)$. Sketch $v(t)$ in the complex I–Q plane. Discuss the results.
   (b) Assume, $\Delta V_{\text{RF}} = \pm 0.9\,\text{V}$, $\Delta V_{\text{LO}} = \pm 0.85\,\text{V}$, $\Delta\phi_{\text{RF}} = \pm 3°$, $\Delta\phi_{\text{LO}} = \pm 5°$. How many possible combinations of these parameters can be used for $v(t)$? Plot $v(t)$ in the complex I–Q plane assuming: (i) positive ΔV_{RF} and $\Delta\phi_{\text{RF}}$ and negative ΔV_{LO} and $\Delta\phi_{\text{LO}}$; (ii) negative ΔV_{RF} and $\Delta\phi_{\text{RF}}$ and positive ΔV_{LO} and $\Delta\phi_{\text{LO}}$; and (iii) positive ΔV_{RF} and negative $\Delta\phi_{\text{RF}}$, ΔV_{LO}, and $\Delta\phi_{\text{LO}}$. Comment on the results.

16.24 Consider a practical quadrature mixer with nonidentical I and Q mixers and assume negligible DC offset. The I and Q output signals of this mixer have different amplitude and phase as

$$v_I(t) = V_I \cos[\phi(t)]$$

and

$$v_Q(t) = V_Q \sin[\phi(t) \pm \Delta\phi]$$

Assume $f = 10$ GHz and the initial phase $\phi_o = 0$. Plot $v(t) = v_I(t) + jv_Q(t)$ in the complex I–Q plane and discuss the results for each of the following cases and comment on the overall responses:

a) $V_I = V_Q$ and $\Delta\phi = 0°$. This is the case of an ideal quadrature mixer with equal-amplitude, 90° out-of-phase LO signals and equal-amplitude, equal-phase RF signals.
b) $V_I = 0.9V_Q$ and $\Delta\phi = 5°$.
c) $V_I = 0.7V_Q$ and $\Delta\phi = 0°$.
d) $V_Q = 0.7V_I$ and $\Delta\phi = 0°$.
e) $V_I = V_Q$ and $\Delta\phi = 10°$.
f) $V_I = V_Q$ and $\Delta\phi = 50°$.
g) $V_I = V_Q$ and $\Delta\phi = 90°$.
h) $V_I = V_Q$ and $\Delta\phi = 180°$.
i) $V_I = V_Q$ and $\Delta\phi = -50°$.
j) $V_I = 0.4V_Q$ and $\Delta\phi = 50°$.

APPENDIX

RFIC DESIGN EXAMPLE: MIXER

The design of radio frequency integrated circuits (RFICs) is relatively complicated, involving many steps as mentioned in Chapter 15, from the design of constituent circuit elements, initial circuit design, analysis and simulation to final circuit design, analysis, simulation and optimization, layout, post-layout simulation and optimization, and final layout. In this appendix, the design of an RFIC double-balanced mixer employing a Gilbert cell, which is perhaps one of the most complicated RFICs, is presented as a way to illustrate the design process of RFICs. It is particularly noted that the main objective of the mixer design presented in this appendix is to show a design process for mixers, in particular, and RFICs in general. Attempts are not made to improve the mixer topology or to achieve optimum performance for the circuit.

A1.1 CIRCUIT DESIGN SPECIFICATIONS AND GENERAL DESIGN INFORMATION

Table A1.1 shows the design specifications for the mixer. The mixer is designed to generate a 24.5-GHz radio frequency (RF) signal from a 21-GHz intermediate frequency (IF) and 3.5-GHz local oscillator (LO) signals. It is noted that the design specifications for mixers are different for their intended applications. In this design example, the mixer is used as an up-converter to generate an RF signal. Therefore, it is designed to have high linearity, high harmonic and spur rejection, and reasonable noise figure.

The mixer is designed using Jazz SBC18H2 BiCMOS process [1]. The SiGe HBT transistors used in the mixer have a cut-off frequency of 200 GHz, which is well suited for 24.5 GHz operations. The design is carried out using both circuit simulator (Cadence [2]) and EM simulator (IE3D [3]). Figures obtained from the simulations are shown un-altered to keep the context close to actual designs conducted by RFIC designers as much as possible. All on-chip inductors, vias, and interconnects are simulated using IE3D and the resultant *S*-parameters and/or equivalent-circuit models are imported into Cadence and used in the mixer design, simulation and optimization. The mixer layout is done using Cadence.

A1.2 MIXER DESIGN

Figure A1.1 shows the schematic of the mixer employing a double-balanced Gilbert cell [4]. LO leakage is a critical problem in the mixer design. To lessen this problem, 3.5-GHz is chosen for the LO signal, which

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

TABLE A1.1. Mixer Design Specifications

IF frequency (GHz)	21	Noise figure (dB)	<8
LO frequency (GHz)	3.5	Input 1-dB power compression $P_{\text{in,1dB}}$ (dBm)	>0
RF frequency (GHz)	24.5	LO power (dBm)	<0
Power conversion gain (dB)	>20	LO–RF isolation (dB)	>40
Side-band Suppression (dB)	>40	IF–RF isolation (dB)	>15

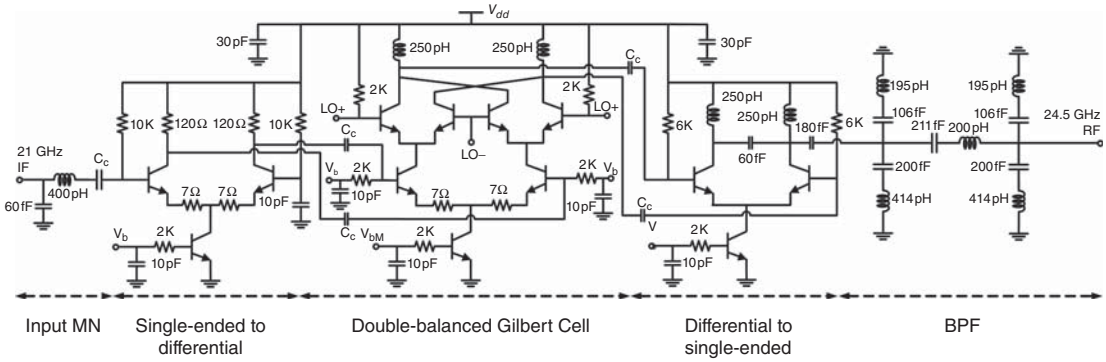


Figure A1.1. Mixer schematic.

is quite far away from the output RF frequency of 24.5 GHz, hence facilitating the removal of the 3.5-GHz LO signal using a low Q band-pass filter (BPF) to improve the purity of the output signal. As shown in Figure A1.1, the mixer consists of an input matching network (MN), (input) single-ended-to-differential active balun, a double-balanced Gilbert cell, an (output) differential-to-single-ended active balun, and a BPF. The input balun is used to accommodate single-ended input IF signals to facilitate on-wafer measurement. The double-balanced Gilbert cell is chosen instead of the single-balanced counterpart to enhance the IF-to-RF and LO-to-RF isolation and even-order suppression. The LC tanks at the outputs of the Gilbert cell are used to boost the gain and form band-pass responses. The output balun is used to convert the differential signals at the output of the Gilbert cell to the single-ended RF signals to be used for subsequent single-ended components. This active balun also increases the gain and output power of the mixer, thus also acting as a (differential) amplifier. The BPF passes the 24.5-GHz output signal while suppressing the LO signal, lower-sideband signal at 17.5-GHz, and cross-channel coupling signal at 35 GHz. It is noted that, when used in systems, the mixer should be preceded with a BPF to reject the image signal at 28 GHz.

At high frequencies, particularly in the high RF range, the effects of ground pads, vias and interconnects, including vias (even very short ones) used for connecting the transistor terminals from one metal layer to another one, on circuit performance are substantial. It is therefore important to conduct EM simulations for these elements and include the simulated results (e.g., S-parameters) in the circuit simulation. For instance, the vias used for connecting the transistor terminals from Metal 1 or 2 to Metal 6 in the mixer are simulated using IE3D and their S-parameters are included in the transistor's model (i.e., considering these vias as part of the transistor) to take into account the vias' effects. It is noted that parasitic components from vias cannot be extracted from the Cadence RCX extraction feature in the current Cadence version and, hence, EM simulations are the most accurate approach.

Figure A1.2 shows the model of the transistors used in the mixer design including the S-parameters of the base, collector and emitter interconnecting vias. Large capacitors (C) and inductors (L) are used to block the unwanted DC and AC signals, respectively. It is noted that the post-layout simulation result for the mixer using this model and the EM-simulated S-parameters of other passive components, including interconnects, are more accurate and reliable than the post-layout simulation using the models for the transistors and passive elements from the PDK, and those extracted by the Cadence RCX extraction.

The following describes the design of the mixer's constituent components.

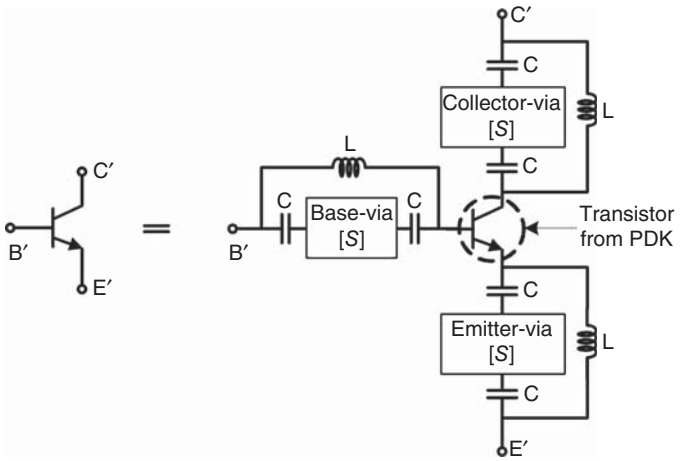


Figure A1.2. Transistor model including S-parameters of interconnecting vias

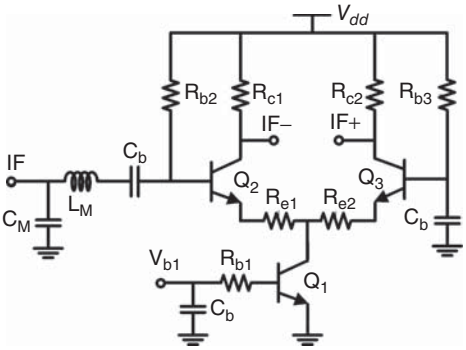


Figure A1.3. Single-ended to differential active balun. IF is the input, and IF+ and IF− are the output.

A1.2.1 Single-Ended to Differential Input Active Balun

An active balun, as shown in Figure A1.3, is designed to convert a 21-GHz input IF signal from single-ended to differential. It consists of a differential pair (Q_2 and Q_3) and a current source (Q_1). One input of the differential pair is single-ended and another is grounded through a bypass capacitor. Resistor loads of R_{c1} and R_{c2} are used instead of inductors to save the area. The amplitude and phase balance of this input balun is strongly affected by the parasitic capacitors existing at the common node, hence by the size of transistor Q_1 of the current source. Therefore, a trade-off between the amplitude and phase balance and the gain of the input balun is needed. A small size of 6 μm is chosen for transistor Q_1 to obtain a low parasitic capacitance while providing a DC current of 6 mA to the differential pair. The input MN, consisting of L_M and C_M, is designed to match the input of the input balun to the 50-Ω source. The values for R_{c1} and R_{c2} are chosen so that the DC voltage dropped in R_c is 0.6 V, hence providing enough headroom voltages for transistors Q_1, Q_2, and Q_3. Transistors Q_2 and Q_3 are biased at DC currents of 3 mA, resulting in highest possible cut-off frequency (f_T) for the transistors. Their sizes are chosen as 3 μm. The degenerative resistors R_{e1} and R_{e2} are used to increase the linearity of the input balun. Table A1.2 shows the component values of the input balun.

A1.2.2 Double-Balanced Gilbert Cell

The double-balanced Gilbert mixer cell shown in Figure A1.4 is designed for high gain and linearity, while achieving a reasonable noise figure. It consists of the IF gain stage (Q_5 and Q_6), LO switching stage (Q_7, Q_8, Q_9, and Q_{10}), and current source (Q_4). The inductors L_{c1} and L_{c2} are used to resonate all parasitic capacitors at the outputs of the mixer and form the tuned loads. Compared to using resistor loads, the tuned load boosts

TABLE A1.2. Element Values of the Input Active Balun

Circuit element	Value	Element	Value	Element	Emitter area
R_{c1}, R_{c2}	200 Ω	C_b	5 pF	Q_1	0.15×6 μm^2
R_{b1}	2 kΩ	V_{dd}	1.8 V	Q_2, Q_3	0.15×3 μm^2
R_{b2}, R_{b3}	12 kΩ	V_{b1}	1.3 V		
L_M	400 pH	C_M	60 pF		

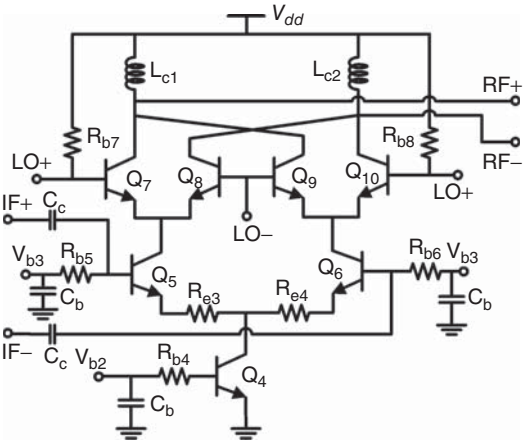


Figure A1.4. Double-balanced Gilbert mixer.

the gain, increases the headroom for transistors, hence increasing the linearity of the mixer, and partly reject the unwanted output frequency components through its band-pass characteristic.

As seen in Eq. (13.40), the gain of the mixer is proportional to the transconductance g_m of the transistors in the (IF) gain stage and the equivalent resistance R_L of the two parallel loading inductors L_{c1} and L_{c2}. The equivalent load R_L can be expressed as $R_L = Q\omega L_c$, where Q is the quality factor of the inductor ($L_c = L_{c1} = L_{c2}$). Inductors L_{c1} and L_{c2} are designed on the thick topmost metal layer M6 and optimized using IE3D to achieve a high quality factor of 16 which helps improve the gain for the mixer.

Large sizes and high bias currents for transistors Q_5 and Q_6 in the IF gain stage are desired for high gain, high linearity, and low noise for the mixer. As there is always a trade-off between the gain and power consumption for mixer design, an iterative optimization is performed to choose proper sizes and bias currents for the transistors. As a result of this analysis, transistors having configuration of one emitter, two base and two collector contacts (CBEBC) with an emitter area of 0.15×8 μm^2 are chosen for Q_5 and Q_6. They are biased at a current of 6.4 mA for maximum f_T. Degenerative resistors R_{e2} of 10 Ω are used to further improve the linearity of the IF stage. Since the noise contribution of the mixer mainly comes from the IF stage as long as the transistors in the LO stage are switched abruptly, large-size Q_5 and Q_6 having two base contacts would lower the thermal noise contribution due to small base resistors.

Transistor Q_4 in the current source has a size of 0.15×10 μm^2 and is biased at 12.8 mA to provide sufficient current for the mixer cell. The small size of Q_4 results in small parasitic capacitance and hence increasing the common-mode rejection for the mixer.

The sizes of the transistors (Q_7, Q_8, Q_9, and Q_{10}) in the LO switching stage are chosen to be small with multibase and collector configuration for fast switching. Smaller-size transistors result in lower LO power needed at an expense of lower linearity due to lower current-handling capability. Fast switching for the LO switching stage results in improved linearity and reduced noise contribution from the LO switching transistors. Simulations show that a transistor size of 0.15×4 μm^2 gives lowest noise figure with a low LO power of −4 dBm. All the bypass capacitors C_b's have values of 5 pF. The coupling capacitors C_c's in the IF stage are 200 fF. The mixer core consumes a 12.8-mA DC current from a supply voltage of 1.8 V.

TABLE A1.3. Circuit Element Values of the Gilbert Mixer Cell

Element	Value	Element	Value	Transistor	Emitter area
L_{c1}	250 Ω	C_c	200 fF	Q_4	0.15×10 μm^2
$R_{b3,}$ R_{b4}	2 kΩ	V_{dd}	1.8 V	Q_5, Q_6	0.15×8 μm^2
R_{b5}	1 kΩ	R_{e2}	10 Ω	Q_7, Q_8, Q_9, Q_{10}	0.15×4 μm^2

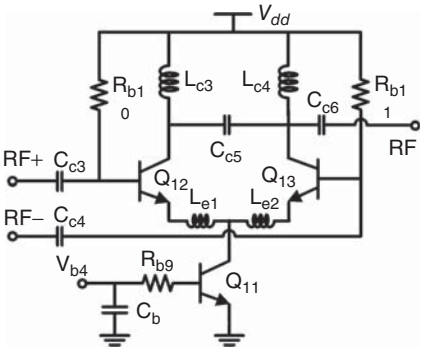


Figure A1.5. Differential to single-ended active balun. RF+ and RF– are the input and RF is the output.

Table A1.3 summarizes the values of the elements used in the Gilbert mixer cell.

A1.2.3 Differential to Single-Ended Output Active Balun

A differential to single-ended active balun, as shown in Figure A1.5, is used at the output of the mixer core to convert differential to single-ended signals, as well as to increase the mixer's conversion gain and output power. It functions indeed as a differential amplifier. Transistors Q_{12} and Q_{13} are designed with a large emitter area of 0.15×8 μm^2 and biased at a DC current of 7 mA for high gain and linearity. Transistor Q_{11} biased at 14 mA functions as a DC current source for this output balun. The load inductors L_{c3} and L_{c4} are optimized to tune all the parasitic capacitors at the collectors of Q_{12} and Q_{13}, as well as form the LC tanks for signal selectivity and side-band and LO suppression. L_{c3}, L_{c4} are also used with capacitor C_{c3} to form a current combiner functioning as a passive balun to convert the differential signal to single-ended signal. Capacitors C_{c3}, C_{c4}, and C_{c6} are used to provide the matching at the input and output of the output balun, respectively. Degenerative inductors L_{e1} and L_{e2} are used to increase the linearity of the output balun. There is a trade-off between the gain and linearity of the output balun due to values of the degenerative inductors. The values of L_{e1} and L_{e2} are chosen to be 25 pH through several iterations. These inductors are implemented using simple metal traces (lines). The values of the elements used in the output balun are shown in Table A1.4.

A1.2.4 Band-Pass Filter

Figure A1.6 shows the schematic of the BPF used at the output of the differential to single-ended output active balun to select the 24.5-GHz RF signal and suppress the lower sideband at 17.5 GHz and the cross channel

TABLE A1.4. Circuit Element Values of the Output Active Balun

Element	Value	Element	Value	Transistor	Emitter area
$L_{c3,}$ L_{c4}	250 Ω	C_{c3}, C_{c4}	200 fF	Q_{11}	0.15×10 μm^2
$R_{b10,}$ R_{b11}	6 KΩ	C_5	60 fF	Q_{12}, Q_{13}	0.15×8 μm^2
R_{b9}	2 KΩ	L_{e1}, L_{e2}	25 pH	V_{dd}	1.8 V

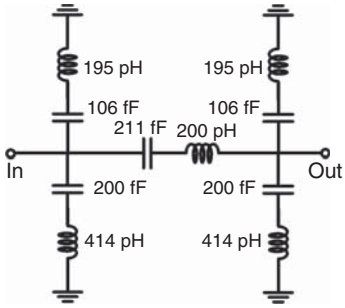


Figure A1.6. Band-pass filter.

leakage at 35 GHz. Simulation results show that the BPF exhibits an insertion loss of 1.8 dB at 24.5 GHz and rejection of 34 and 37 dB at 17.5 and 35 GHz, respectively.

A1.3 MIXER OPTIMIZATION AND LAYOUT

The design of mixers at high frequencies in the RF regime is not straight forward as there are many trade-off parameters in the design and EM effects resulting from layouts. Mixer design, hence, requires various iterative optimization and layout and, in general, is performed in three stages: initial, intermediate, and final stages.

In the initial design, ideal resistors, capacitors, and inductors as well as transistors from the PDK of the employed process can be used to design and optimize the mixer to meet its required specifications. Optimization of the mixer requires understanding of trade-off between the mixer parameters including gain, linearity, noise figure, power consumption, LO power level, and actual implementation of passive components (which affect the chip area).

In the intermediate design stage, after the optimization of the mixer using ideal components and transistor models from the PDK in the initial design, the transistors are replaced with the model described in Figure A1.2, the designed ideal capacitors are replaced with real capacitors from the PDK, and the designed ideal inductors are replaced by their *S*-parameters obtained from the EM simulated design. As expected, the performance of the resultant mixer circuit is normally different from that of the mixer based on ideal components in the initial design. Therefore, further optimization is needed for the mixer to meet the required specifications. A layout for the mixer is then prepared.

In the final design stage, the interconnects resulting from the mixer layout are modeled accurately using IE3D and included in the mixer simulation. This step, requiring the use of EM simulators, is important for the mixer design. During the layout process, some already designed components, particularly inductors, may need to be modified due to constraints in the layout. The altered components then need to be simulated using an EM simulator and included in the mixer simulation. Further optimization of the mixer may be needed. Completion of the mixer design is done in an interactive way, requiring several iterations going back and forth among the layout and EM and circuit simulations. During this process, the mixer layout also needs to be optimized to improve the mixer performance by various ways, such as reduction of parasitics, better layout symmetry, etc.

Figure A1.7 shows the final layout of the entire mixer. The layout is done as symmetrical as possible in order to preserve the symmetry needed for the Gilbert mixer core to work properly. The size of the whole mixer is around 1.2 mm $\times$ 0.9 mm, while that of the mixer core is about 600 μm $\times$ 350 μm. As can been seen, most of the chip area is occupied by the passive components (mainly inductors and interconnects). Metal 6 (the topmost metal layer) is used for the IF, LO, and RF signals. A robust ground consisting of six stacked metal layers is placed around the mixer. Metal 5 and metal 6 are used for the interconnects at the LO and RF port. Poly blocking layers are inserted underneath all the inductors to prevent the poly dummies added during fabrication from affecting the inductor performance.

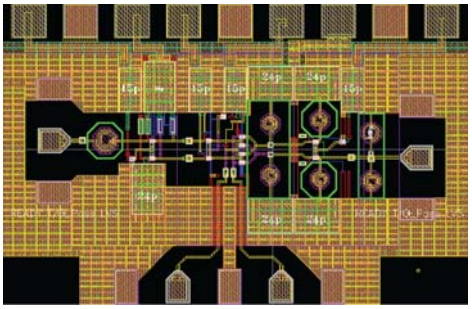


Figure A1.7. Mixer layout.

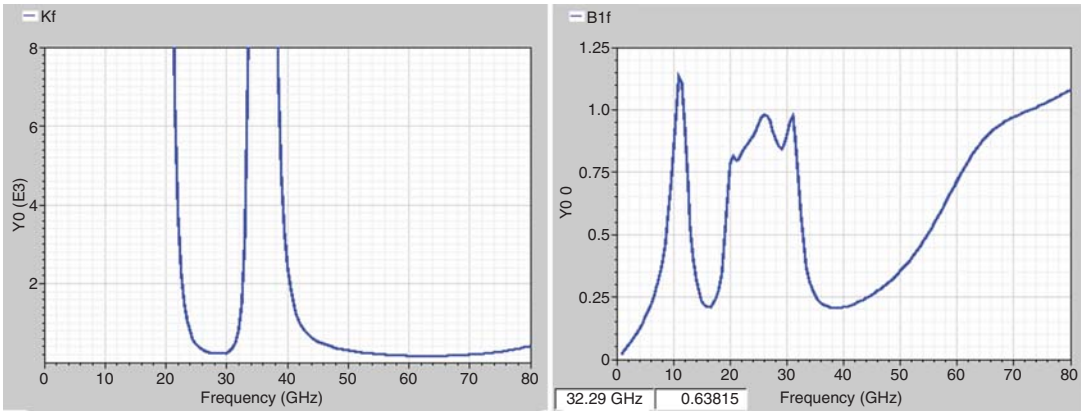


Figure A1.8. Mixer's stability factors K (a) and B1 (b). Input power is −40 dBm.

A1.4 SIMULATION RESULTS

A1.4.1 Stability

Stability is an important issue and should be the first to check in the mixer design. Stability can be evaluated using *S*-parameter simulation with the LO port terminated. When the LO port is terminated, the mixer can be considered a two-port amplifier with the IF and RF ports being the input and output ports, respectively, and its stability is checked as an amplifier. Simulations done in Cadence with different loads for the LO port show that the mixer is unconditionally stable from DC to 80 GHz. Figure A1.8 shows the results with the LO port shorted, showing that the mixer's stability factors K and B_1 given in Eq. (10.4) are larger than 1 and 0, respectively.

A1.4.2 Return Loss

Figure A1.9 shows the simulated input (IF) and output (RF) return losses of the mixer. The input and output return losses are more than 22 and 10 dB at the IF and RF frequencies of 21 and 24.5 GHz, respectively.

A1.4.3 Conversion Gain

The mixer's conversion gain versus LO power is simulated using the harmonic-balance tool available in Cadence to determine the optimum LO power that provides maximum gain. The input IF signal is −40 dBm at 21 GHz. The LO signal is fixed at 3.5 GHz and its power is swept from −10 to 3 dBm in 1-dBm steps. The conversion gain is calculated as the ratio of the 24.5-GHz RF output power and 21-GHz IF input power. Figure A1.10(a) shows the simulated conversion gains versus LO power. The results show that a maximum

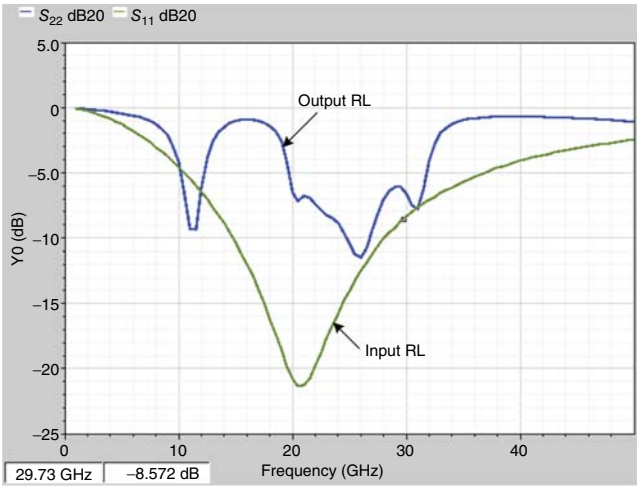


Figure A1.9. Mixer's input (IF) and output (RF) return loss (RL).

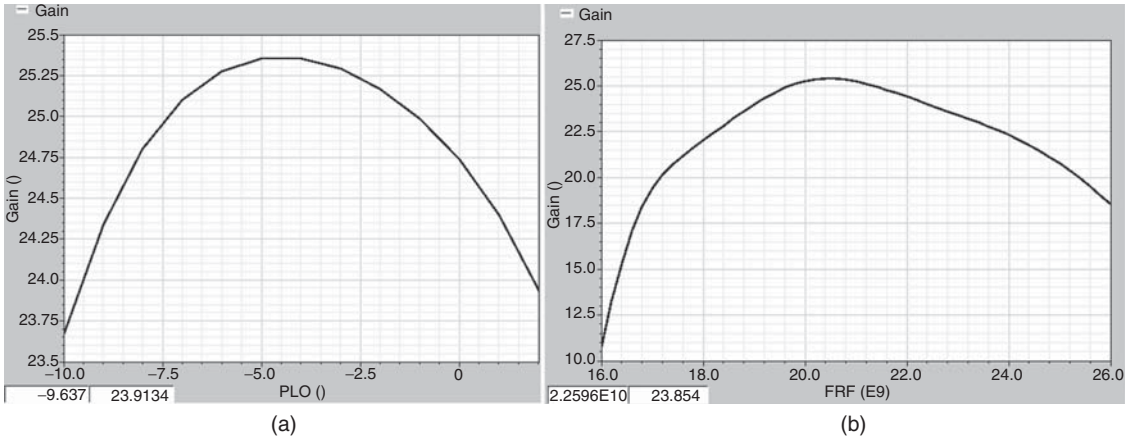


Figure A1.10. Mixer's conversion gain versus LO power (a) and IF frequency (b).

gain of 25.3 dB is obtained for the mixer with an LO power of −4 dBm. This LO power is used for other simulations. The small LO power is achieved by using small-size transistors in the LO stage.

The simulation for the mixer's conversion gain versus IF frequency is carried out using the harmonic-balance analysis to characterize the frequency response o the mixer with respect to the IF frequency. The IF power is set at −40 dBm and the IF frequency is swept from 16 to 26 GHz. The LO frequency and power are fixed at 3.5 GHz and −4 dBm, respectively. Figure A1.10(b) shows the simulated conversion gains versus IF frequency. As can be seen, the mixer exhibits a maximum gain of 25.5 dB at an IF frequency of 21 GHz.

A1.4.4 Noise Figure

Figure A1.11 shows the simulated noise figure of the complete mixer versus the input IF frequency using the harmonic-balance analysis. The LO frequency and power are set at 3.5 GHz and −4 dBm, respectively, and the IF signal at −40 dBm is swept from 15 to 30 GHz. At 21 GHz, the noise figure is 7.6 dB which, although not very good, is a reasonable value for the mixer used as an up-converter and meets the design requirement for the noise figure.

A1.4.5 Other Mixer Performance

Figure A1.12 shows the simulated gain, RF output power (at 24.5 GHz), IF-to-RF isolation, LO-to-RF isolation, and lower-sideband suppression versus the IF input power. In these simulations, the LO frequency and

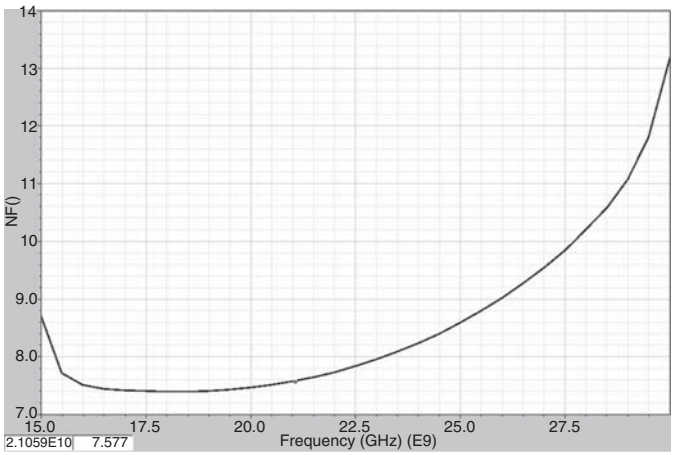


Figure A1.11. Mixer's noise figure versus IF frequency.

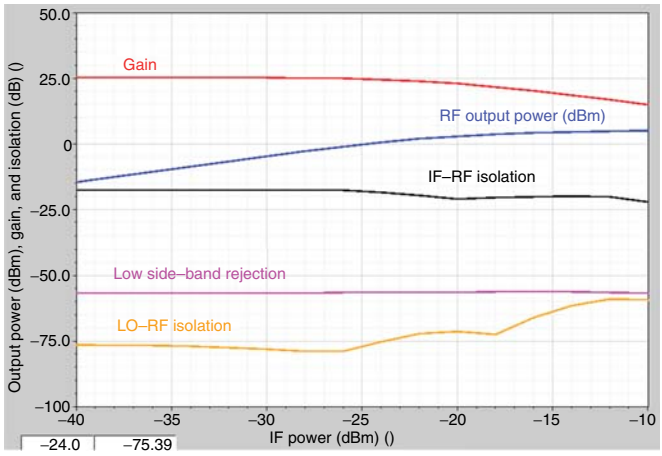


Figure A1.12. Mixer's gain, output power, isolation and lower-sideband suppression.

power are fixed at 3.5 GHz and −4 dBm, respectively, and the IF signal is set at 21 GHz, while its power is swept from −40 to −10 dBm. The simulation results show that the mixer exhibits a maximum gain of 25.3 dB, 1-dB output power compression ($P_{\text{out-1 dB}}$) of 1.28 dBm, lower-sideband suppression of 56 dB, IF−RF isolation higher than 17 dB, and LO−RF isolation more than 60 dB.

Figure A1.13 displays the output signal spectrum of the mixer with 3.5-GHz LO signal of −4-dBm and 21-GHz IF signal of −23 dBm. The spectrum shows a 24.5-GHz RF tone with power of 1.28 dBm and harmonic and inter-modulation products suppressed by more than 25 dB from the 24.5-GHz signal. By changing the bias voltage V_{b2} of the Gilbert cell, indicated in Figure A1.4, from 0.78 to 1.3 V, the gain of the mixer changes from −10 to +25.3 dB, which signifies that the designed mixer can work as a variable-gain mixer with a substantial gain-tuning range.

The performance of the mixer is summarized and compared with the design specifications in Table A1.5. It shows that the designed mixer meets the required specifications.

A1.5 MEASURED RESULTS

Figure A1.14 shows the mixer fabricated on the Jazz SBC18H2 BiCMOS process. The fabricated mixer was measured on-wafer using a Rhode & Schwarz vector network analyzer and Cascade probe station. The short-open-load-thru calibration method along with Microtech's impedance standard substrate standards

TABLE A1.5. Simulated Performance versus Specifications

Parameter	Design specification	Simulated results
IF frequency (GHz)	21	21
LO frequency (GHz)	3.5	3.5
RF frequency (GHz)	24.5	24.5
Power conversion gain (dB)	20	25.3
Side-band suppression (dB)	40	56
Noise figure (dB)	8	7.6
$P_{\text{in,1dB}}$ (dBm)	>0	1.28
LO power (dBm)	<0	−4
LO–RF isolation (dB)	>40	60
IF–RF isolation (dB)	>15	17

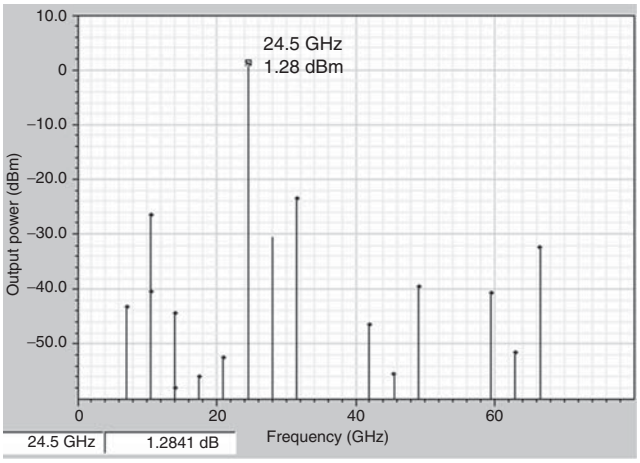


Figure A1.13. Mixer's output spectrum.

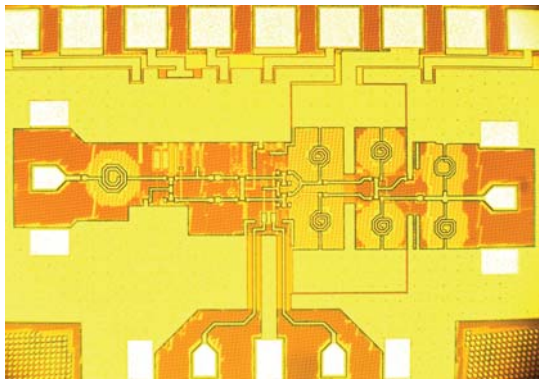


Figure A1.14. Mixer's microphotograph.

was used. The mixer's stability was first confirmed by calculating the stability factors K (> 1) and B_1 (> 0) using the measured S-parameters. The mixer consumes a current of 40 mA from a power supply of 1.8 V. The measured and simulated results agree reasonably well, validating our design. Some of the measured results are described as follows.

Figure A1.15 shows the simulated and measured input and output return losses versus IF and RF frequencies. Figure A1.16 shows the simulated and measured conversion gain and output power versus IF input

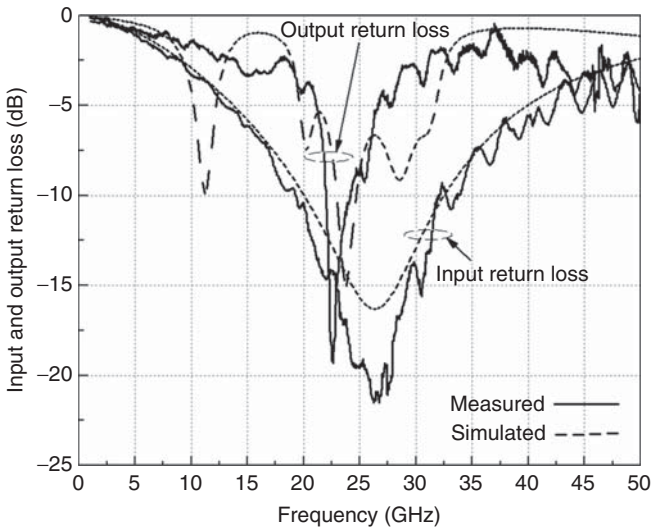


Figure A1.15. Simulated and measured input and output return losses of mixer.

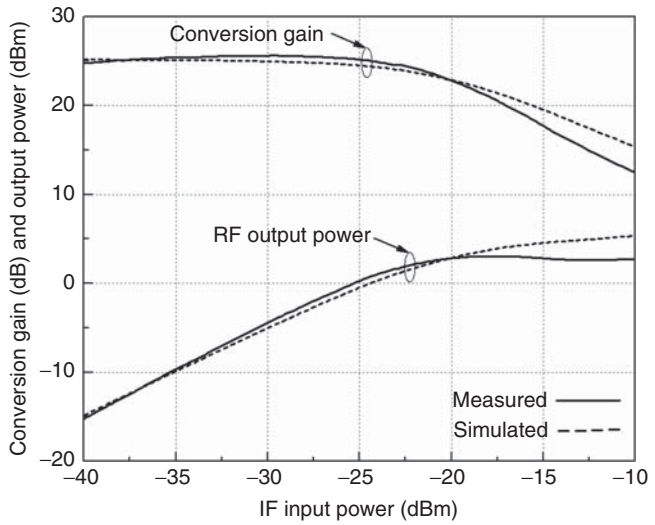


Figure A1.16. Simulated and measured conversion gain and output power of mixer.

power, where the LO frequency and power are fixed at 3.5 GHz and −2 dBm, respectively, while the IF frequency is set to 21 GHz.

REFERENCES

1. Jazz Semiconductor, 4321 Jamboree Road, Newport Beach, CA 92660.
2. Cadence Design Systems, Inc., 2655 Seely Avenue, San Jose, CA 95134.
3. IE3D, Zeland Software, Inc., Fremont, CA, 94538.
4. B. Gilbert, "A Precise Four-Quadrant Multiplier with Subnanosecond Response," *IEEE J. Solid-State Circuits*, Vol. 3, pp. 365–373, Dec. 1968.

INDEX

Note: Page numbers followed by f refer to Figures and those by t refer to Tables

Radio-Frequency Integrated-Circuit Engineering, First Edition. Cam Nguyen.

D

M

P

S

T

WILEY SERIES IN MICROWAVE AND OPTICAL ENGINEERING

KAI CHANG, Editor
Texas A&M University

FIBER-OPTIC COMMUNICATION SYSTEMS, Fourth Edition • *Govind P. Agrawal*

ASYMMETRIC PASSIVE COMPONENTS IN MICROWAVE INTEGRATED CIRCUITS • *Hee-Ran Ahn*

COHERENT OPTICAL COMMUNICATIONS SYSTEMS • *Silvello Betti, Giancarlo De Marchis, and Eugenio Iannone*

PHASED ARRAY ANTENNAS: FLOQUET ANALYSIS, SYNTHESIS, BFNs, AND ACTIVE ARRAY SYSTEMS • *Arun K. Bhattacharyya*

HIGH-FREQUENCY ELECTROMAGNETIC TECHNIQUES: RECENT ADVANCES AND APPLICATIONS • *Asoke K. Bhattacharyya*

RADIO PROPAGATION AND ADAPTIVE ANTENNAS FOR WIRELESS COMMUNICATION LINK NETWORKS: TERRESTRIAL, ATMOSPHERIC, AND IONOSPHERIC, Second Edition • *Nathan Blaunstein and Christos G. Christodoulou*

COMPUTATIONAL METHODS FOR ELECTROMAGNETICS AND MICROWAVES • *Richard C. Booton, Jr.*

ELECTROMAGNETIC SHIELDING • *Salvatore Celozzi, Rodolfo Araneo, and Giampiero Lovat*

MICROWAVE RING CIRCUITS AND ANTENNAS • *Kai Chang*

MICROWAVE SOLID-STATE CIRCUITS AND APPLICATIONS • *Kai Chang*

RF AND MICROWAVE WIRELESS SYSTEMS • *Kai Chang*

RF AND MICROWAVE CIRCUIT AND COMPONENT DESIGN FOR WIRELESS SYSTEMS • *Kai Chang, Inder Bahl, and Vijay Nair*

MICROWAVE RING CIRCUITS AND RELATED STRUCTURES, Second Edition • *Kai Chang and Lung-Hwa Hsieh*

MULTIRESOLUTION TIME DOMAIN SCHEME FOR ELECTROMAGNETIC ENGINEERING • *Yinchao Chen, Qunsheng Cao, and Raj Mittra*

DIODE LASERS AND PHOTONIC INTEGRATED CIRCUITS, Second Edition • *Larry Coldren, Scott Corzine, and Milan Masanovic*

EM DETECTION OF CONCEALED TARGETS • *David J. Daniels*

RADIO FREQUENCY CIRCUIT DESIGN • *W. Alan Davis and Krishna Agarwal*

RADIO FREQUENCY CIRCUIT DESIGN, Second Edition • *W. Alan Davis*

FUNDAMENTALS OF OPTICAL FIBER SENSORS • *Zujie Fang, Ken K. Chin, Ronghui Qu, and Haiwen Cai*

MULTICONDUCTOR TRANSMISSION-LINE STRUCTURES: MODAL ANALYSIS TECHNIQUES • *J. A. Brandão Faria*

PHASED ARRAY-BASED SYSTEMS AND APPLICATIONS • *Nick Fourikis*

SOLAR CELLS AND THEIR APPLICATIONS, Second Edition • *Lewis M. Fraas and Larry D. Partain*

FUNDAMENTALS OF MICROWAVE TRANSMISSION LINES • *Jon C. Freeman*

OPTICAL SEMICONDUCTOR DEVICES • *Mitsuo Fukuda*

MICROSTRIP CIRCUITS • *Fred Gardiol*

HIGH-SPEED VLSI INTERCONNECTIONS, Second Edition • *Ashok K. Goel*

FUNDAMENTALS OF WAVELETS: THEORY, ALGORITHMS, AND APPLICATIONS, Second Edition • *Jaideva C. Goswami and Andrew K. Chan*

HIGH-FREQUENCY ANALOG INTEGRATED CIRCUIT DESIGN • *Ravender Goyal (ed.)*

RF AND MICROWAVE TRANSMITTER DESIGN • *Andrei Grebennikov*

ANALYSIS AND DESIGN OF INTEGRATED CIRCUIT ANTENNA MODULES • *K. C. Gupta and Peter S. Hall*

PHASED ARRAY ANTENNAS, Second Edition • *R. C. Hansen*

STRIPLINE CIRCULATORS • *Joseph Helszajn*

THE STRIPLINE CIRCULATOR: THEORY AND PRACTICE • *Joseph Helszajn*

LOCALIZED WAVES • *Hugo E. Hernández-Figueroa, Michel Zamboni-Rached, and Erasmo Recami (eds.)*

MICROSTRIP FILTERS FOR RF/MICROWAVE APPLICATIONS, Second Edition • *Jia-Sheng Hong*

MICROWAVE APPROACH TO HIGHLY IRREGULAR FIBER OPTICS • *Huang Hung-Chia*

NONLINEAR OPTICAL COMMUNICATION NETWORKS • *Eugenio Iannone, Francesco Matera, Antonio Mecozzi, and Marina Settembre*

FINITE ELEMENT SOFTWARE FOR MICROWAVE ENGINEERING • *Tatsuo Itoh, Giuseppe Pelosi, and Peter P. Silvester (eds.)*

INFRARED TECHNOLOGY: APPLICATIONS TO ELECTROOPTICS, PHOTONIC DEVICES, AND SENSORS • *A. R. Jha*

SUPERCONDUCTOR TECHNOLOGY: APPLICATIONS TO MICROWAVE, ELECTRO-OPTICS, ELECTRICAL MACHINES, AND PROPULSION SYSTEMS • *A. R. Jha*

TIME AND FREQUENCY DOMAIN SOLUTIONS OF EM PROBLEMS USING INTEGTRAL EQUATIONS AND A HYBRID METHODOLOGY • *B. H. Jung, T. K. Sarkar, S. W. Ting, Y. Zhang, Z. Mei, Z. Ji, M. Yuan, A. De, M. Salazar-Palma, and S. M. Rao*

OPTICAL COMPUTING: AN INTRODUCTION • *M. A. Karim and A. S. S. Awwal*

INTRODUCTION TO ELECTROMAGNETIC AND MICROWAVE ENGINEERING • *Paul R. Karmel, Gabriel D. Colef, and Raymond L. Camisa*

MILLIMETER WAVE OPTICAL DIELECTRIC INTEGRATED GUIDES AND CIRCUITS • *Shiban K. Koul*

ADVANCED INTEGRATED COMMUNICATION MICROSYSTEMS • *Joy Laskar, Sudipto Chakraborty, Manos Tentzeris, Franklin Bien, and Anh-Vu Pham*

MICROWAVE DEVICES, CIRCUITS AND THEIR INTERACTION • *Charles A. Lee and G. Conrad Dalman*

ADVANCES IN MICROSTRIP AND PRINTED ANTENNAS • *Kai-Fong Lee and Wei Chen (eds.)*

SPHEROIDAL WAVE FUNCTIONS IN ELECTROMAGNETIC THEORY • *Le-Wei Li, Xiao-Kang Kang, and Mook-Seng Leong*

MICROWAVE NONCONTACT MOTION SENSING AND ANALYSIS • *Changzhi Li and Jenshan Lin*

COMPACT MULTIFUNCTIONAL ANTENNAS FOR WIRELESS SYSTEMS • *Eng Hock Lim and Kwok Wa Leung*

ARITHMETIC AND LOGIC IN COMPUTER SYSTEMS • *Mi Lu*

OPTICAL FILTER DESIGN AND ANALYSIS: A SIGNAL PROCESSING APPROACH • *Christi K. Madsen and Jian H. Zhao*

THEORY AND PRACTICE OF INFRARED TECHNOLOGY FOR NONDESTRUCTIVE TESTING • *Xavier P. V. Maldague*

METAMATERIALS WITH NEGATIVE PARAMETERS: THEORY, DESIGN, AND MICROWAVE APPLICATIONS • Ricardo Marqués, Ferran Martín, and Mario Sorolla

OPTOELECTRONIC PACKAGING • *A. R. Mickelson, N. R. Basavanhally, and Y. C. Lee (eds.)*

OPTICAL CHARACTER RECOGNITION • *Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada*

ANTENNAS FOR RADAR AND COMMUNICATIONS: A POLARIMETRIC APPROACH • *Harold Mott*

INTEGRATED ACTIVE ANTENNAS AND SPATIAL POWER COMBINING • *Julio A. Navarro and Kai Chang*

ANALYSIS METHODS FOR RF, MICROWAVE, AND MILLIMETER-WAVE PLANAR TRANSMISSION LINE STRUCTURES • *Cam Nguyen*

LASER DIODES AND THEIR APPLICATIONS TO COMMUNICATIONS AND INFORMATION PROCESSING • *Takahiro Numai*

FREQUENCY CONTROL OF SEMICONDUCTOR LASERS • *Motoichi Ohtsu (ed.)*

INVERSE SYNTHETIC APERTURE RADAR IMAGING WITH MATLAB ALGORITHMS • *Caner Özdemir*

SILICA OPTICAL FIBER TECHNOLOGY FOR DEVICE AND COMPONENTS: DESIGN, FABRICATION, AND INTERNATIONAL STANDARDS • *Un-Chul Paek and Kyunghwan Oh*

WAVELETS IN ELECTROMAGNETICS AND DEVICE MODELING • *George W. Pan*

OPTICAL SWITCHING • Georgios Papadimitriou, Chrisoula Papazoglou, and Andreas S. Pomportsis

MICROWAVE IMAGING • *Matteo Pastorino*

ANALYSIS OF MULTICONDUCTOR TRANSMISSION LINES • *Clayton R. Paul*

INTRODUCTION TO ELECTROMAGNETIC COMPATIBILITY, Second Edition • *Clayton R. Paul*

ADAPTIVE OPTICS FOR VISION SCIENCE: PRINCIPLES, PRACTICES, DESIGN AND APPLICATIONS • *Jason Porter, Hope Queener, Julianna Lin, Karen Thorn, and Abdul Awwal (eds.)*

ELECTROMAGNETIC OPTIMIZATION BY GENETIC ALGORITHMS • *Yahya Rahmat-Samii and Eric Michielssen (eds.)*

INTRODUCTION TO HIGH-SPEED ELECTRONICS AND OPTOELECTRONICS • *Leonard M. Riaziat*

NEW FRONTIERS IN MEDICAL DEVICE TECHNOLOGY • *Arye Rosen and Harel Rosen (eds.)*

ELECTROMAGNETIC PROPAGATION IN MULTI-MODE RANDOM MEDIA • *Harrison E. Rowe*

ELECTROMAGNETIC PROPAGATION IN ONE-DIMENSIONAL RANDOM MEDIA • *Harrison E. Rowe*

HISTORY OF WIRELESS • *Tapan K. Sarkar, Robert J. Mailloux, Arthur A. Oliner, Magdalena Salazar-Palma, and Dipak L. Sengupta*

PHYSICS OF MULTIANTENNA SYSTEMS AND BROADBAND PROCESSING • *Tapan K. Sarkar, Magdalena Salazar-Palma, and Eric L. Mokole*

SMART ANTENNAS • *Tapan K. Sarkar, Michael C. Wicks, Magdalena Salazar-Palma, and Robert J. Bonneau*

NONLINEAR OPTICS • *E. G. Sauter*

APPLIED ELECTROMAGNETICS AND ELECTROMAGNETIC COMPATIBILITY • *Dipak L. Sengupta and Valdis V. Liepa*

COPLANAR WAVEGUIDE CIRCUITS, COMPONENTS, AND SYSTEMS • *Rainee N. Simons*

ELECTROMAGNETIC FIELDS IN UNCONVENTIONAL MATERIALS AND STRUCTURES • *Onkar N. Singh and Akhlesh Lakhtakia (eds.)*

ANALYSIS AND DESIGN OF AUTONOMOUS MICROWAVE CIRCUITS • *Almudena Suárez*

ELECTRON BEAMS AND MICROWAVE VACUUM ELECTRONICS • *Shulim E. Tsimring*

FUNDAMENTALS OF GLOBAL POSITIONING SYSTEM RECEIVERS: A SOFTWARE APPROACH, Second Edition • *James Bao-yen Tsui*

SUBSURFACE SENSING • *Ahmet S. Turk, A. Koksal Hocaoglu, and Alexey A. Vertiy (eds.)*

RF/MICROWAVE INTERACTION WITH BIOLOGICAL TISSUES • *André Vander Vorst, Arye Rosen, and Youji Kotsuka*

InP-BASED MATERIALS AND DEVICES: PHYSICS AND TECHNOLOGY • *Osamu Wada and Hideki Hasegawa (eds.)*

COMPACT AND BROADBAND MICROSTRIP ANTENNAS • *Kin-Lu Wong*

DESIGN OF NONPLANAR MICROSTRIP ANTENNAS AND TRANSMISSION LINES • *Kin-Lu Wong*

PLANAR ANTENNAS FOR WIRELESS COMMUNICATIONS • *Kin-Lu Wong*

FREQUENCY SELECTIVE SURFACE AND GRID ARRAY • *T. K. Wu (ed.)*

PHOTONIC SENSING: PRINCIPLES AND APPLICATIONS FOR SAFETY AND SECURITY MONITORING • *Gaozhi Xiao and Wojtek J. Bock*

ACTIVE AND QUASI-OPTICAL ARRAYS FOR SOLID-STATE POWER COMBINING • *Robert A. York and Zoya B. Popović (eds.)*

OPTICAL SIGNAL PROCESSING, COMPUTING AND NEURAL NETWORKS • *Francis T. S. Yu and Suganda Jutamulia*

ELECTROMAGNETIC SIMULATION TECHNIQUES BASED ON THE FDTD METHOD • *Wenhua Yu, Xiaoling Yang, Yongjun Liu, and Raj Mittra*

SiGe, GaAs, AND InP HETEROJUNCTION BIPOLAR TRANSISTORS • *Jiann Yuan*

PARALLEL SOLUTION OF INTEGRAL EQUATION-BASED EM PROBLEMS • *Yu Zhang and Tapan K. Sarkar*

ELECTRODYNAMICS OF SOLIDS AND MICROWAVE SUPERCONDUCTIVITY • *Shu-Ang Zhou*

MICROWAVE BANDPASS FILTERS FOR WIDEBAND COMMUNICATIONS • *Lei Zhu, Sheng Sun, and Rui Li*

FUNDAMENTALS OF MICROWAVE PHOTONICS • *Vincent Jude Urick Jr., Jason Dwight McKinney, and Keith Jake Williams*

RADIO-FREQUENCY INTEGRATED-CIRCUIT ENGINEERING • *Cam Nguyen*