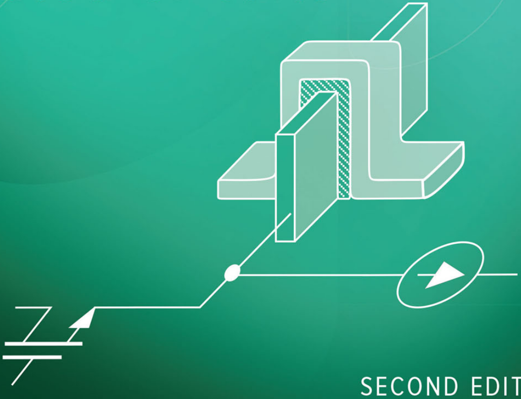
Behzad Razavi

DESIGN OF Analog CMOS Integrated Circuits



Mc Graw Hill Education

SECOND EDITION

Design of Analog CMOS Integrated Circuits

Second Edition

Behzad Razavi

Professor of Electrical Engineering
University of California, Los Angeles

Mc
Graw
Hill
Education

DESIGN OF ANALOG CMOS INTEGRATED CIRCUITS, SECOND EDITION
Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 QVS/QVS 1 0 9 8 7 6

ISBN 978-0-07-252493-2
MHID 0-07-252493-6

Senior Vice President, Products & Markets: *Kurt L. Strand*
Vice President, General Manager, Products & Markets: *Marty Lange*
Vice President, Content Design & Delivery: *Kimberly Meriwether David*
Managing Director: *Thomas Timp*
Global Brand Manager: *Raghu Srinivasan*
Director, Product Development: *Rose Koos*
Product Developer: *Vincent Bradshaw*
Marketing Manager: *Nick McFadden*
Director of Digital Content: *Chelsea Haupt, Ph. D*
Director, Content Design & Delivery: *Linda Avenarius*
Program Manager: *Faye M. Herrig*
Content Project Managers: *Heather Ervolino; Sandra Schnee*
Buyer: *Jennifer Pickel*
Content Licensing Specialists: *Lorraine Buczek* (*Text*)
Compositor: *MPS Limited*
Printer: *Quad/Graphics*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Cataloging-in-Publication Data

Razavi, Behzad.
Design of analog CMOS integrated circuits / Behzad Razavi, professor of electrical engineering, University of California, Los Angeles. – Second edition.
pages cm
Includes bibliographical references and index.
ISBN 978-0-07-252493-2 (alk. paper) – ISBN 0-07-252493-6 (alk. paper) 1. Analog CMOS integrated circuits. 2. Linear integrated circuits–Design and construction. 3. Metal oxide semiconductors, Complementary. I. Title.
TK7874.654.R39 2017
621.3815–dc23 2015035303

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

mheducation.com/highered

To the memory of my parents

Preface to the Second Edition

When I submitted proposals to publishers for the first edition of this book, they posed two questions to me: (1) What is the future demand for analog books in a digital world? and (2) Is it wise to publish a book dealing solely with CMOS? The words "analog" and "CMOS" in the book's title were both in question.

Fortunately, the book resonated with students, instructors, and engineers. It has been adopted by hundreds of universities around the world, translated to five languages, and cited 6,500 times.

While many fundamentals of analog design have not changed since the first edition was introduced, several factors have called for a second: migration of CMOS technologies to finer geometries and lower supply voltages, new approaches to analysis and design, and the need for more detailed treatments of some topics. This edition provides:

- Greater emphasis on modern CMOS technology, culminating in a new chapter, Chapter 11, on design methodologies and step-by-step op amp design in nanometer processes
- Extensive study of feedback through the approaches by Bode and Middlebrook
- A new section on the analysis of stability using Nyquist's approach—as the oft-used Bode method falls short in some common systems
- Study of FinFETs
- Sidebars highlighting important points in nanometer design
- A new section on biasing techniques
- Study of low-voltage bandgap circuits
- More than 100 new examples

Some instructors ask why we begin with square-law devices. This is for two reasons: (1) such a path serves as an intuitive entry point and provides considerable value in the analysis of amplifiers in terms of allowable voltage swings, and (2) despite their very short channel lengths, FinFETs—the devices used in 16-nm nodes and below—exhibit nearly square-law characteristics.

This book is accompanied with a solutions manual and a new set of PowerPoint slides, available at www.mhhe.com/razavi.

Behzad Razavi
July 2015

Preface to the First Edition

In the past two decades, CMOS technology has rapidly embraced the field of analog integrated circuits, providing low-cost, high-performance solutions and rising to dominate the market. While silicon bipolar and III-V devices still find niche applications, only CMOS processes have emerged as a viable choice for the integration of today's complex mixed-signal systems. With channel lengths projected to scale down to 0.05 μm, CMOS technology will continue to serve circuit design for another two decades.

Analog circuit design itself has evolved with the technology as well. High-voltage, high-power analog circuits containing a few tens of transistors and processing small, continuous-time signals have gradually been replaced by low-voltage, low-power systems comprising thousands of devices and processing large, mostly discrete-time signals. For example, many analog techniques used only ten years ago have been abandoned because they do not lend themselves to low-voltage operation.

This book deals with the analysis and design of analog CMOS integrated circuits, emphasizing fundamentals as well as new paradigms that students and practicing engineers need to master in today's industry. Since analog design requires both intuition and rigor, each concept is first introduced from an intuitive perspective and subsequently treated by careful analysis. The objective is to develop both a solid foundation and methods of analyzing circuits by inspection so that the reader learns what approximations can be made in which circuits and how much error to expect in each approximation. This approach also enables the reader to apply the concepts to bipolar circuits with little additional effort.

I have taught most of the material in this book both at UCLA and in industry, polishing the order, the format, and the content with every offering. As the reader will see throughout the book, I follow four "golden rules" in writing (and teaching): (1) I explain *why* the reader needs to know the concept that is to be studied; (2) I put myself in the reader's position and predict the questions that he/she may have while reading the material for the first time; (3) With Rule 2 in mind, I pretend to know only as much as the (first-time) reader and try to "grow" with him/her, thereby experiencing the same thought process; (4) I begin with the "core" concept in a simple (even imprecise) language and gradually add necessary modifications to arrive at the final (precise) idea. The last rule is particularly important in teaching circuits because it allows the reader to observe the evolution of a topology and hence learn both analysis and synthesis.

The text comprises 16 chapters whose contents and order are carefully chosen to provide a natural flow for both self-study and classroom adoption in quarter or semester systems. Unlike some other books on analog design, we cover only a *bare minimum* of MOS device physics at the beginning, leaving more advanced properties and fabrication details for later chapters. To an expert, the elementary device physics treatment my appear oversimplified, but my experience suggests that (a) first-time readers simply do not absorb the high-order device effects and fabrication technology before they study circuits because they do not see the relevance; (b) if properly presented, even the simple treatment proves adequate for a substantial coverage of basic circuits; (c) readers learn advanced device phenomena and processing steps much more readily *after* they have been exposed to a significant amount of circuit analysis and design.

Chapter 1 provides the reader with motivation for learning the material in this book. Chapter 2 describes basic physics and operation of MOS devices.

Chapters 3 through 5 deal with single-stage and differential amplifiers and current mirrors, respectively, developing efficient analytical tools for quantifying the behavior of basic circuits by inspection.

Chapters 6 and 7 introduce two imperfections of circuits, namely, frequency response and noise. Noise is treated at an early stage so that it "sinks in" as the reader accounts for its effects in subsequent circuit developments.

Chapters 8 through 10 describe feedback, operational amplifiers, and stability in feedback systems, respectively. With the useful properties of feedback analyzed, the reader is motivated to design high-performance, stable op amps and understand the trade-offs between speed, precision, and power dissipation.

Chapters 11 through 13 deal with more advanced topics: bandgap references, elementary switched-capacitor circuits, and the effect of nonlinearity and mismatch. These three subjects are included here because they prove essential in most analog and mixed-signal systems today.

Chapter 14 is concerned with high-order MOS device effects and models, emphasizing the circuit design implications. If preferred, the chapter can directly follow Chapter 2 as well. Chapter 15 describes CMOS fabrication technology with a brief overview of layout design rules.

Chapter 16 presents the layout and packaging of analog and mixed-signal circuits. Many practical issues that directly impact the performance of the circuit are described and various techniques are introduced.

The reader is assumed to have a basic knowledge of electronic circuits and devices, e.g., *pn* junctions, the concept of small-signal operation, equivalent circuits, and simple biasing. For a senior-level elective course, Chapters 1 through 8 can be covered in a quarter and Chapters 1 through 10 in a semester. For a first-year graduate course, Chapters 1 through 11 plus one of Chapters 12, 13, or 14 can be taught in one quarter, and almost the entire book in one semester.

The problem sets at the end of each chapter are designed to extend the reader's understanding of the material and complement it with additional practical considerations. A solutions manual will be available for instructors.

Behzad Razavi
July 2000

Acknowledgments for the Second Edition

The second edition was enthusiastically and meticulously reviewed by a large number of individuals in academia and industry. It is my pleasure to acknowldege their contributions:

Saheed Adeolu Tijani (University of Pavia)
Firooz Aflatouni (University of Pennsylvania)
Pietro Andreani (Lund University)
Emily Allstot (University of Washington)
Tejasvi Anand (University of Illinois, Urbana-Champaign)
Afshin Babveyh (Stanford)
Nima Baniasadi (UC Berkeley)
Sun Yong Cho (Seoul National University)
Min Sung Chu (Seoul National University)
Yi-Ying Cheng (UCLA)
Jeny Chu (UCLA)
Milad Darvishi (Qualcomm)
Luis Fei (Intel)
Andrea Ghilioni (University of Pavia)
Chengkai Gu (UCLA)
Payam Heydari (UC Irvine)
Cheng-En Hsieh (National Taiwan University)
Po-Chiun Huang (National Tsing-Hua University)
Deog-Kyoon Jeong (Seoul National University)
Nader Kalantari (Broadcom)

Alireza Karimi (UC Irvine)
Ehsan Kargaran (University of Pavia)
Sotirios Limotyrakis (Qualcomm Atheros)
Xiaodong Liu (Lund University)
Nima Maghari (University of Florida)
Shahriar Mirabbasi (University of British Columbia)
Hossein Mohammadnezhad (UC Irvine)
Amir Nikpaik (University of British Columbia)
Aria Samiei (University of Southern California)
Kia Salimi (IMEC)
Alireza Sharif-Bakhtiar (University of Toronto)
Guanghua Shu (University of Illinois, Urbana-Champaign)
David Su (Qualcomm Atheros)
Siyu Tan (Lund University)
Jeffrey Wang (University of Toronto)
Tzu-Chao Yan (National Chiao-Tung University)
Ehzan Zhian Tabasy (University of Texas A&M)

In addition, my colleague Jason Woo explained to me many subtleties of nanometer devices and their physics. I wish to thank all.

The production of the book has been in the hands of Heather Ervolino and Vincent Bradshaw of McGraw-Hill, who tirelessly attended to every detail over a six-month period. I would like to thank both.

Finally, I wish to thank my wife, Angelina, for her continual help with typing and organizing the chapters.

Acknowledgments for the First Edition

Writing a book begins with a great deal of excitement. However, after two years of relentless writing, drawing, and revising, when the book exceeds 700 pages and it is almost impossible to make the equations and subscripts and superscripts in the last chapter consisent with those in the first, the author begins to feel streaks of insanity, realizing that the book will never finish without the support of many other people.

This book has benefited from the contributions of many individuals. A number of UCLA students read the first draft and the preview edition sentence by sentence. In particular, Alireza Zolfaghari, Ellie Cijvat, and Hamid Rafati meticulously read the book and found several hundred errors (some quite subtle). Also, Emad Hegazi, Dawei Guo, Alireza Razzaghi, Jafar Savoj, and Jing Tian made helpful suggestions regarding many chapters. I thank all.

Many experts in academia and industry read various parts of the book and provided useful feedback. Among them are Brian Brandt (National Semiconductor), Matt Corey (National Semiconductor), Terri Fiez (Oregon State University), Ian Galton (UC San Diego), Ali Hajimiri (Caltech), Stacy Ho (Analog Devices), Yin Hu (Texas Instruments), Shen-Iuan Liu (National Taiwan University), Joe Lutsky (National Semiconductor), Amit Mehrotra (University of Illinois, Urbana-Champaign), David Robertson (Analog Devices), David Su (T-Span), Tao Sun (National Semiconductor), Robert Taft (National Semiconductor), and Masoud Zargari (T-Span). Jason Woo (UCLA) patiently endured and answered my questions about device physics. I thank all.

Ramesh Harjani (University of Minnesota), John Nyenhius (Purdue University), Norman Tien (Cornell University), and Mahmoud Wagdy (California State University, Long Beach) reviewed the book proposal and made valuable sugegstions. I thank all.

My wife, Angelina, has made many contributions to this book, from typing chapters to finding numerous errors and raising questions that made me reexamine my own understanding. I am very grateful to her.

The timely production of the book was made possible by the hard work of the staff at McGraw-Hill, particularly Catherine Fields, Michelle Flomenhoft, Heather Burbridge, Denise Santor-Mitzit, and Jim Labeots. I thank all.

I learned analog design from two masters: Mehrdad Sharif-Bakhtiar (Sharif University of Technology) and Bruce Wooley (Stanford University), and it is only appropriate that I express my gratitude to them here. What I inherited from them will be inherited by many generations of students.

About the Author

Behzad Razavi received the BSEE degree from Sharif University of Technology in 1985 and the MSEE and PhDEE degrees from Stanford University in 1988 and 1992, respectively. He was with AT&T Bell Laboratories and Hewlett-Packard Laboratories until 1996. Since 1996, he has been Associate Professor and subsequently Professor of Electrical Engineering at University of California, Los Angeles. His current research includes wireless transceivers, frequency synthesizers, phase-locking and clock recovery for high-speed data communications, and data converters.

Professor Razavi was an Adjunct Professor at Princeton University from 1992 to 1994, and at Stanford University in 1995. He served on the Technical Program Committees of the International Solid-State Circuits Conference (ISSCC) from 1993 to 2002 and VLSI Circuits Symposium from 1998 to 2002. He has also served as Guest Editor and Associate Editor of the *IEEE Journal of Solid-State Circuits*, *IEEE Transactions on Circuits and Systems*, and *International Journal of High Speed Electronics*.

Professor Razavi received the Beatrice Winner Award for Editorial Excellence at the 1994 ISSCC, the best paper award at the 1994 European Solid-State Circuits Conference, the best panel award at the 1995 and 1997 ISSCC, the TRW Innovative Teaching Award in 1997, the best paper award at the IEEE Custom Integrated Circuits Conference in 1998, and the McGraw-Hill First Edition of the Year Award in 2001. He was the corecipient of both the Jack Kilby Outstanding Student Paper Award and the Beatrice Winner Award for Editorial Excellence at the 2001 ISSCC. He received the Lockheed Martin Excellence in Teaching Award in 2006, the UCLA Faculty Senate Teaching Award in 2007, and the CICC Best Invited Paper Award in 2009 and in 2012. He was the corecipient of the 2012 VLSI Circuits Symposium Best Student Paper Award and the 2013 CICC Best Paper Award. He was also recognized as one of the top 10 authors in the 50-year history of ISSCC. He received the 2012 Donald Pederson Award in Solid-State Circuits and the American Society for Engineering Education PSW Teaching Award in 2014.

Professor Razavi has served as an IEEE Distinguished Lecturer and is a Fellow of IEEE. He is the author of *Principles of Data Conversion System Design*, *RF Microelectronics*, *Design of Analog CMOS Integrated Circuits*, *Design of Integrated Circuits for Optical Communications*, and *Fundamentals of Microelectronics,* and the editor of *Monolithic Phase-Locked Loops and Clock Recovery Circuits* and *Phase-Locking in High-Performance Systems.*

Brief Contents

Contents

CHAPTER

1

Introduction to Analog Design

1.1 ■ Why Analog?

We are surrounded by "digital" devices: digital cameras, digital TVs, digital communications (cell phones and WiFi), the Internet, etc. Why, then, are we still interested in analog circuits? Isn't analog design old and out of fashion? Will there even be jobs for analog designers ten years from now?

Interestingly, these questions have been raised about every five years over the past 50 years, but mostly by those who either did not understand analog design or did not want to deal with its challenges. In this section, we learn that analog design is still essential, relevant, and challenging and will remain so for decades to come.

1.1.1 Sensing and Processing Signals

Many electronic systems perform two principal functions: they sense (receive) a signal and subsequently process and extract information from it. Your cell phone receives a radio-frequency (RF) signal and, after processing it, provides voice or data information. Similarly, your digital camera senses the light intensity emitted from various parts of an object and processes the result to extract an image.

We know intuitively that the complex task of *processing* is preferably carried out in the digital domain. In fact, we may wonder whether we can directly digitize the signal and avoid *any* operations in the analog domain. Figure 1.1 shows an example where the RF signal received by the antenna is digitized by an analog-to-digital converter (ADC) and processed entirely in the digital domain. Would this scenario send analog and RF designers to the unemployment office?

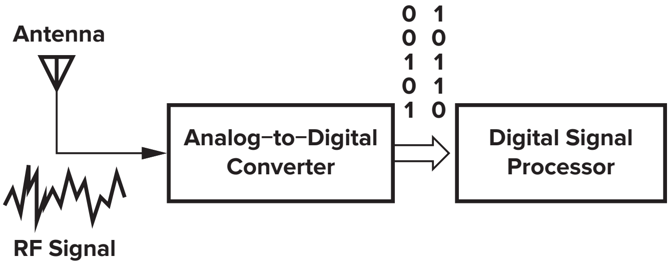


Figure 1.1 Hypothetical RF receiver with direct signal digitization.

The answer is an emphatic no. An ADC that could digitize the minuscule RF signal[1] would consume much more power than today's cell phone receivers. Furthermore, even if this approach were seriously considered, only *analog* designers would be able to develop the ADC. The key point offered by this example is that the sensing *interface* still demands high-performance analog design.

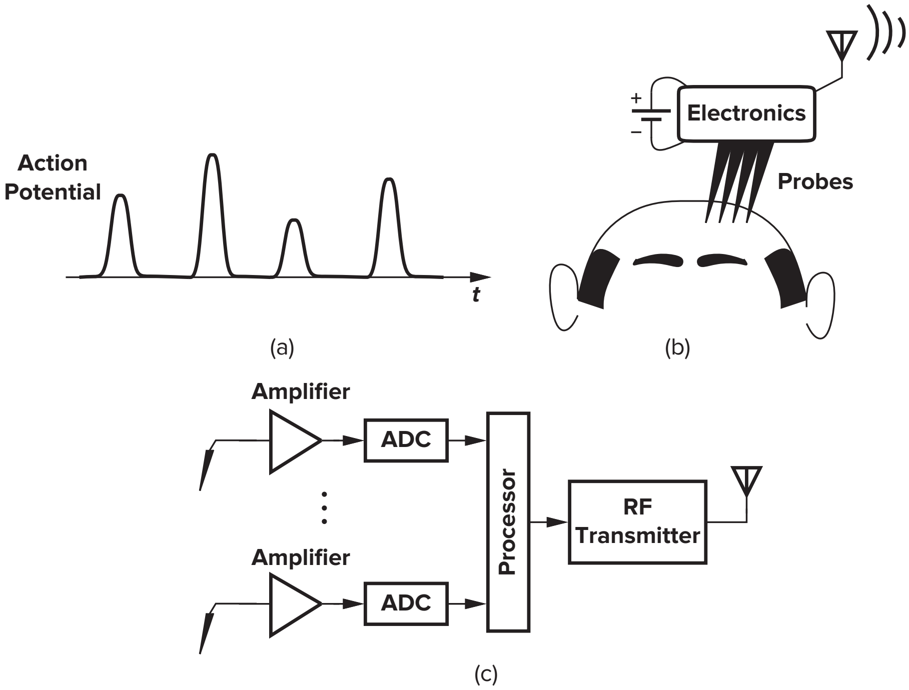


Figure 1.2 (a) Voltage waveform generated as a result of neural activity, (b) use of probes to measure action potentials, and (c) processing and transmission of signals.

Another interesting example of sensing challenges arises in the study of the brain signals. Each time a neuron in your brain "fires," it generates an electric pulse with a height of a few millivolts and a duration of a few hundred microseconds [Fig. 1.2(a)]. To monitor brain activities, a neural recording system may employ tens of "probes" (electrodes) [Fig. 1.2(b)], each sensing a series of pulses. The signal produced by each probe must now be amplified, digitized, and transmitted *wirelessly* so that the patient is free to move around [Fig. 1.2(c)]. The sensing, processing, and transmission electronics in this environment must consume a low amount of power for two reasons: (1) to permit the use of a small battery for days or weeks, and (2) to minimize the rise in the chip's temperature, which could otherwise damage the patient's tissue. Among the functions shown in Fig. 1.2(c), the amplifiers, the ADCs, and the RF transmitter—all analog circuits—consume most of the power.

1.1.2 When Digital Signals Become Analog

The use of analog circuits is not limited to analog signals. If a digital signal is so small and/or so distorted that a digital gate cannot interpret it correctly, then the analog designer must step in. For example, consider a long USB cable carrying data rate of hundreds of megabits per second between two laptops. As shown in Fig. 1.3, Laptop 1 delivers the data to the cable in the form of a sequence of ONEs and ZERO.

[1] And withstand large unwanted signals.

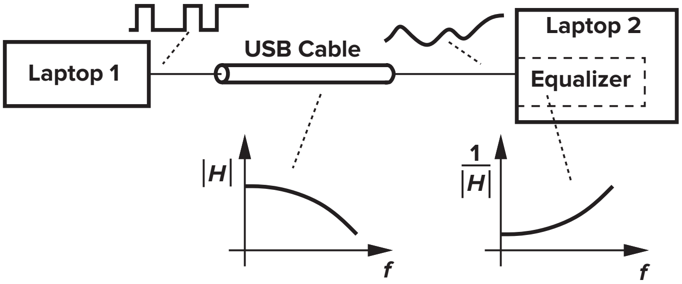


Figure 1.3 Equalization to compensate for high-frequency attenuation in a USB cable.

Unfortunately, the cable exhibits a finite bandwidth, attenuating high frequencies and distorting the data as it reaches Laptop 2. This device must now perform sensing and processing, the former requiring an analog circuit (called an “equalizer”) that corrects the distortion. For example, since the cable attenuates high frequencies, we may design the equalizer to *amplify* such frequencies, as shown conceptually by the $1/|H|$ plot in Fig. 1.3.

The reader may wonder whether the task of equalization in Fig. 1.3 could be performed in the digital domain. That is, could we directly digitize the received distorted signal, digitally correct for the cable’s limited bandwidth, and then carry out the standard USB signal processing? Indeed, this is possible if the ADC required here demands less power and less complexity than the analog equalizer. Following a detailed analysis, the analog designer decides which approach to adopt, but we intuitively know that at very high data rates, e.g., tens of gigabits per second, an analog equalizer proves more efficient than an ADC.

The above equalization task exemplifies a general trend in electronics: at lower speeds, it is more efficient to digitize the signal and perform the required function(s) in the digital domain, whereas at higher speeds, we implement the function(s) in the analog domain. The speed boundary between these two paradigms depends on the nature of the problem, but it has risen over time.

1.1.3 Analog Design Is in Great Demand

Despite tremendous advances in semiconductor technology, analog design continues to face new challenges, thus calling for innovations. As a gauge of the demand for analog circuits, we can consider the papers published by industry and academia at circuits conferences and see what percentage fall in our domain. Figure 1.4 plots the number of analog papers published at the International Solid-State Circuits

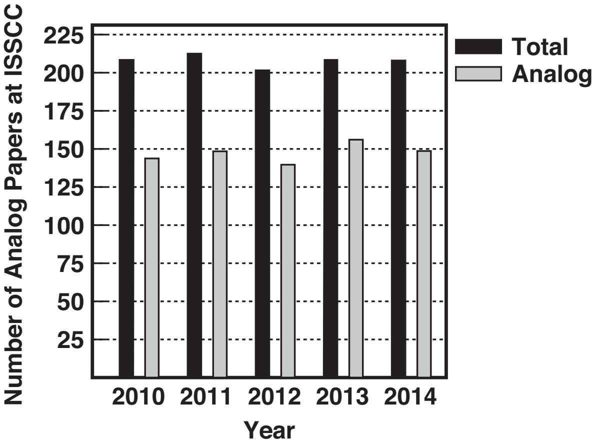


Figure 1.4 Number of analog papers published at the ISSCC in recent years.

Conference (ISSCC) in recent years, where "analog" is defined as a paper requiring the knowledge in this book. We observe that the *majority* of the papers involve analog design. This is true even though analog circuits are typically quite a lot less complex than digital circuits; an ADC contains several thousand transistors whereas a microprocessor employs billions.

1.1.4 Analog Design Challenges

Today's analog designers must deal with interesting and difficult problems. Our study of devices and circuits in this book will systematically illustrate various issues, but it is helpful to take a brief look at what lies ahead.

Transistor Imperfections As a result of scaling, MOS transistors continue to become *faster,* but at the cost of their "analog" properties. For example, the maximum voltage gain that a transistor can provide declines with each new generation of CMOS technology. Moreover, a transistor's characteristics may depend on its *surroundings*, i.e., the size, shape, and distance of other components around it on the chip.

Declining Supply Voltages As a result of device scaling, the supply voltage of CMOS circuits has inevitably fallen from about 12 V in the 1970s to about 0.9 V today. Many circuit configurations have not survived this supply reduction and have been discarded. We continue to seek new topologies that operate well at low voltages.

Power Consumption The semiconductor industry, more than ever, is striving for low-power design. This effort applies both to portable devices—so as to increase their battery lifetime—and to larger systems—so as to reduce the cost of heat removal and ease their drag on the earth's resources. MOS device scaling directly lowers the power consumption of digital circuits, but its effect on analog circuits is much more complicated.

Circuit Complexity Today's analog circuits may contain tens of thousands of transistors, demanding long and tedious simulations. Indeed, modern analog designers must be as adept at SPICE as at higher-level simulators such as MATLAB.

PVT Variations Many device and circuit parameters vary with the fabrication process, supply voltage, and ambient temperature. We denote these effects by PVT and design circuits such that their performance is acceptable for a specified range of PVT variations. For example, the supply voltage may vary from 1 V to 0.95 V and the temperature from 0° to 80°. Robust analog design in CMOS technology is a challenging task because device parameters vary significantly across PVT.

1.2 ■ Why Integrated?

The idea of placing multiple electronic devices on the same substrate was conceived in the late 1950s. In 60 years, the technology has evolved from producing simple chips containing a handful of components to fabricating flash drives with one trillion transistors as well as microprocessors comprising several billion devices. As Gordon Moore (one of the founders of Intel) predicted in the early 1970s, the number of transistors per chip has continued to double approximately every one and a half years. At the same time, the minimum dimension of transistors has dropped from about 25 μm in 1960 to about 12 nm in the year 2015, resulting in a tremendous improvement in the speed of integrated circuits.

Driven primarily by the memory and microprocessor market, integrated-circuit technologies have also embraced analog design, affording a complexity, speed, and precision that would be impossible to achieve using discrete implementations. We can no longer build a discrete prototype to predict the behavior and performance of modern analog circuits.

1.3 ■ Why CMOS?

The idea of metal-oxide-silicon field-effect transistors (MOSFETs) was patented by J. E. Lilienfeld in the early 1930s—well before the invention of the bipolar transistor. Owing to fabrication limitations, however, MOS technologies became practical only much later, in the early 1960s, with the first several generations producing only n-type transistors. It was in the mid-1960s that complementary MOS (CMOS) devices (i.e., with both n-type and p-type transistors) were introduced, initiating a revolution in the semiconductor industry.

CMOS technologies rapidly captured the digital market: CMOS gates dissipated power only during switching and required very few devices, two attributes in sharp contrast to their bipolar or GaAs counterparts. It was also soon discovered that the dimensions of MOS devices could be scaled down more easily than those of other types of transistors.

The next obvious step was to apply CMOS technology to analog design. The low cost of fabrication and the possibility of placing both analog and digital circuits on the same chip so as to improve the overall performance and/or reduce the cost of packaging made CMOS technology attractive. However, MOSFETs were slower and noisier than bipolar transistors, finding limited application.

How did CMOS technology come to dominate the analog market as well? The principal force was device scaling because it continued to improve the speed of MOSFETs. The intrinsic speed of MOS transistors has increased by orders of magnitude in the past 60 years, exceeding that of bipolar devices even though the latter have also been scaled (but not as fast).

Another critical advantage of MOS devices over bipolar transistors is that the former can operate with lower supply voltages. In today's technology, CMOS circuits run from supplies around 1 V and bipolar circuits around 2 V. The lower supplies have permitted a smaller power consumption for complex integrated circuits.

1.4 ■ Why This Book?

The design of analog circuits itself has evolved together with the technology and the performance requirements. As the device dimensions shrink, the supply voltage of intergrated circuits drops, and analog and digital circuits are fabricated on one chip, many design issues arise that were previously unimportant. Such trends demand that the analysis and design of circuits be accompanied by an in-depth understanding of new technology-imposed limitations.

Good analog design requires intuition, rigor, and creativity. As analog designers, we must wear our engineer's hat for a quick and intuitive understanding of a large circuit, our mathematician's hat for quantifying subtle, yet important effects in a circuit, and our artist's hat for inventing new circuit topologies.

This book describes modern analog design from both intuitive and rigorous angles. It also fosters the reader's creativity by carefully guiding him or her through the evolution of each circuit and presenting the thought process that occurs during the development of new circuit techniques.

1.5 ■ Levels of Abstraction

Analysis and design of integrated circuits often require thinking at various levels of abstraction. Depending on the effect or quantity of interest, we may study a complex circuit at device physics level, transistor level, architecture level, or system level. In other words, we may consider the behavior of individual devices in terms of their internal electric fields and charge transport [Fig. 1.5(a)], the interaction of a group of devices according to their electrical characteristics [Fig. 1.5(b)], the function of several building blocks operating as a unit [Fig. 1.5(c)], or the performance of the system in terms of that of its constituent subsystems

[Fig. 1.5(d)]. Switching between levels of abstraction becomes necessary in both understanding the details of the operation and optimizing the overall performance. In fact, in today's IC industry, the interaction among all groups, from device physicists to system designers, is essential to achieving high performance and low cost. In this book, we begin with device physics and develop increasingly more complex circuit topologies.

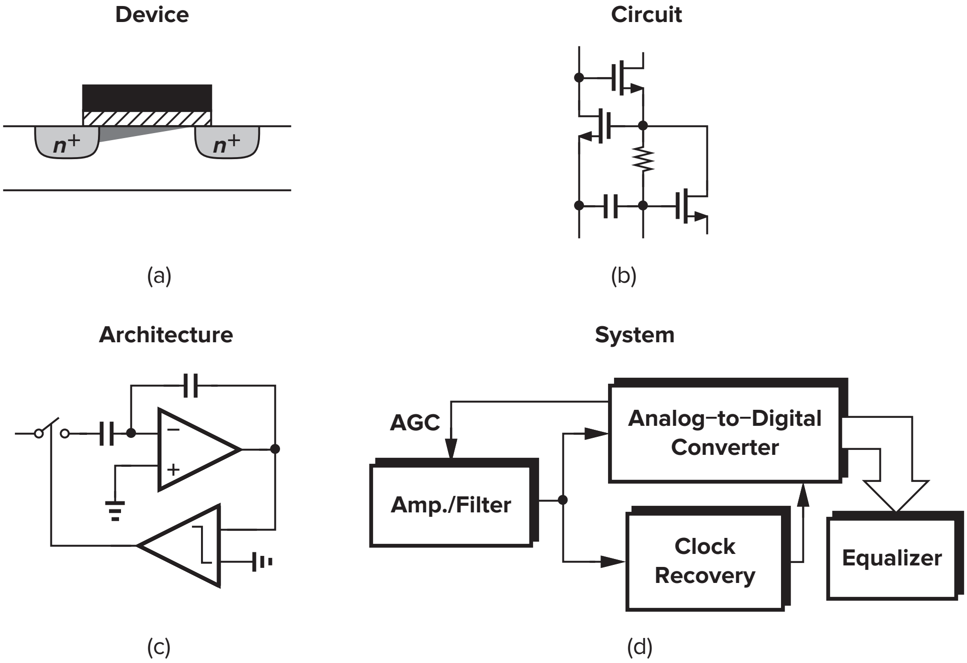


Figure 1.5 Abstraction levels in circuit design: (a) device level, (b) circuit level, (c) architecture level, (d) system level.

CHAPTER

2

Basic MOS Device Physics

In studying the design of integrated circuits (ICs), one of two extreme approaches can be taken, (1) begin with quantum mechanics and understand solid-state physics, semiconductor device physics, device modeling, and finally the design of circuits; or (2) treat each semiconductor device as a black box whose behavior is described in terms of its terminal voltages and currents and design circuits with little attention to the internal operation of the device. Experience shows that neither approach is optimum. In the first case, the reader cannot see the relevance of all the physics to designing circuits, and in the second, he or she is constantly mystified by the contents of the black box.

In today's IC industry, a solid understanding of semiconductor devices is essential—more so in analog design than in digital design, because in the former, transistors are not considered to be simple switches, and many of their second-order effects directly impact the performance. Furthermore, as each new generation of IC technologies scales the devices, these effects become more significant. Since the designer must often decide which effects can be neglected in a given circuit, insight into device operation proves invaluable.

In this chapter, we study the physics of MOSFETs at an elementary level, covering the bare minimum that is necessary for basic analog design. The ultimate goal is still to develop a circuit model for each device by formulating its operation, but this is accomplished through a good understanding of the underlying principles. After studying many analog circuits in Chapters 3 through 14 and gaining motivation for a deeper understanding of devices, we return to the subject in Chapter 17 and deal with other aspects of MOS operation.

We begin our study with the structure of MOS transistors and derive their I/V characteristics. Next, we describe second-order effects such as body effect, channel-length modulation, and subthreshold conduction. We then identify the parasitic capacitances of MOSFETs, derive a small-signal model, and present a simple SPICE model. We assume that the reader is familiar with such basic concepts as doping, mobility, and *pn* junctions.

2.1 ■ General Considerations

2.1.1 MOSFET as a Switch

Before delving into the actual operation of the MOSFET, we consider a simplistic model of the device so as to gain a feel for what the transistor is expected to be and which aspects of its behavior are important.

Shown in Fig. 2.1 is the symbol for an *n*-type MOSFET, revealing three terminals: gate (G), source (S), and drain (D). The latter two are interchangeable because the device is symmetric. When operating

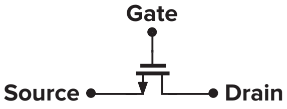


Figure 2.1 Simple view of a MOS device.

as a switch, the transistor "connects" the source and the drain together if the gate voltage, V_G, is "high" and isolates the source and the drain if V_G is "low."

Even with this simplified view, we must answer several questions. For what value of V_G does the device turn on? In other words, what is the "threshold" voltage? What is the resistance between S and D when the device is on (or off)? How does this resistance depend on the terminal voltages? Can we always model the path between S and D by a simple linear resistor? What limits the speed of the device?

While all of these questions arise at the circuit level, they can be answered only by analyzing the structure and physics of the transistor.

2.1.2 MOSFET Structure

Figure 2.2 shows a simplified structure of an n-type MOS (NMOS) device. Fabricated on a p-type substrate (also called the "bulk" or the "body"), the device consists of two heavily-doped n regions forming the source and drain terminals, a heavily-doped (conductive) piece of polysilicon[1] (simply called "poly") operating as the gate, and a thin layer of silicon dioxide (SiO_2) (simply called "oxide") insulating the gate from the substrate. The useful action of the device occurs in the substrate region under the gate oxide. Note that the structure is symmetric with respect to S and D.

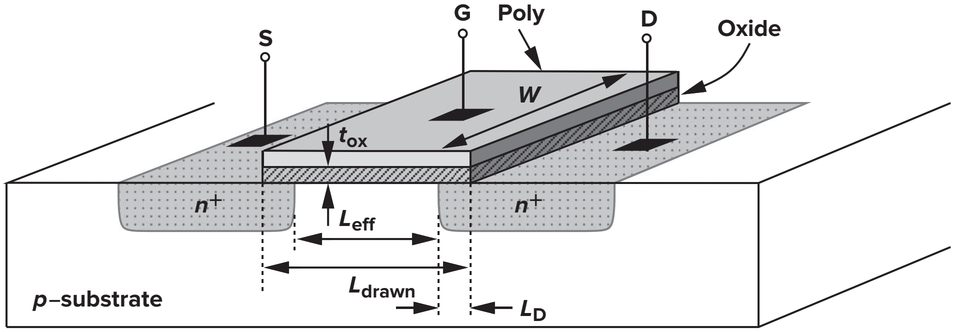


Figure 2.2 Structure of a MOS device.

The lateral dimension of the gate along the source-drain path is called the length, L, and that perpendicular to the length is called the width, W. Since the S/D junctions "side-diffuse" during fabrication, the actual distance between the source and the drain is slightly less than L. To avoid confusion, we write, $L_{eff} = L_{drawn} - 2L_D$, where L_{eff} is the "effective" length, L_{drawn} is the total length,[2] and L_D is the amount of side diffusion. As we will see later, L_{eff} and the gate oxide thickness, t_{ox}, play an important role in the performance of MOS circuits. Consequently, the principal thrust in MOS technology development is to reduce both of these dimensions from one generation to the next without degrading other parameters of the device. Typical values at the time of this writing are $L_{eff} \approx 10$ nm and $t_{ox} \approx 15$ Å. In the remainder of this book, we denote the effective length by L unless otherwise stated.

[1] Polysilicon is silicon in amorphous (non crystal) form. As explained in Chapter 18, when the gate silicon is grown on top of the oxide, it cannot form a crystal. The gate was originally made of metal [hence the term "metal-oxide-semiconductor" (MOS)] and is returning to metal in recent generations.

[2] The subscript "drawn" is used because this is the dimension that we draw in the layout of the transistor (Sec. 2.4.1).

If the MOS structure is symmetric, why do we call one n region the source and the other the drain? This becomes clear if the source is defined as the terminal that provides the charge carriers (electrons in the case of NMOS devices) and the drain as the terminal that collects them. Thus, as the voltages at the three terminals of the device vary, the source and the drain may exchange roles. These concepts are practiced in the problems at the end of the chapter.

We have thus far ignored the substrate on which the device is fabricated. In reality, the substrate potential greatly influences the device characteristics. That is, the MOSFET is a *four*-terminal device. Since in typical MOS operation, the S/D junction diodes must be reverse-biased, we assume that the substrate of NMOS transistors is connected to the most negative supply in the system. For example, if a circuit operates between zero and 1.2 volts, $V_{sub,NMOS} = 0$. The actual connection is usually provided through an ohmic p^+ region, as depicted in the side view of the device in Fig. 2.3.

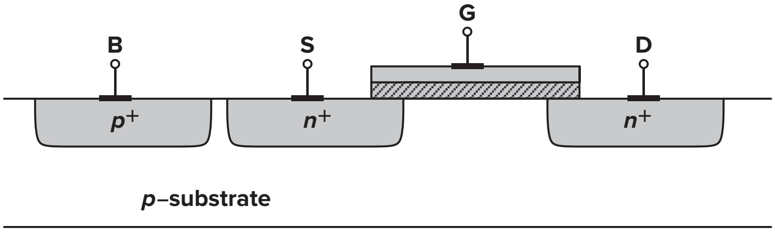


Figure 2.3 Substrate connection.

In complementary MOS (CMOS) technologies, both NMOS and PMOS transistors are available. From a simplistic viewpoint, the PMOS device is obtained by negating all of the doping types (including the substrate) [Fig. 2.4(a)], but in practice, NMOS and PMOS devices must be fabricated on the same wafer, i.e., the same substrate. For this reason, one device type can be placed in a "local substrate," usually called a "well." In today's CMOS processes, the PMOS device is fabricated in an n-well [Fig. 2.4(b)]. Note that the n-well must be connected to a potential such that the S/D junction diodes of the PMOS transistor remain reverse-biased under all conditions. In most circuits, the n-well is tied to the most positive supply voltage. For the sake of brevity, we sometimes call NMOS and PMOS devices "NFETs" and "PFETs," respectively.

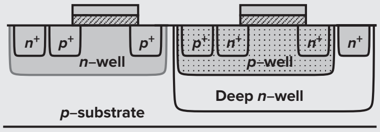
Figure 2.4(b) indicates an interesting difference between NMOS and PMOS transistors: while all NFETs share the same substrate, each PFET can have an independent n-well. This flexibility of PFETs is exploited in some analog circuits.

2.1.3 MOS Symbols

The circuit symbols used to represent NMOS and PMOS transistors are shown in Fig. 2.5. The symbols in Fig. 2.5(a) contain all four terminals, with the substrate denoted by "B" (bulk) rather than "S" to avoid confusion with the source. The source of the PMOS device is positioned on top as a visual aid because it has a higher potential than its gate. Since in most circuits the bulk terminals of NMOS and PMOS devices are tied to ground and V_{DD}, respectively, we usually omit these connections in drawing [Fig. 2.5(b)]. In digital circuits, it is customary to use the "switch" symbols depicted in Fig. 2.5(c) for the two types, but we prefer those in Fig. 2.5(b) because the visual distinction between S and D proves helpful in understanding the operation of circuits.

Nanometer Design Notes

Some modern CMOS processes offer a "deep n-well," an n-well that contains an NMOS device and its p-type bulk. As shown below, the NMOS transistor's bulk is now localized and need not be tied to that of other NMOS devices. But the design incurs substantial area overhead because the deep n-well must extend beyond the p-well by a certain amount and must maintain a certain distance to the regular n-well.



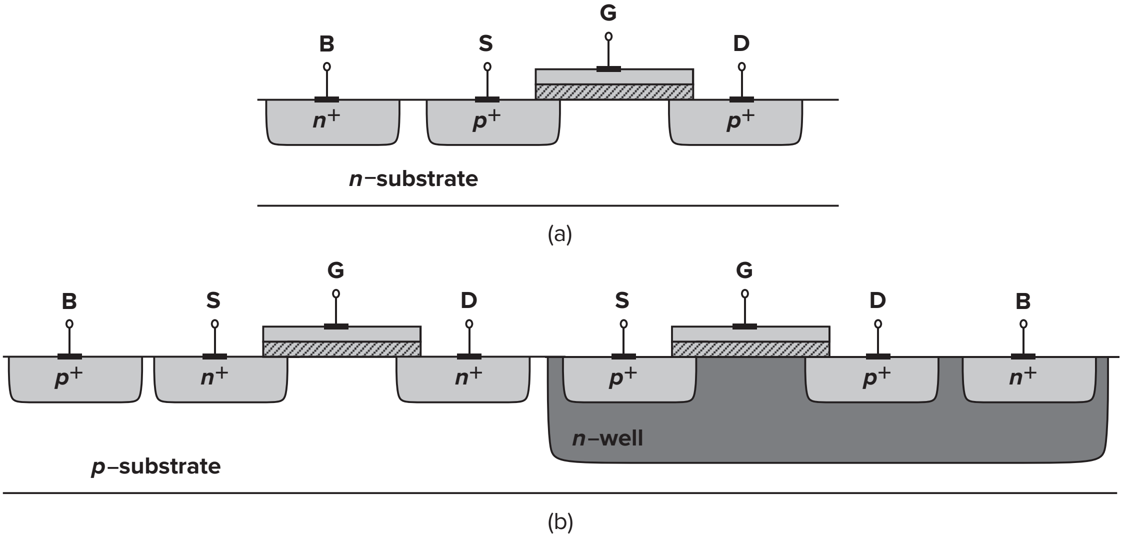


Figure 2.4 (a) Simple PMOS device; (b) PMOS inside an *n*-well.

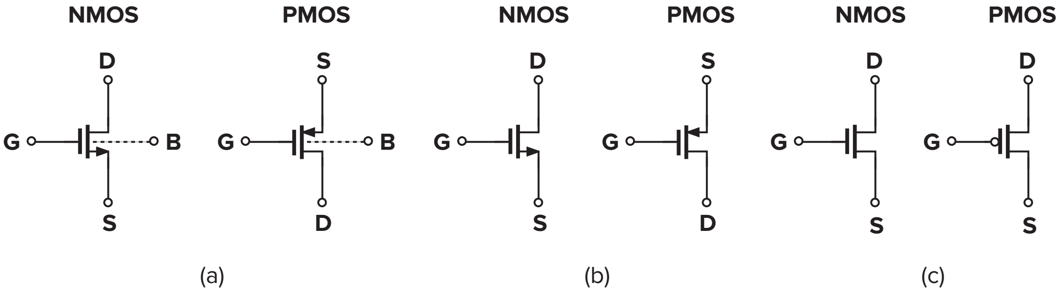


Figure 2.5 MOS symbols.

2.2 ■ MOS I/V Characteristics

In this section, we analyze the generation and transport of charge in MOSFETs as a function of the terminal voltages. Our objective is to derive equations for the I/V characteristics such that we can elevate our abstraction from device physics level to circuit level.

2.2.1 Threshold Voltage

Consider an NFET connected to external voltages as shown in Fig. 2.6(a). What happens as the gate voltage, V_G, increases from zero? Since the gate, the dielectric, and the substrate form a capacitor, as V_G becomes more positive, the holes in the *p*-substrate are repelled from the gate area, leaving negative ions behind so as to mirror the charge on the gate. In other words, a depletion region is created [Fig. 2.6(b)]. Under this condition, no current flows because no charge carriers are available.

As V_G increases, so do the width of the depletion region and the potential at the oxide-silicon interface. In a sense, the structure resembles a voltage divider consisting of two capacitors in series: the gate-oxide capacitor and the depletion-region capacitor [Fig. 2.6(c)]. When the interface potential reaches a sufficiently positive value, electrons flow from the source to the interface and eventually to the drain.

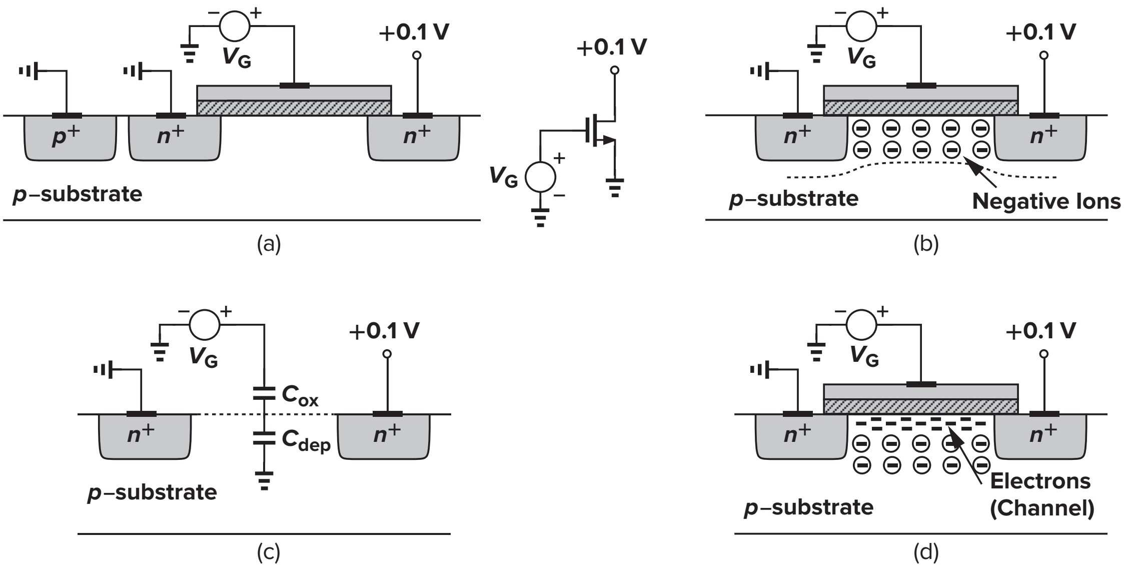


Figure 2.6 (a) A MOSFET driven by a gate voltage; (b) formation of depletion region; (c) onset of inversion; (d) formation of inversion layer.

Thus, a "channel" of charge carriers is formed under the gate oxide between S and D, and the transistor is "turned on." We say the interface is "inverted." For this reason, the channel is also called the "inversion layer." The value of V_G for which this occurs is called the "threshold voltage," V_{TH}. If V_G rises further, the charge in the depletion region remains relatively constant while the channel charge density continues to increase, providing a greater current from S to D.

In reality, the turn-on phenomenon is a gradual function of the gate voltage, making it difficult to define V_{TH} unambiguously. In semiconductor physics, the V_{TH} of an NFET is usually defined as the gate voltage for which the interface is "as much n-type as the substrate is p-type." It can be proved [1] that[3]

$$V_{TH} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}} \tag{2.1}$$

where Φ_{MS} is the difference between the work functions of the polysilicon gate and the silicon substrate, $\Phi_F = (kT/q)\ln(N_{sub}/n_i)$, k is Boltzmann's constant, q is the electron charge, N_{sub} is the doping density of the substrate, n_i is the density of electrons in undoped silicon, Q_{dep} is the charge in the depletion region, and C_{ox} is the gate-oxide capacitance per unit area. From pn junction theory, $Q_{dep} = \sqrt{4q\epsilon_{si}|\Phi_F|N_{sub}}$, where ϵ_{si} denotes the dielectric constant of silicon. Since C_{ox} appears very frequently in device and circuit calculations, it is helpful to remember that for $t_{ox} \approx 20$ Å, $C_{ox} \approx 17.25$ fF/μm^2. The value of C_{ox} can then be scaled proportionally for other oxide thicknesses.

In practice, the "native" threshold value obtained from the above equation may not be suited to circuit design, e.g., $V_{TH} = 0$ and the device does not turn off for $V_G \geq 0$.[4] For this reason, the threshold voltage is typically adjusted by implantation of dopants into the channel area during device fabrication, in essence altering the doping level of the substrate near the oxide interface. For example, as shown in Fig. 2.7, if a thin sheet of p^+ is created, the gate voltage required to deplete this region increases.

[3] Charge trapping in the oxide is neglected here.

[4] Called a "depletion-mode" FET, such a device was used in old technologies. NFETs with a positive threshold are called "enhancement-mode" devices.

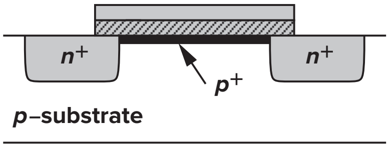


Figure 2.7 Implantation of $p+$ dopants to alter the threshold.

The above definition is not directly applicable to the *measurement* of V_{TH}. In Fig. 2.6(a), only the drain current can indicate whether the device is "on" or "off," failing to reveal at what V_{GS} the interface is as much n-type as the bulk is p-type. As a result, the calculation of V_{TH} from I/V measurements is somewhat ambiguous. We will return to this point later, but assume for now that the device turns on *abruptly* for $V_{GS} \geq V_{TH}$.

The turn-on phenomenon in a PMOS device is similar to that of NFETs, but with all the polarities reversed. As shown in Fig. 2.8, if the gate-source voltage becomes sufficiently *negative*, an inversion layer consisting of holes is formed at the oxide-silicon interface, providing a conduction path between the source and the drain. That is, the threshold voltage of a PMOS device is typically negative.

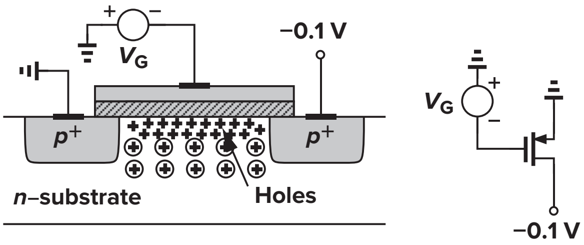


Figure 2.8 Formation of inversion layer in a PFET.

2.2.2 Derivation of I/V Characteristics

In order to obtain the relationship between the drain current of a MOSFET and its terminal voltages, we make two observations.

First, consider a semiconductor bar carrying a current I [Fig. 2.9(a)]. If the mobile charge density along the direction of current is Q_d coulombs per meter and the velocity of the charge is v meters per second, then

$$I = Q_d \cdot v \tag{2.2}$$

To understand why, we measure the total charge that passes through a cross section of the bar in unit time. With a velocity v, all of the charge enclosed in v meters of the bar must flow through the cross section in

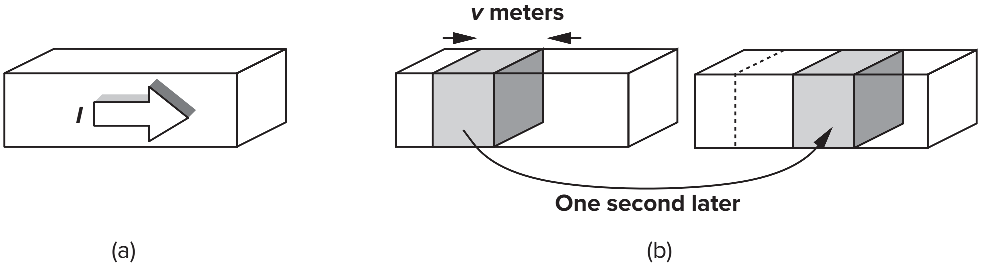


Figure 2.9 (a) A semiconductor bar carrying a current I; (b) snapshots of the carriers one second apart.

one second [Fig. 2.9(b)]. Since the charge density is Q_d, the total charge in v meters equals $Q_d \cdot v$. This lemma proves useful in analyzing semiconductor devices.

Second, to utilize the above lemma, we must determine the mobile charge density in a MOSFET. To this end, consider an NFET whose source and drain are connected to ground [Fig. 2.10(a)]. What is the charge density in the inversion layer? Since we assume that the onset of inversion occurs at $V_{GS} = V_{TH}$, the inversion charge density produced by the gate-oxide capacitance is proportional to $V_{GS} - V_{TH}$. For $V_{GS} \geq V_{TH}$, any charge placed on the gate must be mirrored by the charge in the channel, yielding a uniform channel charge density (charge per unit length along the source-drain path) equal to

$$Q_d = WC_{ox}(V_{GS} - V_{TH}) \tag{2.3}$$

where C_{ox} is multiplied by W to represent the total capacitance per unit length.

Now suppose, as depicted in Fig. 2.10(b), that the drain voltage is greater than zero. Since the channel potential varies from zero at the source to V_D at the drain, the local voltage *difference* between the gate and the channel varies from V_G (near the source) to $V_G - V_D$ (near the drain). Thus, the charge density at a point x along the channel can be written as

$$Q_d(x) = WC_{ox}[V_{GS} - V(x) - V_{TH}] \tag{2.4}$$

where $V(x)$ is the channel potential at x. From (2.2), the current is given by

$$I_D = -WC_{ox}[V_{GS} - V(x) - V_{TH}]v \tag{2.5}$$

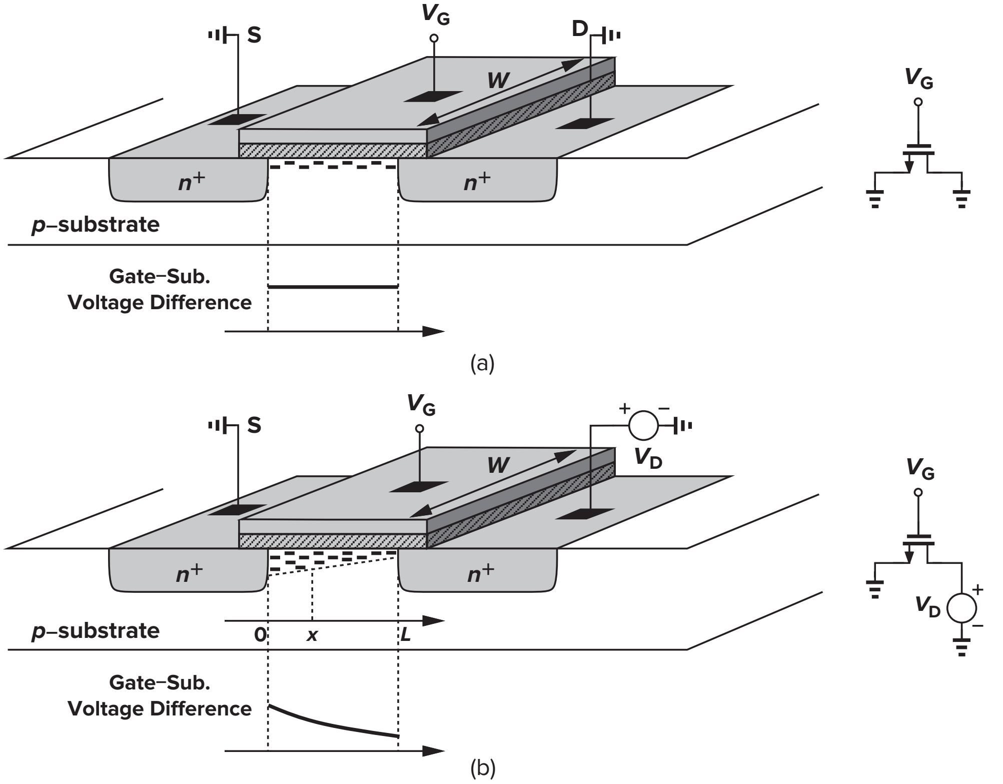


Figure 2.10 Channel charge with (a) equal source and drain voltages and (b) unequal source and drain voltages.

where the negative sign is inserted because the charge carriers are negative. Note that v denotes the velocity of the electrons in the channel. For semiconductors, $v = \mu E$, where μ is the mobility of charge carriers and E is the electric field. Noting that $E(x) = -dV/dx$ and representing the mobility of electrons by μ_n, we have

$$I_D = WC_{ox}[V_{GS} - V(x) - V_{TH}]\mu_n \frac{dV(x)}{dx} \tag{2.6}$$

subject to boundary conditions $V(0) = 0$ and $V(L) = V_{DS}$. While $V(x)$ can be easily found from this equation, the quantity of interest is in fact I_D. Multiplying both sides by dx and performing integration, we obtain

$$\int_{x=0}^{L} I_D dx = \int_{V=0}^{V_{DS}} WC_{ox}\mu_n[V_{GS} - V(x) - V_{TH}]dV \tag{2.7}$$

Since I_D is constant along the channel,

$$I_D = \mu_n C_{ox} \frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{2.8}$$

Note that L is the effective channel length.

Figure 2.11 plots the parabolas given by (2.8) for different values of V_{GS}, indicating that the "current capability" of the device increases with V_{GS}. Calculating $\partial I_D/\partial V_{DS}$, the reader can show that the peak of each parabola occurs at $V_{DS} = V_{GS} - V_{TH}$ and the peak current is

$$I_{D,max} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.9}$$

We call $V_{GS} - V_{TH}$ the "overdrive voltage" and W/L the "aspect ratio." If $V_{DS} \le V_{GS} - V_{TH}$, we say the device operates in the "triode region."[5]

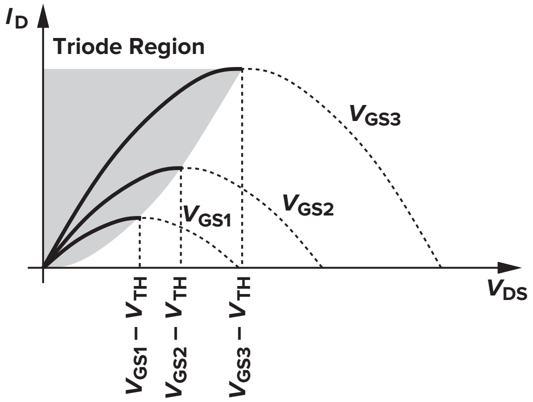


Figure 2.11 Drain current versus drain-source voltage in the triode region.

Equations (2.8) and (2.9) serve as our first step toward CMOS circuit design, describing the dependence of I_D upon the constant of the technology, $\mu_n C_{ox}$, the device dimensions, W and L, and the gate and drain potentials with respect to the source. Note that the integration in (2.7) assumes that μ_n and V_{TH} are independent of x and the gate and drain voltages, an approximation that we will revisit in Chapter 17.

[5] Also called the "linear region."

If in (2.8), $V_{DS} \ll 2(V_{GS} - V_{TH})$, we have

$$I_D \approx \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})V_{DS} \tag{2.10}$$

that is, the drain current is a *linear* function of V_{DS}. This is also evident from the characteristics of Fig. 2.11 for small V_{DS}: as shown in Fig. 2.12, each parabola can be approximated by a straight line. The linear relationship implies that the path from the source to the drain can be represented by a linear resistor equal to

$$R_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})} \tag{2.11}$$

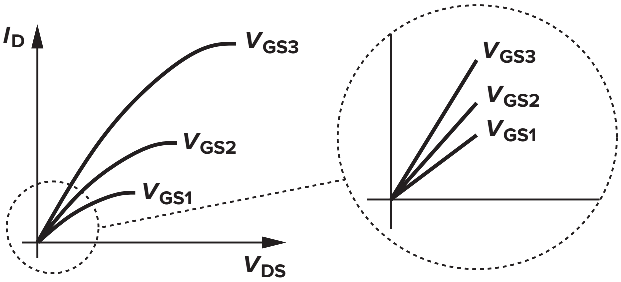


Figure 2.12 Linear operation in deep triode region.

A MOSFET can therefore operate as a resistor whose value is controlled by the overdrive voltage [so long as $V_{DS} \ll 2(V_{GS} - V_{TH})$]. This is conceptually illustrated in Fig. 2.13. Note that in contrast to bipolar transistors, a MOS device may be on even if it carries no current. With the condition $V_{DS} \ll 2(V_{GS} - V_{TH})$, we say the device operates in the deep triode region.

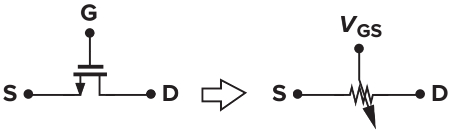


Figure 2.13 MOSFET as a controlled linear resistor.

▶ **Example 2.1**

For the arrangement in Fig. 2.14(a), plot the on-resistance of M_1 as a function of V_G. Assume that $\mu_n C_{ox} = 50\ \mu\text{A/V}^2$, $W/L = 10$, and $V_{TH} = 0.3$ V. Note that the drain terminal is open.

Solution

Since the drain terminal is open, $I_D = 0$ and $V_{DS} = 0$. Thus, if the device is on, it operates in the deep triode region. For $V_G < 1\text{ V} + V_{TH}$, M_1 is off and $R_{on} = \infty$. For $V_G > 1\text{ V} + V_{TH}$, we have

$$R_{on} = \frac{1}{50\ \mu\text{A/V}^2 \times 10(V_G - 1\text{ V} - 0.3\text{ V})} \tag{2.12}$$

The result is plotted in Fig. 2.14(b).

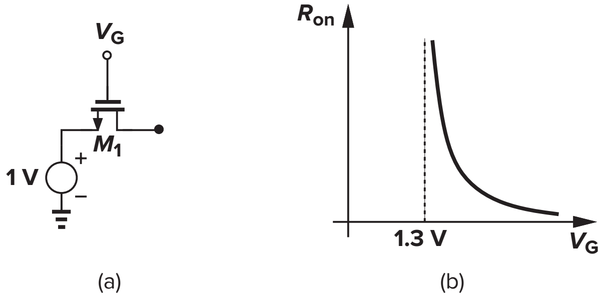


Figure 2.14

MOSFETs operating as controllable resistors play a crucial role in many analog circuits. For example, a voltage-controlled resistor can be used to adjust the frequency of the clock generator in a laptop computer if the system must go into a power saving mode. As studied in Chapter 13, MOSFETs also serve as switches.

What happens if the drain-source voltage in Fig. 2.11 exceeds $V_{GS} - V_{TH}$? In reality, the drain current does *not* follow the parabolic behavior for $V_{DS} > V_{GS} - V_{TH}$. In fact, as shown in Fig. 2.15, I_D becomes relatively constant, and we say the device operates in the "saturation" region.[6] To understand this phenomenon, recall from (2.4) that the local density of the inversion-layer charge is proportional to $V_{GS} - V(x) - V_{TH}$. Thus, if $V(x)$ approaches $V_{GS} - V_{TH}$, then $Q_d(x)$ drops to zero. In other words, as depicted in Fig. 2.16, if V_{DS} is slightly greater than $V_{GS} - V_{TH}$, then the inversion layer stops at $x \leq L$, and we say the channel is "pinched off." As V_{DS} increases further, the point at which Q_d equals zero gradually moves toward the source. Thus, at some point along the channel, the local potential difference between the gate and the oxide-silicon interface is not sufficient to support an inversion layer.

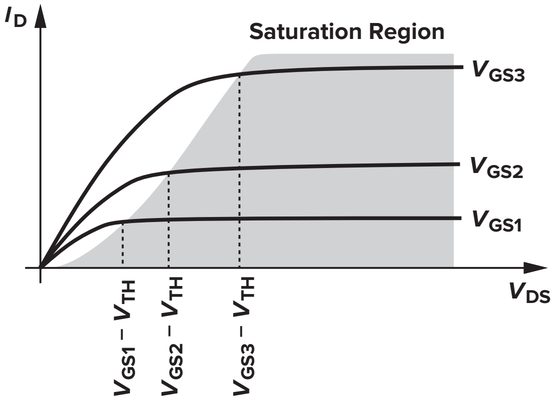


Figure 2.15 Saturation of drain current.

How does the device conduct current in the presence of pinch-off? As the electrons approach the pinch-off point (where $Q_d \to 0$), their velocity rises tremendously ($v = I/Q_d$). Upon passing the pinch-off point, the electrons simply shoot through the depletion region near the drain junction and arrive at the drain terminal.

[6] Note the difference between saturation in bipolar and MOS devices.

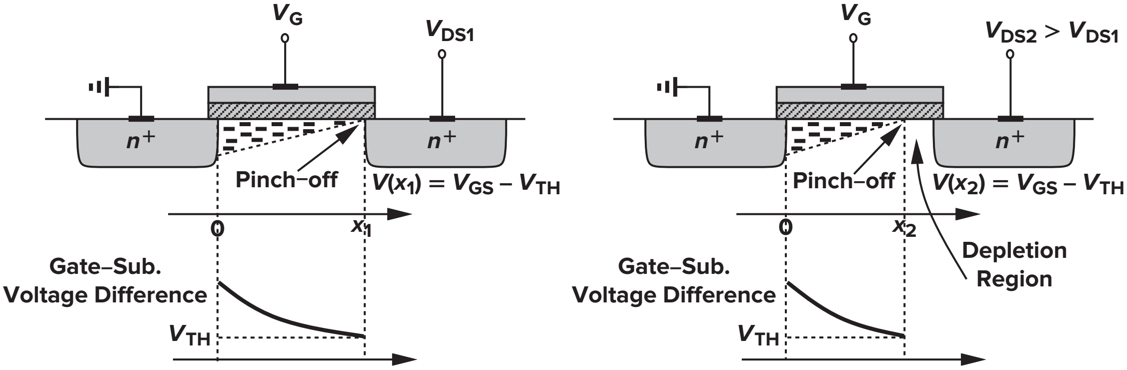


Figure 2.16 Pinch-off behavior.

With the above observations, we reexamine (2.7) for a saturated device. Since Q_d is the density of *mobile* charge, the integral on the left-hand side of (2.7) must be taken from $x = 0$ to $x = L'$, where L' is the point at which Q_d drops to zero (e.g., x_2 in Fig. 2.16), and that on the right from $V(x) = 0$ to $V(x) = V_{GS} - V_{TH}$. As a result,

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.13}$$

indicating that I_D is relatively independent of V_{DS} if L' remains close to L. We say the device exhibits a "square-law" behavior. If I_D is known, then V_{GS} is obtained as

$$V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox}\frac{W}{L'}}} + V_{TH} \tag{2.14}$$

We must emphasize that for the transistor to remain in saturation (as is the case in many analog circuits), the drain-source voltage must be equal to or greater than the overdrive voltage. For this reason, some books write $V_{D,sat} = V_{GS} - V_{TH}$, where $V_{D,sat}$ denotes the minimum V_{DS} necessary for operation in saturation. As seen later in this book, if the signal swings at the drain or the gate cause V_{DS} to fall below $V_{GS} - V_{TH}$, then a number of undesirable effects occur. For this reason, the choice of the overdrive and hence $V_{D,sat}$ translates to a certain voltage "headroom" for the signal swings in the circuit: the larger the $V_{D,sat}$, the less headroom is available for the signals.

Equations (2.8) and (2.13) represent the "large-signal" behavior of NMOS devices; i.e., they can predict the drain current for arbitrary voltages applied to the gate, source, and drain (but only if the device is on). Since the nonlinear nature of these equations makes the analysis difficult, we often resort to linear approximations ("small-signal" models) so as to develop some understanding of a given circuit. This point becomes clear in Sec. 2.4.3.

For PMOS devices, Eqs. (2.8) and (2.13) are respectively written as

$$I_D = -\mu_p C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{2.15}$$

and

$$I_D = -\frac{1}{2}\mu_p C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.16}$$

The negative sign appears here because we assume that I_D flows from the drain to the source, whereas holes flow in the reverse direction. Note that V_{GS}, V_{DS}, V_{TH}, and $V_{GS} - V_{TH}$ are negative for a PMOS transistor that is turned on. Since the mobility of holes is about one-half the mobility of electrons, PMOS devices suffer from lower "current drive" capability.

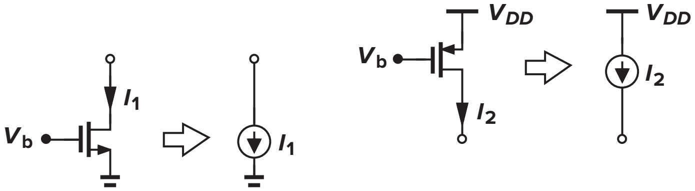


Figure 2.17 Saturated MOSFETs operating as current sources.

With L assumed constant, a saturated MOSFET can be used as a current source connected between the drain and the source (Fig. 2.17), an important component in analog design. Note that the NMOS current source injects current into ground and the PMOS current source draws current from V_{DD}. In other words, only one terminal of each current source is "floating." (It is difficult to design a current source that flows between two arbitrary nodes of a circuit.)

▶ **Example 2.2**

On a V_{DS}-V_{GS} plane, show the regions of operation of an NMOS transistor.

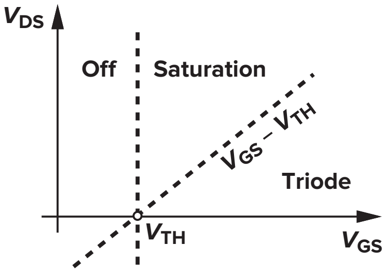


Figure 2.18 V_{DS}-V_{GS} plane showing regions of operation.

Solution

Since the value of V_{DS} with respect to $V_{GS} - V_{TH}$ determines the region of operation, we draw the line $V_{DS} = V_{GS} - V_{TH}$ in the plane, as shown in Fig. 2.18. If $V_{GS} > V_{TH}$, then the region above the line corresponds to saturation, and that below the line corresponds to the triode region. Note that for a given V_{DS}, the device eventually leaves saturation as V_{GS} increases. The minimum allowable V_{DS} for operation in saturation is also called $V_{D,sat}$. It is important to bear in mind that $V_{D,sat} = V_{GS} - V_{TH}$.

◀

The distinction between saturation and triode regions can be confusing, especially for PMOS devices. Intuitively, we note that the channel is pinched off if the difference between the gate and drain voltages is not sufficient to create an inversion layer. As depicted conceptually in Fig. 2.19, as $V_G - V_D$ of an NFET drops below V_{TH}, pinch-off occurs. Similarly, if $V_D - V_G$ of a PFET is not large enough ($< |V_{THP}|$), the device is saturated. Note that this view does not require knowledge of the source voltage. This means that we must know a priori which terminal operates as the drain. The drain is defined as the terminal with a higher (lower) voltage than the source for an NFET (PFET).

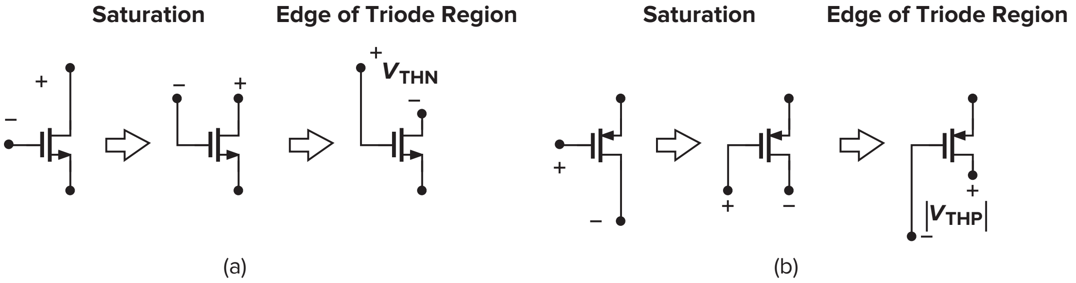


Figure 2.19 Conceptual visualization of saturation and triode regions.

2.2.3 MOS Transconductance

Since a MOSFET operating in saturation produces a current in response to its gate-source overdrive voltage, we may define a figure of merit that indicates how well a device converts a voltage to a current. More specifically, since in processing signals, we deal with the *changes* in voltages and currents, we define the figure of merit as the change in the drain current divided by the change in the gate-source voltage. Called the "transconductance" (and usually defined in the saturation region) and denoted by g_m, this quantity is expressed as

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{VDS\ \text{const.}} \tag{2.17}$$

$$= \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH}) \tag{2.18}$$

In a sense, g_m represents the sensitivity of the device: for a high g_m, a small change in V_{GS} results in a large change in I_D. We express g_m in $1/\Omega$ or in siemens (S); e.g., $g_m = 1/(100\,\Omega) = 0.01$ S. In analog design, we sometimes say a MOSFET operates as a "transconductor" or a "V/I converter" to indicate that it converts a voltage change to a current change. Interestingly, g_m in the saturation region is equal to the inverse of R_{on} in the deep triode region.

The reader can prove that g_m can also be expressed as

$$g_m = \sqrt{2\mu_n C_{ox}\frac{W}{L}I_D} \tag{2.19}$$

$$= \frac{2I_D}{V_{GS} - V_{TH}} \tag{2.20}$$

Plotted in Fig. 2.20, each of the above expressions proves useful in studying the behavior of g_m as a function of one parameter while other parameters remain constant. For example, (2.18) suggests that

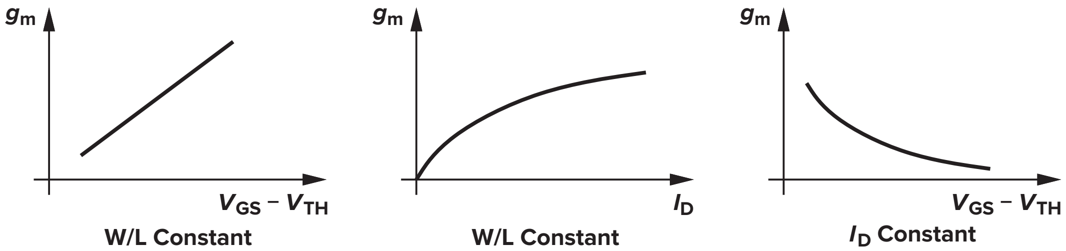


Figure 2.20 Approximate MOS transconductance as a function of overdrive and drain current.

g_m increases with the overdrive if W/L is constant, whereas (2.20) implies that g_m decreases with the overdrive if I_D is constant.

The I_D and $V_{GS} - V_{TH}$ terms in the above g_m equations are *bias* values. For example, a transistor with $W/L = 5\ \mu\text{m}/0.1\ \mu\text{m}$ and biased at $I_D = 0.5$ mA may exhibit a transconductance of $(1/200\ \Omega)$. If a signal is applied to the device, then I_D and $V_{GS} - V_{TH}$ and hence g_m *vary*, but in small-signal analysis, we assume that the signal amplitude is small enough that this variation is negligible.

Equation (2.19) implies that the transconductance can be raised arbitrarily if we increase W/L and keep I_D constant. This result is incorrect and will be revised in Sec. 2.3.

The concept of transconductance can also be applied to a device operating in the triode region, as illustrated in the following example.

▶ **Example 2.3**

For the arrangement shown in Fig. 2.21, plot the transconductance as a function of V_{DS}.

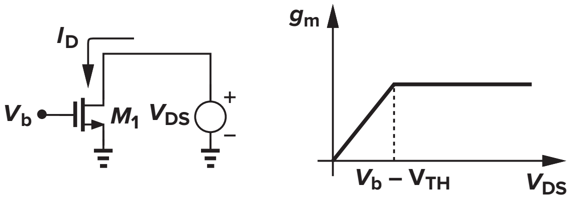


Figure 2.21

Solution

It is simpler to study g_m as V_{DS} decreases from infinity. So long as $V_{DS} \geq V_b - V_{TH}$, M_1 is in saturation, I_D is relatively constant, and, from (2.19), so is g_m. If the drain voltage falls below the gate voltage by more than one threshold, M_1 enters the triode region, and

$$g_m = \frac{\partial}{\partial V_{GS}} \left\{ \frac{1}{2} \mu_n C_{ox} \frac{W}{L} \left[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2 \right] \right\} \tag{2.21}$$

$$= \mu_n C_{ox} \frac{W}{L} V_{DS} \tag{2.22}$$

Thus, as plotted in Fig. 2.21, the transconductance drops in the triode region. For amplification, therefore, we usually employ MOSFETs in saturation.

◀

For a PFET, the transconductance in the saturation region is expressed as $g_m = -\mu_p C_{ox}(W/L)(V_{GS} - V_{TH}) = -2I_D/(V_{GS} - V_{TH}) = \sqrt{2\mu_p C_{ox}(W/L)I_D}$.

2.3 ■ Second-Order Effects

Our analysis of the MOS structure has thus far entailed various simplifying assumptions, some of which are not valid in many analog circuits. In this section, we describe three second-order effects that are essential in our subsequent circuit analyses. Other phenomena that appear in nanometer devices are studied in Chapter 17.

Body Effect In the analysis of Fig. 2.10, we tacitly assumed that the bulk and the source of the transistor were tied to ground. What happens if the bulk voltage of an NFET drops below the source voltage (Fig. 2.22)? Since the S and D junctions remain reverse-biased, we surmise that the device continues to operate properly, but some of its characteristics may change. To understand the effect, suppose

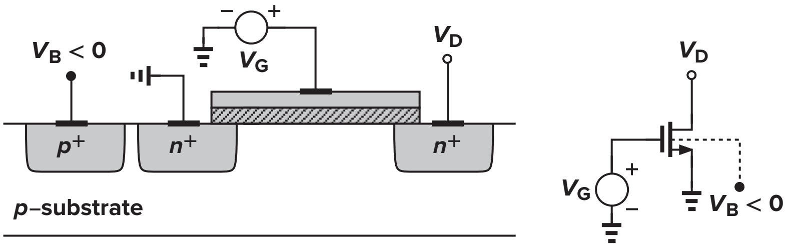


Figure 2.22 NMOS device with negative bulk voltage.

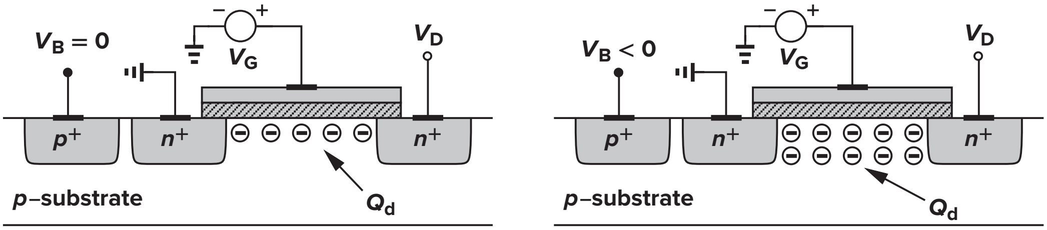


Figure 2.23 Variation of depletion region charge with bulk voltage.

$V_S = V_D = 0$, and V_G is somewhat less than V_{TH}, so that a depletion region is formed under the gate but no inversion layer exists. As V_B becomes more negative, more holes are attracted to the substrate connection, leaving a larger negative charge behind; i.e., as depicted in Fig. 2.23, the depletion region becomes wider. Now recall from Eq. (2.1) that the threshold voltage is a function of the total charge in the depletion region because the gate charge must mirror Q_d before an inversion layer is formed. Thus, as V_B drops and Q_d increases, V_{TH} also increases. This phenomenon is called the "body effect" or the "back-gate effect."

It can be proved that with body effect,

$$V_{TH} = V_{TH0} + \gamma \left(\sqrt{2\Phi_F + V_{SB}} - \sqrt{|2\Phi_F|} \right) \tag{2.23}$$

where V_{TH0} is given by (2.1), $\gamma = \sqrt{2q\epsilon_{si} N_{sub}}/C_{ox}$ denotes the body-effect coefficient, and V_{SB} is the source-bulk potential difference [1]. The value of γ typically lies in the range of 0.3 to 0.4 $\mathrm{V}^{1/2}$.

▶ **Example 2.4**

In Fig. 2.24(a), plot the drain current if V_X varies from $-\infty$ to 0. Assume $V_{TH0} = 0.3$ V, $\gamma = 0.4$ $\mathrm{V}^{1/2}$, and $2\Phi_F = 0.7$ V.

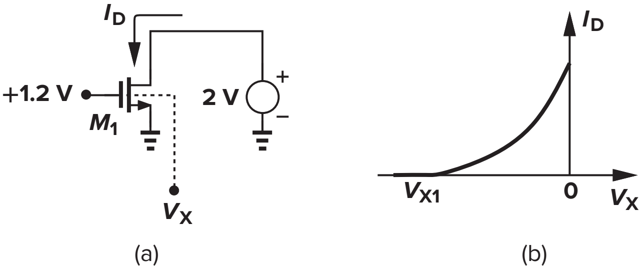


Figure 2.24

Solution

If V_X is sufficiently negative, the threshold voltage of M_1 exceeds 1.2 V and the device is off. That is,

$$1.2\text{ V} = 0.3 + 0.4\left(\sqrt{0.7 - V_{X1}} - \sqrt{0.7}\right) \tag{2.24}$$

and hence $V_{X1} = -8.83$ V. For $V_{X1} < V_X < 0$, I_D increases according to

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[V_{GS} - V_{TH0} - \gamma\left(\sqrt{2\Phi_F - V_X} - \sqrt{2\Phi_F}\right)\right]^2 \tag{2.25}$$

Fig. 2.24(b) shows the resulting behavior.

For body effect to manifest itself, the bulk potential, V_{sub}, need not change: if the source voltage varies with respect to V_{sub}, the same phenomenon occurs. For example, consider the circuit in Fig. 2.25(a), first ignoring body effect. We note that as V_{in} varies, V_{out} closely follows the input because the drain current remains equal to I_1. In fact, we can write

$$I_1 = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{out} - V_{TH})^2 \tag{2.26}$$

concluding that $V_{in} - V_{out}$ is constant if I_1 is constant [Fig. 2.25(b)].

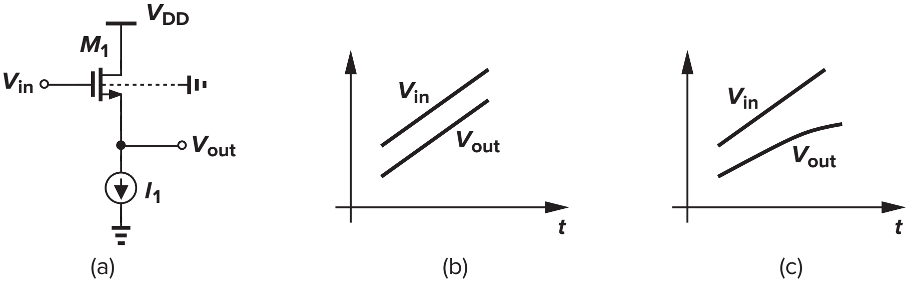


Figure 2.25 (a) A circuit in which the source-bulk voltage varies with input level; (b) input and output voltages with no body effect; (c) input and output voltages with body effect.

Now suppose that the substrate is tied to ground and body effect is significant. Then, as V_{in} and hence V_{out} become more positive, the potential difference between the source and the bulk increases, raising the value of V_{TH}. Equation (2.26) implies that $V_{in} - V_{out}$ must increase so as to maintain I_D constant [Fig. 2.25(c)].

Body effect is usually undesirable. The change in the threshold voltage, e.g., as in Fig. 2.25(a), often complicates the design of analog (and even digital) circuits. Device technologists balance N_{sub} and C_{ox} to obtain a reasonable value for γ.

▶ **Example 2.5**

Equation (2.23) suggests that if V_{SB} becomes *negative*, then V_{TH} *decreases*. Is this correct?

Solution

Yes, it is. If the bulk voltage of an NMOS device rises above its source voltage, V_{TH} falls below V_{TH0}. This observation proves useful in low-voltage design, where the performance of a circuit may suffer due to a high threshold voltage; one can bias the bulk to reduce V_{TH}. Unfortunately, this is not straightforward for NFETs because they typically share one substrate, but it can readily be applied to individual PFETs.

Channel-Length Modulation In the analysis of channel pinch-off in Sec. 2.2, we noted that the actual length of the channel gradually decreases as the potential difference between the gate and the drain decreases. In other words, in (2.13), L' is in fact a function of V_{DS}. This effect is called "channel-length modulation." Writing $L' = L - \Delta L$, i.e., $1/L' \approx (1 + \Delta L/L)/L$, and assuming a first-order relationship between $\Delta L/L$ and V_{DS}, such as $\Delta L/L = \lambda V_{DS}$, we have, in saturation,

$$I_D \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.27}$$

where λ is the "channel-length modulation coefficient." Illustrated in Fig. 2.26, this phenomenon results in a nonzero slope in the I_D/V_{DS} characteristic and hence a nonideal current source between D and S in saturation. The parameter λ represents the *relative* variation in length for a given increment in V_{DS}. Thus, for longer channels, λ is smaller.

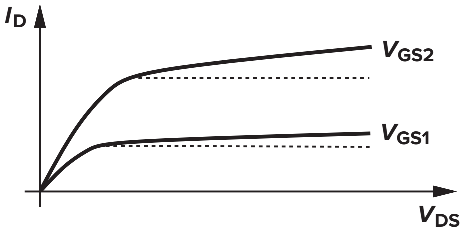


Figure 2.26 Finite saturation region slope resulting from channel-length modulation.

Example 2.6

Is there channel-length modulation in the triode region?

Solution

No, there is not. In the triode region, the channel continuously stretches from the source to the drain, experiencing no pinch-off. Thus, the drain voltage does not modulate the length of the channel.

The reader may then observe a discontinuity in the equations as the device goes from the triode region to saturation:

$$I_{D,tri} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2\right] \tag{2.28}$$

$$I_{D,sat} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.29}$$

The former yields $(1/2)\mu_n C_{ox} W/L(V_{GS} - V_{TH})^2$ at the edge of the triode region, whereas the latter exhibits an additional factor of $1+\lambda V_{DS}$. This discrepancy is removed in more complex models of MOSFETs (Chapter 17).

With channel-length modulation, some of the expressions derived for g_m must be modified. Equations (2.18) and (2.19) are respectively rewritten as

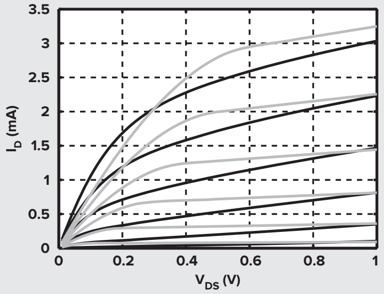
$$g_m = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})(1 + \lambda V_{DS}) \tag{2.30}$$

$$= \sqrt{2\mu_n C_{ox}(W/L)I_D(1 + \lambda V_{DS})} \tag{2.31}$$

while Eq. (2.20) remains unchanged.

Nanometer Design Notes

Nanometer transistors suffer from various imperfections and markedly depart from square-law behavior. Shown below are the actual I-V characteristics of an NFET with $W/L = 5$ μm/40 nm for $V_{GS} = 0.3$ V $\cdots 0.8$ V. Also plotted are the characteristics of a square-law device of the same dimensions. Despite our best efforts to match the latter device to the former, we still observe significant differences.



▶ **Example 2.7**

Keeping all other parameters constant, plot the I_D/V_{DS} characteristic of a MOSFET for $L = L_1$ and $L = 2L_1$.

Solution

Writing

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.32}$$

and $\lambda \propto 1/L$, we note that if the length is doubled, the slope of I_D vs. V_{DS} is divided by *four* because $\partial I_D/\partial V_{DS} \propto \lambda/L \propto 1/L^2$ (Fig. 2.27). (This is true only if $V_{GS} - V_{TH}$ is constant.) For a given gate-source overdrive, a larger L gives a more ideal current source while degrading the current capability of the device. Thus, W may need to be increased proportionally. In fact, if we double W to restore I_D to its original value, the slope also doubles. In other words, for a required drain current and a given overdrive, doubling the length reduces the slope by a factor of 2.

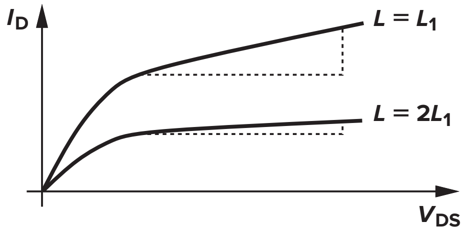


Figure 2.27 Effect of doubling channel length.

◀

The linear approximation $\Delta L/L \propto V_{DS}$ becomes less accurate in short-channel transistors, resulting in a *variable* slope in the saturated I_D/V_{DS} characteristics. We return to this issue in Chapter 17.

The dependence of I_D upon V_{DS} in saturation may suggest that the bias current of a MOSFET can be defined by the proper choice of the drain-source voltage, allowing freedom in the choice of $V_{GS} - V_{TH}$. However, since the dependence on V_{DS} is much weaker, the drain-source voltage is not used to set the current. That is, we always consider $V_{GS} - V_{TH}$ as the current-defining parameter. The effect of V_{DS} on I_D is usually considered an *error*, and it is studied in Chapter 5.

Subthreshold Conduction In our analysis of the MOSFET, we have assumed that the device turns off abruptly as V_{GS} drops below V_{TH}. In reality, for $V_{GS} \approx V_{TH}$, a "weak" inversion layer still exists and some current flows from D to S. Even for $V_{GS} < V_{TH}$, I_D is finite, but it exhibits an *exponential* dependence on V_{GS} [2, 3]. Called "subthreshold conduction," this effect can be formulated for V_{DS} greater than roughly 100 mV as

$$I_D = I_0 \exp\frac{V_{GS}}{\xi V_T} \tag{2.33}$$

where I_0 is proportional to W/L, $\xi > 1$ is a nonideality factor, and $V_T = kT/q$. We also say the device operates in "weak inversion." (Similarly, for $V_{GS} > V_{TH}$, we say the device operates in "strong inversion.") Except for ξ, (2.33) is similar to the exponential I_C/V_{BE} relationship of a bipolar transistor. The key point here is that as V_{GS} falls below V_{TH}, the drain current drops at a finite rate. With typical values of ξ, at room temperature V_{GS} must decrease by approximately 80 mV for I_D to decrease by one decade (Fig. 2.28). For example, if a threshold of 0.3 V is chosen in a process to allow low-voltage operation, then when V_{GS} is reduced to zero, the drain current decreases by only a factor of $10^{0.3\text{ V}/80\text{ mV}} = 10^{3.75} \approx 5.62 \times 10^3$. For example, if the transistor carries about 1 μA for $V_{GS} = V_{TH}$ and we have 100 million such devices, then

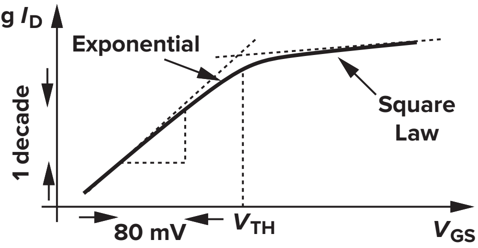


Figure 2.28 MOS subthreshold characteristics.

they draw 18 mA when they are nominally off. Especially problematic in large circuits such as memories, subthreshold conduction can result in significant power dissipation (or loss of analog information).

If a MOS device conducts for $V_{GS} < V_{TH}$, then how do we define the threshold voltage? Indeed, numerous definitions have been proposed. One possibility is to extrapolate, on a logarithmic vertical scale, the weak inversion and strong inversion characteristics and consider their intercept voltage as the threshold (Fig. 2.28).

We now reexamine Eq. (2.19) for the transconductance of a MOS device operating in the subthreshold region. Is it possible to achieve an arbitrarily high transconductance by increasing W while maintaining I_D constant? Is it possible to obtain a *higher* transconductance than that of a bipolar transistor (I_C/V_T) biased at the same current? Equation (2.19) was derived from the square-law characteristic $I_D = (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$. However, if W increases while I_D remains constant, then $V_{GS} \to V_{TH}$ and the device enters the subthreshold region. As a result, the transconductance is calculated from (2.33) to be $g_m = I_D/(\xi V_T)$, revealing that MOSFETs are still inferior to bipolar transistors in this respect.

At what overdrive voltage can we say the transistor goes from strong inversion to weak inversion? While somewhat arbitrary, this transition point can be defined as the overdrive voltage, $(V_{GS} - V_{TH})_1$, at which the corresponding transconductances would become equal for the same drain current:

$$\frac{I_D}{\xi V_T} = \frac{2I_D}{(V_{GS} - V_{TH})_1} \tag{2.34}$$

and hence

$$(V_{GS} - V_{TH})_1 = 2\xi V_T \tag{2.35}$$

For $\xi \approx 1.5$, this amounts to about 80 mV.

The exponential dependence of I_D upon V_{GS} in subthreshold operation may suggest the use of MOS devices in this regime so as to achieve a higher gain. However, since such conditions are met only by a large device width or low drain current, the speed of subthreshold circuits is severely limited.

▶ Example 2.8

Examine the behavior of a MOSFET as the drain “current density,” I_D/W, varies.

Solution

For a given drain current and device width, how do we determine the region of operation? We must consider the equations for both strong and weak inversion:

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.36}$$

$$I_D = \alpha\frac{W}{L}\exp\frac{V_{GS}}{\xi V_T} \tag{2.37}$$

where channel-length modulation is neglected and I_0 in Eq. (2.33) has been expressed as a proportionality factor, α, multiplied by W/L. What happens if the device is in strong inversion and we continue to reduce I_D while W/L is constant? Can V_{GS} simply approach V_{TH} to yield an arbitrarily small value for $(V_{GS} - V_{TH})^2$? Why does the square-law equation not hold as V_{GS} approaches V_{TH}?

To answer these questions, we return to the plot of Fig. 2.28 and observe that only currents beyond a certain level can be supported in strong inversion. In other words, for a given current and W/L, we must obtain V_{GS} from both the square-law and exponential equations and select the lower value:

$$V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox} W/L}} + V_{TH} \tag{2.38}$$

$$V_{GS} = \xi V_T \ln \frac{I_D}{\alpha W/L} \tag{2.39}$$

If I_D remains constant and W increases, V_{GS} falls and the device goes from strong inversion to weak inversion. ◀

Voltage Limitations A MOSFET experiences various undesirable effects if its terminal voltage differences exceed certain limits (if the device is "stressed"). At high gate-source voltages, the gate oxide breaks down irreversibly, damaging the transistor. In short-channel devices, an excessively large drain-source voltage widens the depletion region around the drain so much that it touches that around the source, creating a very large drain current. (This effect is called "punchthrough.") Even without breakdown, MOSFETs' characteristics can change permanently if the terminal voltage differences exceed a specified value. Such effects are described in Chapter 17.

2.4 ■ MOS Device Models

2.4.1 MOS Device Layout

For the developments in subsequent sections, it is beneficial to have some understanding of the layout of a MOSFET. We describe only a simple view here, deferring the fabrication details and structural subtleties to Chapters 18 and 19.

The layout of a MOSFET is determined by both the electrical properties required of the device in the circuit and the "design rules" imposed by the technology. For example, W/L is chosen to set the transconductance or other circuit parameters while the minimum L is dictated by the process. In addition to the gate, the source and drain areas must be defined properly as well.

Shown in Fig. 2.29 are the "bird's-eye view" and the top view of a MOSFET. The gate polysilicon and the source and drain terminals must be tied to metal (aluminum) wires that serve as interconnects with low resistance and capacitance. To accomplish this, one or more "contact windows" must be opened in each region, filled with metal, and connected to the upper metal wires. Note that the gate poly extends beyond the channel area by some amount to ensure reliable definition of the "edge" of the transistor.

The source and drain junctions play an important role in the performance. To minimize the capacitance of S and D, the total area of each junction must be minimized. We see from Fig. 2.29 that one dimension of the junctions is equal to W. The other dimension must be large enough to accommodate the contact windows and is specified by the technology design rules.[7]

[7] This dimension is typically three to four times the minimum allowable channel length.

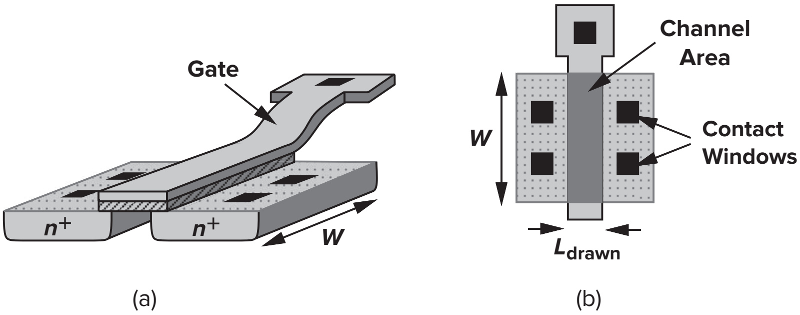


Figure 2.29 Bird's-eye and vertical views of a MOS device.

▶ **Example 2.9**

Draw the layout of the circuit shown in Fig. 2.30(a).

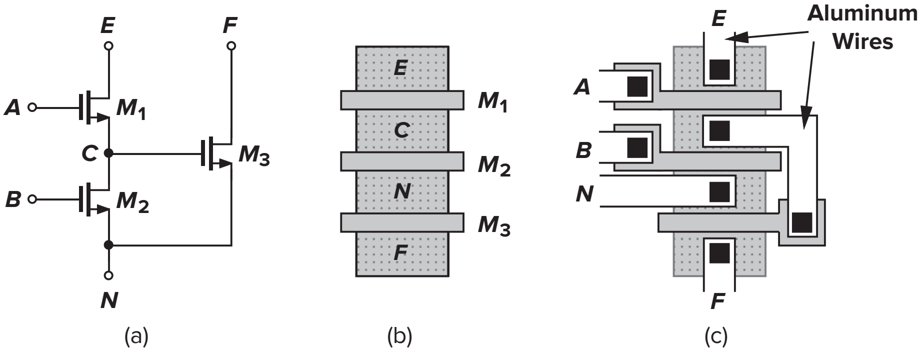


Figure 2.30

Solution

Observing that M_1 and M_2 share the same S/D junctions at node C and M_2 and M_3 also do so at node N, we surmise that the three transistors can be laid out as shown in Fig. 2.30(b). Connecting the remaining terminals, we obtain the layout in Fig. 2.30(c). Note that the gate polysilicon of M_3 cannot be directly tied to the source material of M_1, thus requiring a metal interconnect.

◀

2.4.2 MOS Device Capacitances

The basic quadratic I/V relationships derived in the previous section, along with corrections for body effect and channel-length modulation, provide some understanding of the low-frequency behavior of CMOS circuits. In many analog circuits, however, the capacitances associated with the devices must also be taken into account so as to predict the high-frequency behavior as well.

We expect that a capacitance exists between every two of the four terminals of a MOSFET (Fig. 2.31).[8] Moreover, the value of each of these capacitances may depend on the bias conditions of the transistor.

[8] The capacitance between S and D is negligible.

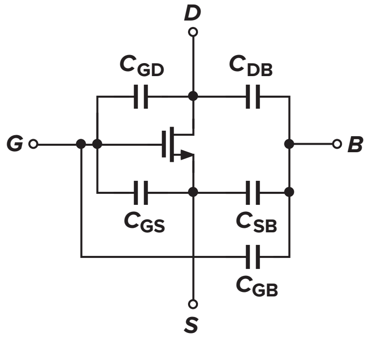


Figure 2.31 MOS capacitances.

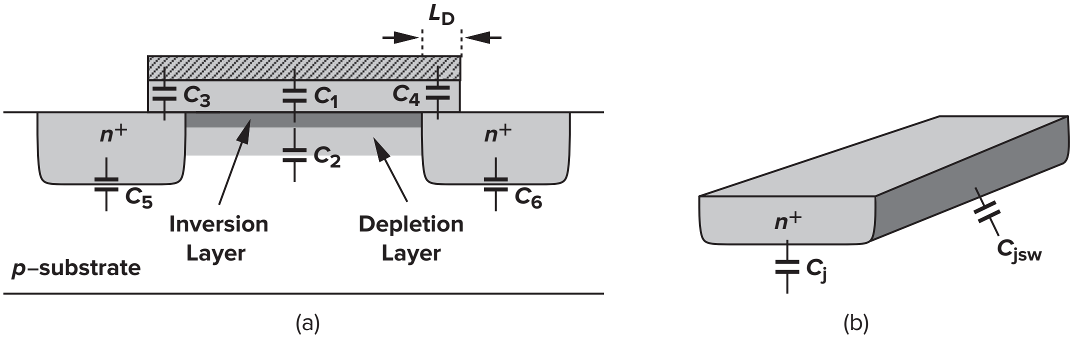
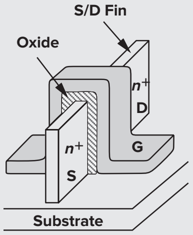


Figure 2.32 (a) MOS device capacitances; (b) decomposition of S/D junction capacitance into bottom-plate and sidewall components.

Nanometer Design Notes

New generations of CMOS technology incorporate the “FinFET” structure. Unlike the conventional “planar” device, the FinFET extends in the third dimension. As shown below, it consists of an n^+ wall (resembling a shark’s fin) and a gate that wraps around the wall. The transistor carries current from the source to the drain on the surfaces of the fin. Owing to the tight confinement of the electric field between the two vertical walls of the gate, the FinFET exhibits less channel-length modulation and subthreshold leakage. But where do the S/D contacts land? What other issues do we face in FinFETs? We return to these questions later in this book.



Considering the physical structure in Fig. 2.32(a), we identify the following: (1) the oxide capacitance between the gate and the channel, $C_1 = WLC_{ox}$; (2) the depletion capacitance between the channel and the substrate, $C_2 = WL\sqrt{q\epsilon_{si}N_{sub}/(4\Phi_F)}$; and (3) the capacitance due to the overlap of the gate poly with the source and drain areas, C_3 and C_4. Owing to fringing electric field lines, C_3 and C_4 cannot be simply written as WL_DC_{ox}, and are usually obtained by more elaborate calculations. The overlap capacitance per unit *width* is denoted by C_{ov} and expressed in F/m (or fF/μm). We simply multiply C_{ov} by W to find the gate-source and gate-drain overlap capacitances. (4) The junction capacitance between the source/drain areas and the substrate. As shown in Fig. 2.32(b), this last capacitance is decomposed into two components: the bottom-plate capacitance associated with the bottom of the junction, C_j, and the sidewall capacitance due to the perimeter of the junction, C_{jsw}. The distinction is necessary because different transistor geometries yield different area and perimeter values for the S/D junctions. We specify C_j and C_{jsw} as capacitance per unit *area* (in F/m^2) and unit *length* (in F/m), respectively. Thus, C_j is multiplied by the S/D area, and C_{jsw} by the S/D perimeter. Note that each junction capacitance can be expressed as $C_j = C_{j0}/[1 + V_R/(\Phi_B)]^m$, where V_R is the reverse voltage across the junction, Φ_B is the junction built-in potential, and m is a power typically in the range of 0.3 and 0.4.

▶ **Example 2.10**

Calculate the source and drain junction capacitances of the two structures shown in Fig. 2.33.

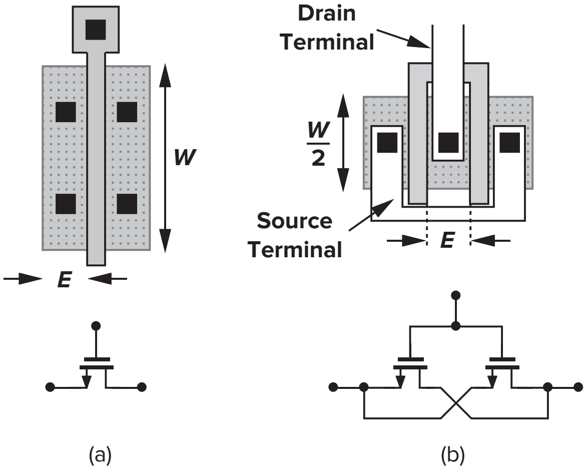


Figure 2.33

Solution

For the transistor in Fig. 2.33(a), we have

$$C_{DB} = C_{SB} = WEC_j + 2(W + E)C_{jsw} \tag{2.40}$$

whereas for that in Fig. 2.33(b),

$$C_{DB} = \frac{W}{2}EC_j + 2\left(\frac{W}{2} + E\right)C_{jsw} \tag{2.41}$$

$$C_{SB} = 2\left[\frac{W}{2}EC_j + 2\left(\frac{W}{2} + E\right)C_{jsw}\right] \tag{2.42}$$

$$= WEC_j + 2(W + 2E)C_{jsw} \tag{2.43}$$

Called a "folded" structure, the geometry in Fig. 2.33(b) exhibits substantially less drain junction capacitance than that in Fig. 2.33(a) while providing the same W/L.

In the above calculations, we have assumed that the total source or drain perimeter, $2(W + E)$, is multiplied by C_{jsw}. In reality, the capacitance of the inner sidewall (under the gate) may be different from that of the other sidewalls.[9] Nonetheless, we typically assume that all four sides have the same C_{jsw}. The error resulting from this assumption is negligible because each node in a circuit is connected to a number of other device capacitances as well.

We now derive the capacitances between terminals of a MOSFET in different regions of operation. If the device is off, $C_{GD} = C_{GS} = C_{ov}W$, and the gate-bulk capacitance consists of the series combination of the gate-oxide capacitance and the depletion-region capacitance [Fig. 2.32(a)], i.e., $C_{GB} = (WLC_{ox})C_d/(WLC_{ox} + C_d)$, where L is the effective length, $C_d = WL\sqrt{q\epsilon_{si}N_{sub}/(4\Phi_F)}$, and

[9] This is because the other sidewalls are surrounded by a "trench" (Chapter 18).

$\epsilon_{si} = \epsilon_{r,si} \times \epsilon_0 = 11.8 \times (8.85 \times 10^{-14})$ F/cm. The value of C_{SB} and C_{DB} is a function of the source and drain voltages with respect to the substrate.

If the device is in the deep triode region, i.e., if S and D have approximately equal voltages, then the gate-channel capacitance, WLC_{ox}, is divided equally between the gate and source terminals and the gate and drain terminals (Fig. 2.34). This is because a change of ΔV in the gate voltage draws equal amounts of charge from S and D. Thus, $C_{GD} = C_{GS} = WLC_{ox}/2 + WC_{ov}$.

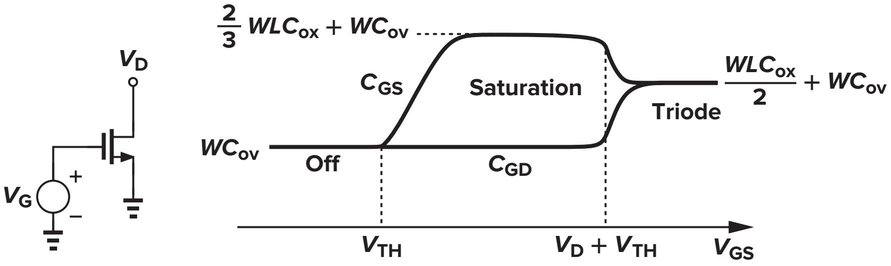


Figure 2.34 Variation of gate-source and gate-drain capacitances versus V_{GS}.

Let us now consider C_{GD} and C_{GS}. If in saturation, a MOSFET exhibits a gate-drain capacitance roughly equal to WC_{ov}. As for C_{GS}, we note that the potential difference between the gate and the channel varies from V_{GS} at the source to V_{TH} at the pinch-off point, resulting in a nonuniform vertical electric field in the gate oxide as we travel from the source to the drain. It can be proved that the equivalent capacitance of this structure, excluding the gate-source overlap capacitance, equals $(2/3)WLC_{ox}$ [1]. Thus, $C_{GS} = 2WL_{eff}C_{ox}/3 + WC_{ov}$. The behavior of C_{GD} and C_{GS} in different regions of operation is plotted in Fig. 2.34. Note that the above equations do not provide a smooth transition from one region of operation to another, creating convergence difficulties in simulation programs. This issue is revisited in Chapter 17.

The gate-bulk capacitance is usually neglected in the triode and saturation regions because the inversion layer acts as a "shield" between the gate and the bulk. In other words, if the gate voltage varies, the charge is supplied by the source and the drain rather than the bulk.

▶ **Example 2.11**

Sketch the capacitances of M_1 in Fig. 2.35 as V_X varies from zero to 3 V. Assume that $V_{TH} = 0.3$ V and $\lambda = \gamma = 0$.

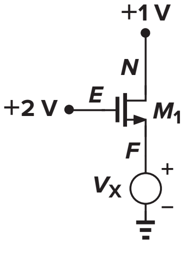


Figure 2.35

Solution

To avoid confusion, we label the three terminals as shown in Fig. 2.35 and denote the bulk by B. For $V_X \approx 0$, M_1 is in the triode region, $C_{EN} \approx C_{EF} = (1/2)WLC_{ox} + WC_{ov}$, and C_{FB} is maximum. The value of C_{NB} is independent of V_X. As V_X exceeds 1 V, the role of the source and drain is exchanged [Fig. 2.36(a)], eventually bringing M_1 out of the triode region for $V_X \geq 2\text{ V} - 0.3$ V. The variation of the capacitances is plotted in Figs. 2.36(b) and (c).

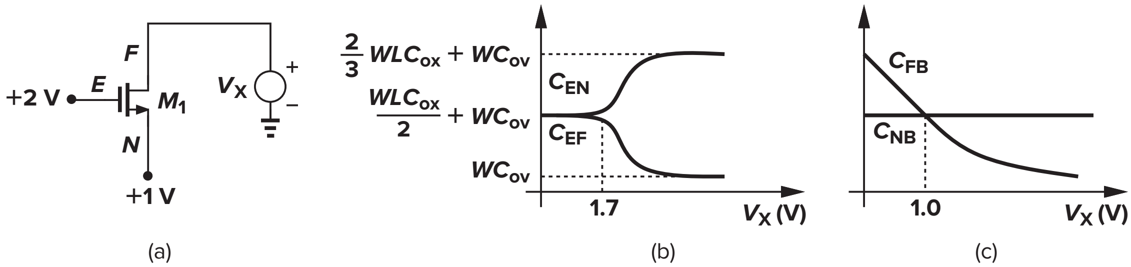


Figure 2.36

2.4.3 MOS Small-Signal Model

The quadratic characteristics described by (2.8) and (2.9) along with the voltage-dependent capacitances derived above form the large-signal model of MOSFETs. Such a model proves essential in analyzing circuits in which the signal significantly disturbs the bias points, particularly if nonlinear effects are of concern. By contrast, if the perturbation in bias conditions is small, a "small-signal" model, i.e., an approximation of the large-signal model around the operating point, can be employed to simplify the calculations. Since in many analog circuits, MOSFETs are biased in the saturation region, we derive the corresponding small-signal model here. For transistors operating as switches, a linear resistor given by (2.11) together with device capacitances serves as a rough small-signal equivalent.

We derive the small-signal model by producing a small increment in one bias parameter and calculating the resulting increment in other bias parameters. Specifically, we (1) apply certain bias voltages to the terminals of the device, (2) increment the potential difference between *two* of the terminals while other terminal voltages remain constant, and (3) measure the resulting change in all terminal currents. If we change the voltage between two terminals by ΔV and measure a current change of ΔI in some branch, we can model the effect by a voltage-dependent current source. Let us apply a change to the gate-source voltage, $\Delta V = V_{GS}$, where V_{GS} is a small-signal quantity.[10] The drain current therefore changes by $g_m V_{GS}$ and is modeled by a voltage-dependent current source tied between the drain and source terminals [Fig. 2.37(a)]. The gate current is very small and its change is negligible, thus requiring no representation here. The result is the small-signal model of an ideal MOSFET—the model that an analog designer applies to most devices in a circuit at first glance.

Owing to channel-length modulation, the drain current also varies with the drain-source voltage. This effect can be modeled by a voltage-dependent current source [Fig. 2.37(b)], but a current source whose value linearly depends on the voltage across it is equivalent to a linear resistor [Fig. 2.37(c)] (why?). Tied between D and S, the resistor is given by

$$r_O = \frac{\partial V_{DS}}{\partial I_D} \tag{2.44}$$

$$= \frac{1}{\partial I_D / \partial V_{DS}} \tag{2.45}$$

$$= \frac{1}{\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \cdot \lambda} \tag{2.46}$$

[10] In this book, we use uppercase letters to denote large-signal or small-signal quantities. The distinction will be clear from the context.

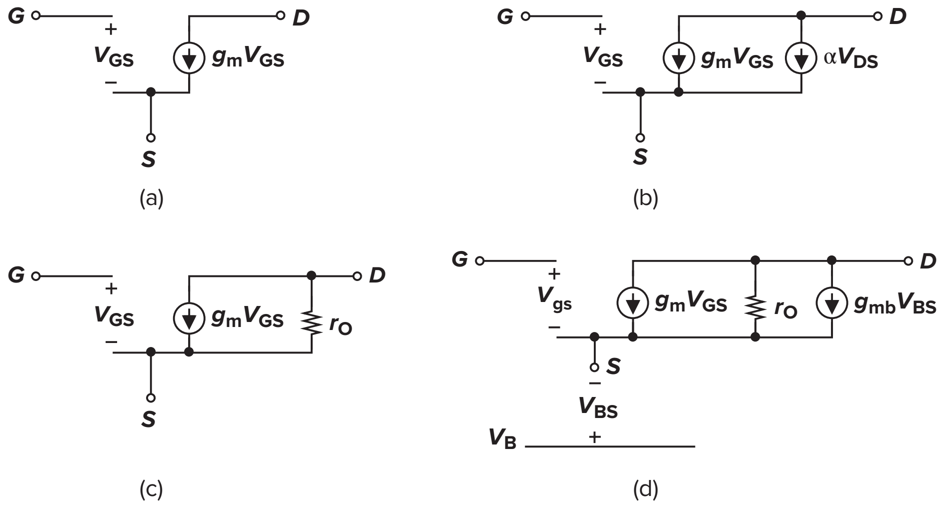


Figure 2.37 (a) Basic MOS small-signal model; (b) channel-length modulation represented by a dependent current source; (c) channel-length modulation represented by a resistor; (d) body effect represented by a dependent current source.

$$\approx \frac{1+\lambda V_{DS}}{\lambda I_D} \tag{2.47}$$

$$\approx \frac{1}{\lambda I_D} \tag{2.48}$$

where it is assumed that $\lambda V_{DS} \ll 1$. As seen throughout this book, the output resistance, r_O, affects the performance of many analog circuits. For example, r_O limits the maximum voltage gain of most amplifiers.

Now recall that the bulk potential influences the threshold voltage and hence the gate-source overdrive. As demonstrated in Example 2.3, with all other terminals held at a constant voltage, the drain current is a function of the bulk voltage. That is, the bulk behaves as a second gate. Modeling this dependence by a current source connected between D and S [Fig. 2.37(d)], we write the value as $g_{mb}V_{bs}$, where $g_{mb} = \partial I_D/\partial V_{BS}$. In the saturation region, g_{mb} can be expressed as

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} \tag{2.49}$$

$$= \mu_n C_{ox}\frac{W}{L}(V_{GS}-V_{TH})\left(-\frac{\partial V_{TH}}{\partial V_{BS}}\right) \tag{2.50}$$

We also have

$$\frac{\partial V_{TH}}{\partial V_{BS}} = -\frac{\partial V_{TH}}{\partial V_{SB}} \tag{2.51}$$

$$= -\frac{\gamma}{2}(2\Phi_F + V_{SB})^{-1/2} \tag{2.52}$$

Thus,

$$g_{mb} = g_m\frac{\gamma}{2\sqrt{2\Phi_F + V_{SB}}} \tag{2.53}$$

$$= \eta g_m \tag{2.54}$$

where $\eta = g_{mb}/g_m$ and is typically around 0.25. As expected, g_{mb} is proportional to γ. Equation (2.53) also suggests that the incremental body effect becomes less pronounced as V_{SB} increases. Note that $g_m V_{GS}$ and $g_{mb} V_{BS}$ have the same polarity, i.e., raising the gate voltage has the same effect as raising the bulk potential.

The model in Fig. 2.37(d) is adequate for most low-frequency small-signal analyses. In reality, each terminal of a MOSFET exhibits a finite ohmic resistance resulting from the resistivity of the material (and the contacts), but proper layout can minimize such resistances. For example, consider the two structures of Fig. 2.33, repeated in Fig. 2.38 along with the gate distributed resistance. We note that folding reduces the gate resistance by a factor of four.

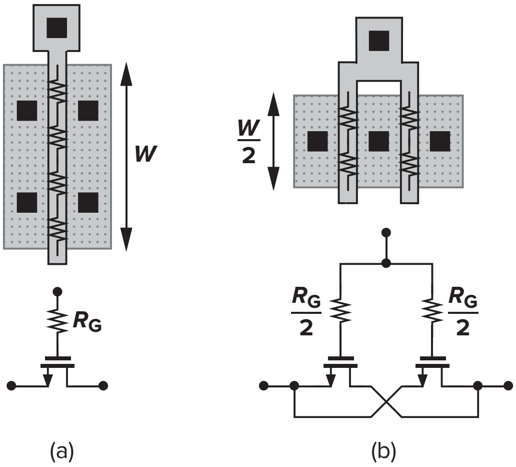


Figure 2.38 Reduction of gate resistance by folding.

Shown in Fig. 2.39, the complete small-signal model includes the device capacitances as well. The value of each capacitance is calculated according to the equations derived in Sec. 2.4.2. The reader may wonder how a complex circuit is analyzed intuitively if each transistor must be replaced by the model of Fig. 2.39. The first step is to determine the *simplest* device model that can represent the role of each transistor with reasonable accuracy. We provide some guidelines for this task at the end of Chapter 3.

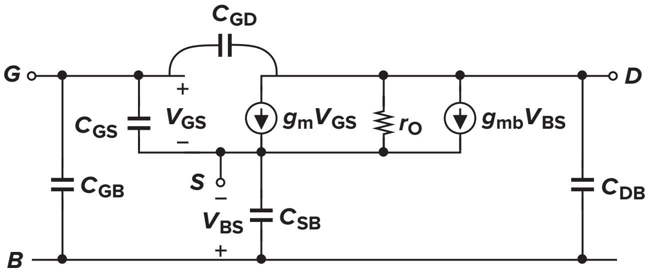


Figure 2.39 Complete MOS small-signal model.

▶ **Example 2.12**

Sketch g_m and g_{mb} of M_1 in Fig. 2.40 as a function of the bias current I_1.

Solution

Since $g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$, we have $g_m \propto \sqrt{I_1}$. The dependence of g_{mb} upon I_1 is less straightforward. As I_1 increases, V_X decreases, and so does V_{SB}.

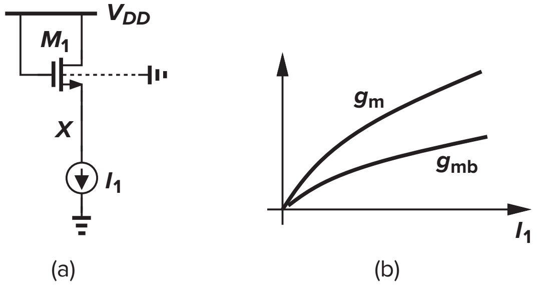


Figure 2.40

PMOS Small-Signal Model The derivation of the small-signal model seeks changes in the terminal currents due to changes in the terminal voltage differences. As such, this derivation yields *exactly* the same model for PMOS devices as for NMOS devices. For example, consider the arrangement shown in Fig. 2.41(a), where the voltage source V_1 is changed by a small amount and the change in I_D is measured (while M_1 remains in saturation). Suppose V_1 becomes more positive, making V_{GS} more *negative*. Since the transistor now has a greater overdrive, it carries a higher current, and hence I_D becomes more *negative*. (Recall that I_D, in the direction shown here, is negative because the actual current of holes flows from the source to the drain.) Thus, a negative ΔV_{GS} leads to a negative ΔI_D. Conversely, a positive ΔV_{GS} produces a positive ΔI_D, as is the case for an NMOS device.

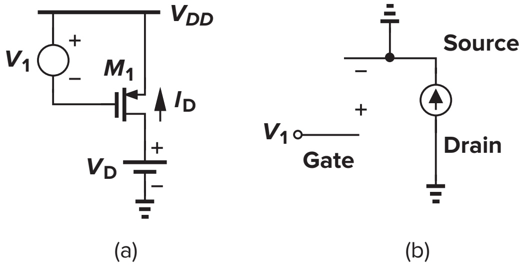


Figure 2.41 (a) Small-signal test of a PMOS device, and (b) small-signal model.

In our circuit diagrams, we usually draw the PMOS devices with their source terminals on top and their drain terminals on the bottom because the former are at a more positive voltage. This practice may cause confusion in drawing small-signal models. Let us draw the small-signal equivalent of the above circuit, assuming no channel-length modulation. Depicted in Fig. 2.41(b), the model shows the voltage-dependent current source pointing *upward*, giving the (wrong) impression that the direction of the current in the PMOS model is the opposite of that in the NMOS model. The reader is cautioned to avoid this confusion and bear in mind that the small-signal models of NMOS and PMOS transistors are identical.

Unless otherwise stated, in this book we assume that the bulk of all NFETs is tied to the most negative supply (usually the ground) and that of PFETs to the most positive supply (usually V_{DD}).

2.4.4 MOS SPICE models

In order to represent the behavior of transistors in circuit simulations, simulators such as SPICE and Cadence require an accurate model for each device. Over the last three decades, MOS modeling has made tremendous progress, reaching sophisticated levels so as to represent high-order effects in short-channel devices.

In this section, we describe the simplest MOS SPICE model, known as "Level 1," and provide typical values for each parameter in the model corresponding to a 0.5-μm technology. Chapter 17 describes more accurate SPICE models. Table 2.1 shows the model parameters for NMOS and PMOS devices. The parameters are defined as follows:

Table 2.1 Level 1 SPICE models for NMOS and PMOS devices.

NMOS Model			
LEVEL = 1	VTO = 0.7	GAMMA = 0.45	PHI = 0.9
NSUB = 9e+14	LD = 0.08e−6	UO = 350	LAMBDA = 0.1
TOX = 9e−9	PB = 0.9	CJ = 0.56e−3	CJSW = 0.35e−11
MJ = 0.45	MJSW = 0.2	CGDO = 0.4e−9	JS = 1.0e−8
PMOS Model			
LEVEL = 1	VTO = −0.8	GAMMA = 0.4	PHI = 0.8
NSUB = 5e+14	LD = 0.09e−6	UO = 100	LAMBDA = 0.2
TOX = 9e−9	PB = 0.9	CJ = 0.94e−3	CJSW = 0.32e−11
MJ = 0.5	MJSW = 0.3	CGDO = 0.3e−9	JS = 0.5e−8

VTO: threshold voltage with zero V_{SB} (unit: V)
GAMMA: body-effect coefficient (unit: $\mathrm{V}^{1/2}$)
PHI: $2\Phi_F$ (unit: V)
TOX: gate-oxide thickness (unit: m)
NSUB: substrate doping (unit: cm^{-3})
LD: source/drain side diffusion (unit: m)
UO: channel mobility (unit: cm^2/V/s)
LAMBDA: channel-length modulation coefficient (unit: V^{-1})
CJ: source/drain bottom-plate junction capacitance per unit area (unit: F/m^2)
CJSW: source/drain sidewall junction capacitance per unit length (unit: F/m)
PB: source/drain junction built-in potential (unit: V)
MJ: exponent in CJ equation (unitless)
MJSW: exponent in CJSW equation (unitless)
CGDO: gate-drain overlap capacitance per unit width (unit: F/m)
CGSO: gate-source overlap capacitance per unit width (unit: F/m)
JS: source/drain leakage current per unit area (unit: A/m^2)

2.4.5 NMOS Versus PMOS Devices

In most CMOS technologies, PMOS devices are quite inferior to NMOS transistors. For example, due to the lower mobility of holes, $\mu_p C_{ox} \approx 0.5\mu_n C_{ox}$, yielding low current drive and transconductance. Moreover, for given dimensions and bias currents, NMOS transistors exhibit a higher output resistance, providing more ideal current sources and higher gain in amplifiers. For these reasons, incorporating NFETs rather than PFETs wherever possible is preferred.[11]

2.4.6 Long-Channel Versus Short-Channel Devices

In this chapter, we have employed a very simple view of MOSFETs so as to understand the basic principles of their operation. Most of our treatment is valid for "long-channel" devices, e.g., transistors having a minimum length of a few microns. Many of the relationships derived here must be reexamined and revised for short-channel MOSFETs. Furthermore, the SPICE models necessary for simulation of today's devices

[11] One exception is when flicker noise is critical (Chapter 7).

are much more sophisticated than the Level 1 model. For example, the intrinsic gain, $g_m r_O$, calculated from the device parameters in Table 2.1 is much higher than actual values. These issues are studied in Chapter 17.

The reader may wonder why we begin with a simplistic view of devices if such a view does not lead to high accuracy in predicting the performance of circuits. The key point is that the simple model provides a great deal of intuition that is necessary in analog design. As we will see throughout this book, we often encounter a trade-off between intuition and rigor, and our approach is to establish the intuition first and gradually complete our understanding so as to achieve rigor as well.

2.5 ■ Appendix A: FinFETs

New CMOS technology generations ("nodes") have migrated from the two-dimensional transistor structure to a three-dimensional geometry called the "FinFET." This device exhibits superior performance as channel lengths fall below approximately 20 nm. In fact, FinFET I/V characteristics are closer to square-law behavior, making our simple large-signal mode relevant again.

Shown in Fig. 2.42(a), the FinFET consists of a vertical silicon "fin," a dielectric (e.g., oxide) layer deposited over the fin, and a polysilicon or metal gate created over the dielectric layer. Controlled by the gate voltage, the current flows from one end of the fin to the other. The top view looks similar to that of a planar MOSFET [Fig. 2.42(b)].

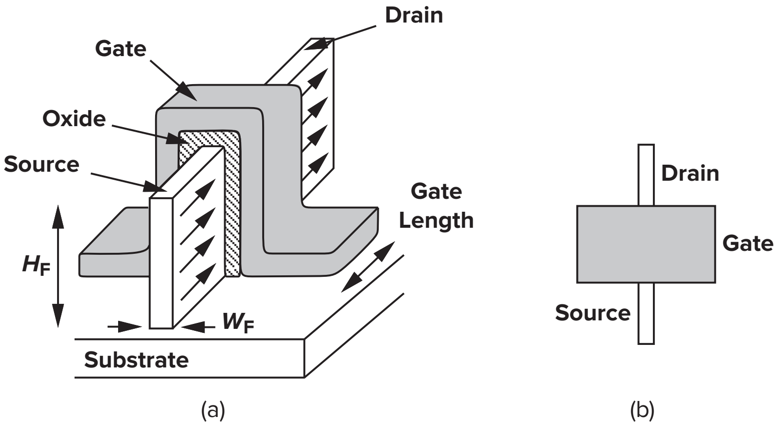


Figure 2.42 (a) FinFET structure, and (b) top view.

As depicted in Fig. 2.42(a), the gate length can be readily identified, but how about the gate width? We note that the current flows on *three* facets of the fin. The equivalent channel width is therefore equal to the sum of the fin's width, W_F, and twice its height, H_F: $W = W_F + 2H_F$. Typically, $W_F \approx 6$ nm and $H_F \approx 50$ nm.

Since H_F is not under the circuit designer's control, it appears that W_F can be chosen so that $W_F + 2H_F$ yields the desired transistor width. However, W_F affects device imperfections such as source and drain series resistance, channel-length modulation, subthreshold conduction, etc. For this reason, the fin width is also fixed, dictating discrete values for the transistor width. For example, if $W_F + 2H_F = 100$ nm, then wider transistors can be obtained only by increasing the number of fins and only in 100-nm increments (Fig. 2.43). The spacing between the fins, S_F, also plays a significant role in the performance and is typically fixed.

Due to the small dimensions of the intrinsic FinFET, the gate and S/D contacts must be placed away from the core of the device. Figure 2.44 shows the details for a single- and a double-fin structure.

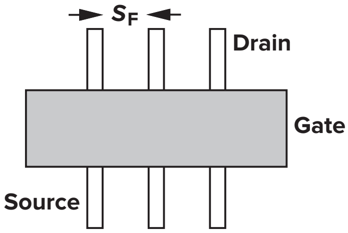


Figure 2.43 FinFET with multiple fins.

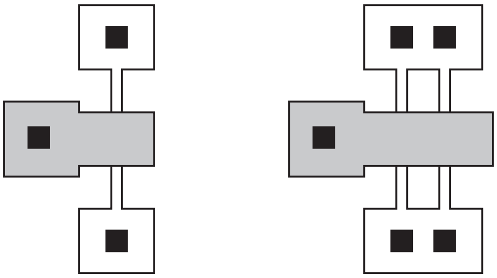


Figure 2.44 Layout of single- and double-fin transistors.

2.6 ■ Appendix B: Behavior of a MOS Device as a Capacitor

In this chapter, we have limited our treatment of MOS devices to a basic level. However, the behavior of a MOSFET as a capacitor merits some attention. Recall that if the source, drain, and bulk of an NFET are grounded and the gate voltage rises, an inversion layer begins to form for $V_{GS} \approx V_{TH}$. We also noted that for $0 < V_{GS} < V_{TH}$, the device operates in the subthreshold region.

Now consider the NFET of Fig. 2.45. The transistor can be considered a two-terminal device, and hence its capacitance can be examined for different gate voltages. Let us begin with a very *negative* gate-source voltage. The negative potential on the gate attracts the holes in the substrate to the oxide interface. We say that the MOSFET operates in the "accumulation" region. The two-terminal device can be viewed as a capacitor having a unit-area capacitance of C_{ox} because the two "plates" of the capacitor are separated by t_{ox}.

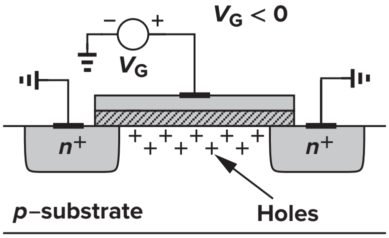


Figure 2.45 NMOS operating in accumulation mode.

As V_{GS} rises, the density of holes at the interface falls, a depletion region begins to form under the oxide, and the device enters weak inversion. In this mode, the capacitance consists of the series combination of C_{ox} and C_{dep}. Finally, as V_{GS} exceeds V_{TH}, the oxide-silicon interface sustains a channel and the unit-area capacitance returns to C_{ox}. Figure 2.46 plots the behavior.

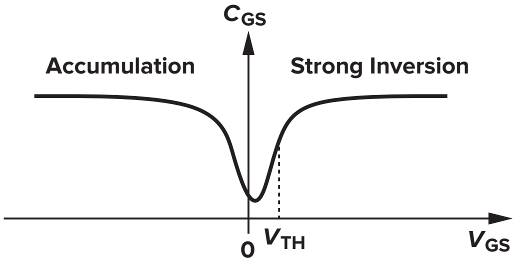


Figure 2.46 Capacitance-voltage characteristic of an NMOS device.

References

[1] R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed. (New York: Wiley, 1986).
[2] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. (Boston: McGraw-Hill, 1999).
[3] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices* (New York: Cambridge University Press, 1998).

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume $V_{DD} = 3$ V where necessary.

2.1. For $W/L = 50/0.5$, plot the drain current of an NFET and a PFET as a function of $|V_{GS}|$ as $|V_{GS}|$ varies from 0 to 3 V. Assume that $|V_{DS}| = 3$ V.

2.2. For $W/L = 50/0.5$ and $|I_D| = 0.5$ mA, calculate the transconductance and output impedance of both NMOS and PMOS devices. Also, find the "intrinsic gain," defined as $g_m r_O$.

2.3. Derive expressions for $g_m r_O$ in terms of I_D and W/L. Plot $g_m r_O$ as a function of I_D with L as a parameter. Note that $\lambda \propto 1/L$.

2.4. Plot I_D versus V_{GS} for a MOS transistor (a) with V_{DS} as a parameter, and (b) with V_{BS} as a parameter. Identify the break points in the characteristics.

2.5. Sketch I_X and the transconductance of the transistor as a function of V_X for each circuit in Fig. 2.47 as V_X varies from 0 to V_{DD}. In part (a), assume that V_X varies from 0 to 1.5 V.

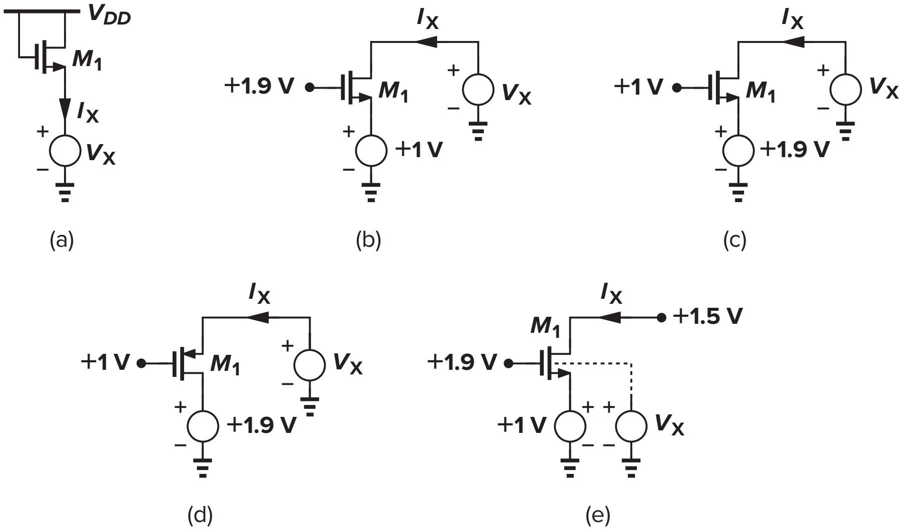


Figure 2.47

2.6. Sketch I_X and the transconductance of the transistor as a function of V_X for each circuit in Fig. 2.48 as V_X varies from 0 to V_{DD}.

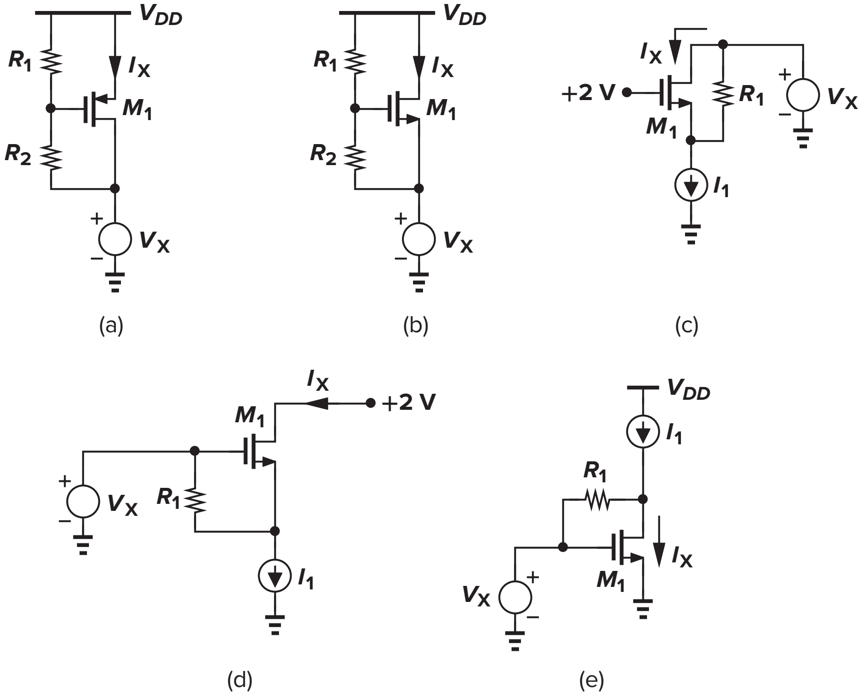


Figure 2.48

2.7. Sketch V_{out} as a function of V_{in} for each circuit in Fig. 2.49 as V_{in} varies from 0 to V_{DD}.

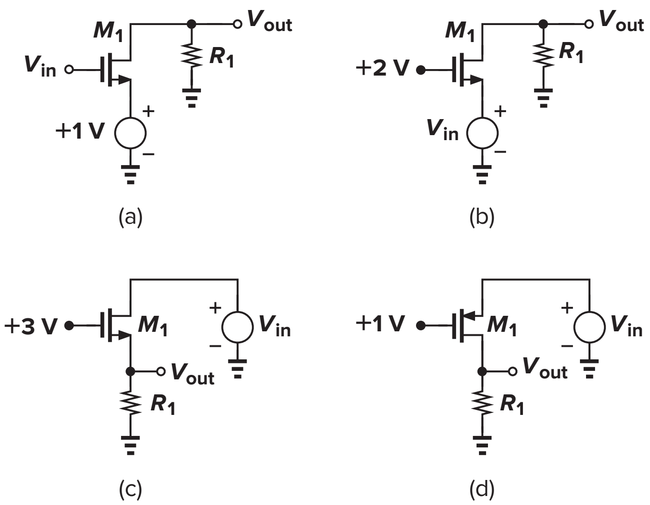
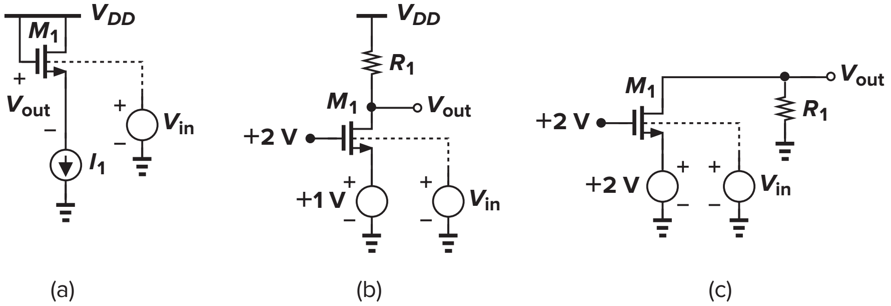


Figure 2.49

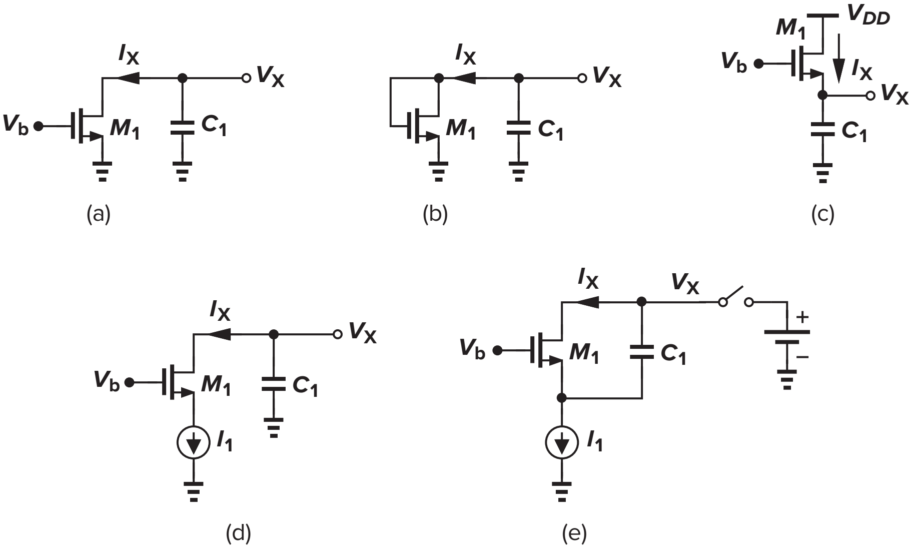
2.8. Sketch V_{out} as a function of V_{in} for each circuit in Fig. 2.50 as V_{in} varies from 0 to V_{DD}.



(a) (b) (c)

Figure 2.50

2.9. Sketch V_X and I_X as a function of time for each circuit in Fig. 2.51. The initial voltage of C_1 is equal to 3 V. In part (e), assume that the switch turns off at $t = 0$.

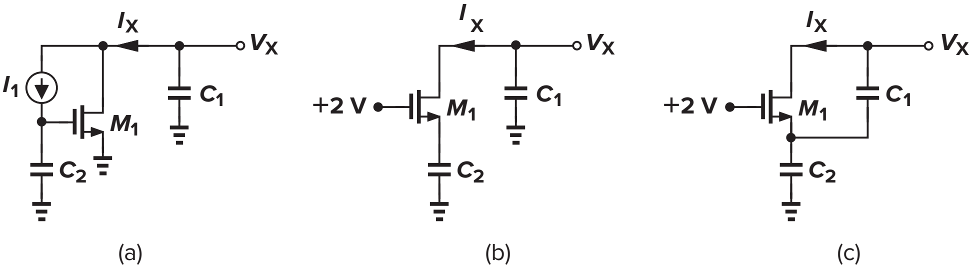


(a) (b) (c)

(d) (e)

Figure 2.51

2.10. Sketch V_X and I_X as a function of time for each circuit in Fig. 2.52. The initial voltage of each capacitor is shown.



(a) (b) (c)

Figure 2.52

2.11. Sketch V_X as a function of time for each circuit in Fig. 2.53. The initial voltage of each capacitor is shown.

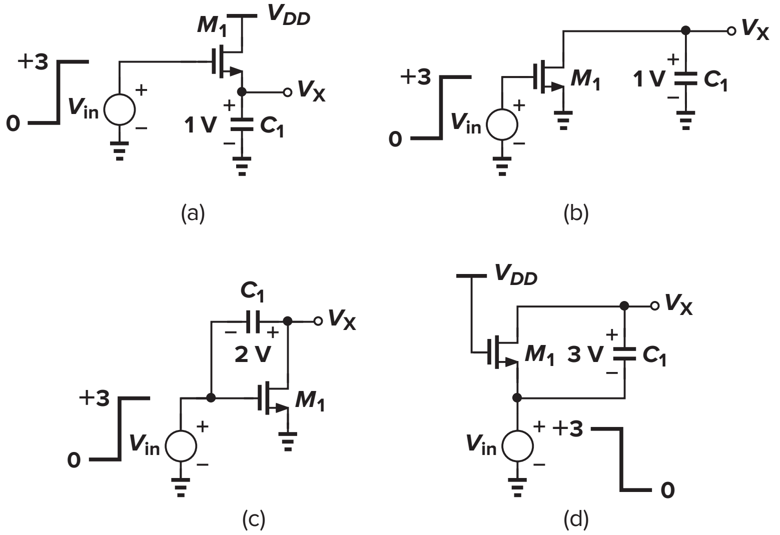


Figure 2.53

2.12. Sketch V_X as a function of time for each circuit in Fig. 2.54. The initial voltage of each capacitor is shown.

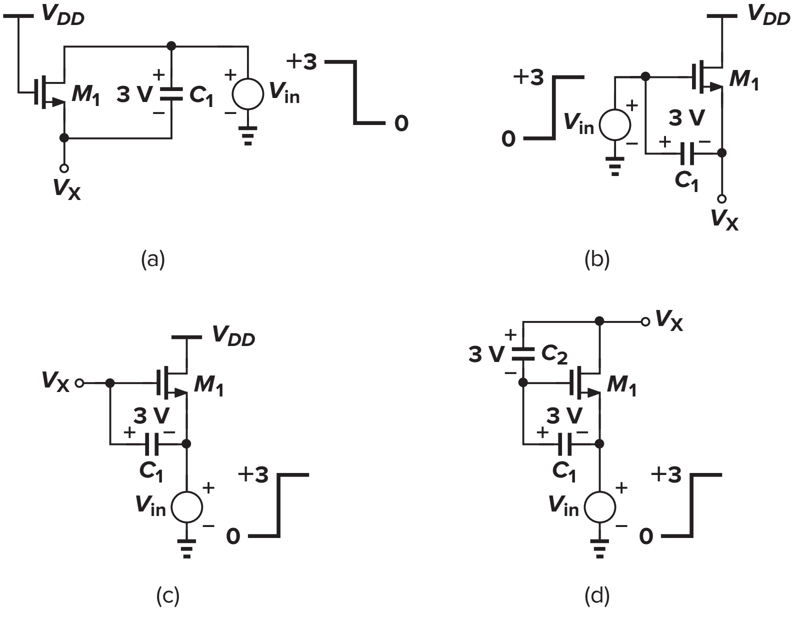


Figure 2.54

2.13. The transit frequency, f_T, of a MOSFET is defined as the frequency at which the small-signal current gain of the device drops to unity while the source and drain terminals are held at ac ground.

(a) Prove that

$$f_T = \frac{g_m}{2\pi(C_{GD} + C_{GS})} \tag{2.55}$$

Note that f_T does not include the effect of the S/D junction capacitance.

(b) Suppose the gate resistance, R_G, is significant and the device is modeled as a distributed set of n transistors, each with a gate resistance equal to R_G/n. Prove that the f_T of the device is independent of R_G and still equal to the value given above.

(c) For a given bias current, the minimum allowable drain-source voltage for operation in saturation can be reduced only by increasing the width and hence the capacitances of the transistor. Using square-law characteristics, prove that

$$f_T = \frac{\mu_n}{2\pi}\frac{V_{GS} - V_{TH}}{L^2} \tag{2.56}$$

This relation indicates how the speed is limited as a device is designed to operate with lower supply voltages.

2.14. Calculate the f_T of a MOS device in the subthreshold region and compare the result with that obtained in Prob. 2.13.

2.15. For a saturated NMOS device having $W = 50\ \mu$m and $L = 0.5\ \mu$m, calculate all the capacitances. Assume that the minimum (lateral) dimension of the S/D areas is 1.5 μm and that the device is folded as shown in Fig. 2.33(b). What is the f_T if the drain current is 1 mA?

2.16. Consider the structure shown in Fig. 2.55. Determine I_D, as a function of V_{GS} and V_{DS}, and prove that the structure can be viewed as a single transistor having an aspect ratio $W/(2L)$. Assume that $\lambda = \gamma = 0$.

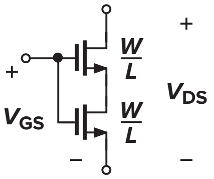


Figure 2.55

2.17. For an NMOS device operating in saturation, plot W/L versus $V_{GS} - V_{TH}$ if (a) I_D is constant, and (b) g_m is constant.

2.18. Explain why the structures shown in Fig. 2.56 cannot operate as current sources even though the transistors are in saturation.

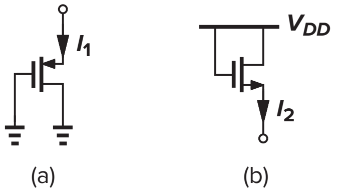


Figure 2.56

2.19. Considering the body effect as "back-gate effect," explain intuitively why γ is directly proportional to $\sqrt{N_{sub}}$ and inversely proportional to C_{ox}.

2.20. A "ring" MOS structure is shown in Fig. 2.57. Explain how the device operates and estimate its equivalent aspect ratio. Compare the drain junction capacitance of this structure with that of the devices shown in Fig. 2.33.

2.21. Suppose we have received an NMOS transistor in a package with four unmarked pins. Describe the minimum number of dc measurement steps using an ohmmeter that is necessary to determine the gate, source/drain, and bulk terminals of the device.

2.22. Repeat Prob. 2.21 if the type of the device (NFET or PFET) is not known.

2.23. For an NMOS transistor, the threshold voltage is known, but $\mu_n C_{ox}$ and W/L are not. Assume that $\lambda = \gamma = 0$. If we cannot measure C_{ox} independently, is it possible to devise a sequence of dc measurement tests to determine $\mu_n C_{ox}$ and W/L? What if we have two transistors and we know that one has twice the aspect ratio of the other?

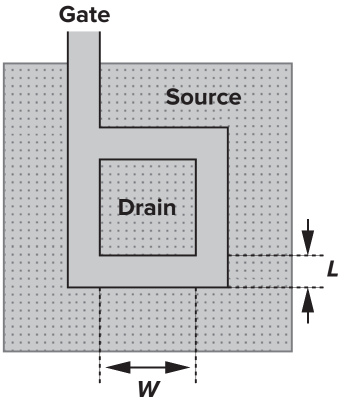


Figure 2.57

2.24. Sketch I_X versus V_X for each of the composite structures shown in Fig. 2.58 with V_G as a parameter. Also, sketch the equivalent transconductance. Assume that $\lambda = \gamma = 0$.

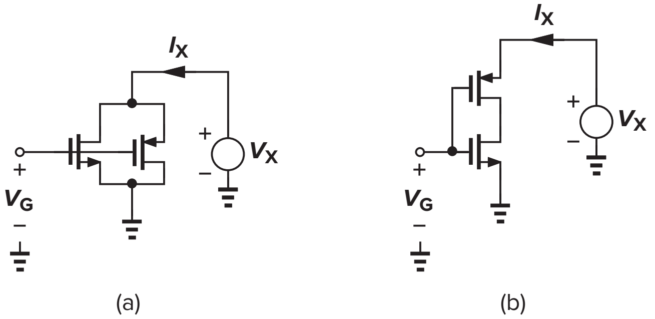


Figure 2.58

2.25. An NMOS current source with $I_D = 0.5$ mA must operate with drain-source voltages as low as 0.4 V. If the minimum required output impedance is 20 kΩ, determine the width and length of the device. Calculate the gate-source, gate-drain, and drain-substrate capacitance if the device is folded as in Fig. 2.33 and $E = 3\ \mu$m.

2.26. Consider the circuit shown in Fig. 2.59, where the initial voltage at node X is equal to V_{DD}. Assuming that $\lambda = \gamma = 0$ and neglecting other capacitances, plot V_X and V_Y versus time if (a) V_{in} is a positive step with amplitude $V_0 > V_{TH}$, and (b) V_{in} is a negative step with amplitude $V_0 = V_{TH}$.

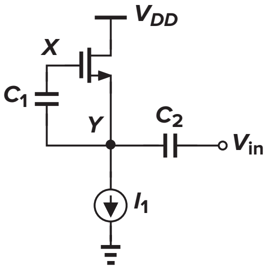


Figure 2.59

2.27. An NMOS device operating in the subthreshold region has a ζ of 1.5. What variation in V_{GS} results in a tenfold change in I_D? If $I_D = 10\ \mu$A, what is g_m?

2.28. Consider an NMOS device with $V_G = 1.5$ V and $V_S = 0$. Explain what happens if we continually decrease V_D below zero or increase V_{sub} above zero.

2.29. Consider the arrangement shown in Fig. 2.60. Explain what happens to the pinch-off point as V_G increases.

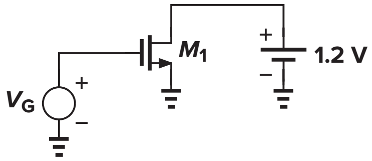


Figure 2.60

2.30. From Fig. 2.20, plot I_D vs. $V_{GS} - V_{TH}$ if W/L is constant, $V_{GS} - V_{TH}$ vs. I_D if W/L is constant, and W/L vs. $V_{GS} - V_{TH}$ if I_D is constant.

2.31. Plotted in Fig. 2.61 are the charactersitics of a square-law NMOS device with $W/L_{drawn} = 5$ μm/40 nm and $t_{ox} = 18$ Å. Here, V_{GS} is incremented in equal steps. Estimate μ_n, V_{TH}, λ, and the V_{GS} steps.

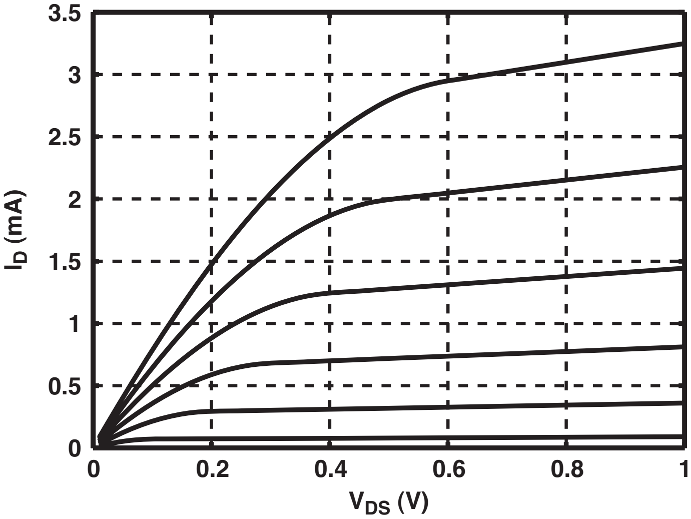


Figure 2.61

CHAPTER

3

Single-Stage Amplifiers

Amplification is an essential function in most analog (and many digital) circuits. We amplify an analog or digital signal because it may be too small to drive a load, overcome the noise of a subsequent stage, or provide logical levels to a digital circuit. Amplification also plays a critical role in feedback systems (Chapter 8).

In this chapter, we study the low-frequency behavior of single-stage CMOS amplifiers. Analyzing both the large-signal and the small-signal characteristics of each circuit, we develop intuitive techniques and models that prove useful in understanding more complex systems. An important part of a designer's job is to use proper approximations so as to create a simple mental picture of a complicated circuit. The intuition thus gained makes it possible to formulate the behavior of most circuits by inspection rather than by lengthy calculations.

Following a brief review of basic concepts, we describe in this chapter four types of amplifiers: common-source and common-gate topologies, source followers, and cascode configurations. In each case, we begin with a simple model and gradually add second-order phenomena such as channel-length modulation and body effect.

3.1 ■ Applications

Do you carry an amplifier? In all likelihood, yes. Your mobile phone, laptop, and digital camera all incorporate various types of amplifiers. The receiver in your phone must sense and amplify small signals received by the antenna, thus requiring a "low-noise" amplifier (LNA) at the front end (Fig. 3.1). As the signal travels down the receive chain, it must be further amplified by additional stages so as to reach an acceptably high level. This proves difficult because, in addition to the small desired signal, the antenna picks up other strong signals ("interferers") that are transmitted by various other users in the same vicinity. Your phone's transmitter, too, employs amplifiers: to amplify the signal generated by the microphone and, eventually, the signal delivered to the antenna. The "power amplifier" (PA) necessary for such delivery draws the most energy from the battery and still presents interesting challenges.

3.2 ■ General Considerations

An ideal amplifier generates an output, $y(t)$, that is a linear replica of the input, $x(t)$:

$$y(t) = \alpha_1 x(t) \tag{3.1}$$

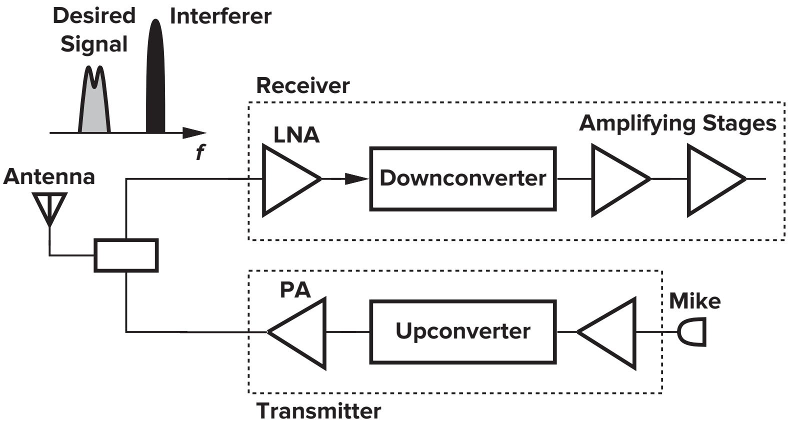


Figure 3.1 General RF transceiver.

where α_1 denotes the gain. Since the output signal is in fact superimposed on a bias (dc operating) point, α_0, we can write the overall output as $y(t) = \alpha_0 + \alpha_1 x(t)$. In this case, the input-output (large-signal) characteristic of the circuit is a straight line [Fig. 3.2(a)]. However, as the signal excursions become larger and the bias point of the transistor(s) is disturbed substantially, the gain (the slope of the characteristic) begins to *vary* [Fig. 3.2(b)]. We approximate this nonlinear characteristic by a polynomial:

$$y(t) = \alpha_0 + \alpha_1 x(t) + \alpha_2 x^2(t) + \cdots + \alpha_n x^n(t) \tag{3.2}$$

A nonlinear amplifier distorts the signal of interest or creates unwanted interactions among several signals that may coexist at the input. We return to the problem of nonlinearity in Chapter 14.

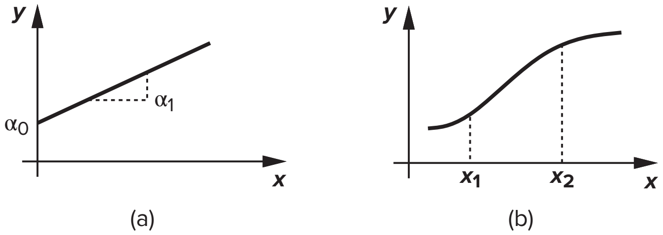


Figure 3.2 Input-output characteristic of a (a) linear and (b) nonlinear system.

What aspects of the performance of an amplifier are important? In addition to gain and speed, such parameters as power dissipation, supply voltage, linearity, noise, or maximum voltage swings may be important. Furthermore, the input and output impedances determine how the circuit interacts with the preceding and subsequent stages. In practice, most of these parameters trade with each other, making the design a multidimensional optimization problem. Illustrated in the "analog design octagon" of Fig. 3.3, such trade-offs present many challenges in the design of high-performance amplifiers, requiring intuition and experience to arrive at an acceptable compromise.

Table 3.1 gives a preview of the amplifier topologies studied in this chapter, indicating the much wider use of the common-source (CS) stage than other circuit configurations. For these amplifiers, we must (1) set up proper bias conditions so that each transistor provides the necessary transconductance and output resistance with certain quiescent currents and voltages, and (2) analyze the circuit's behavior as the input and output signals cause small or large departures from the bias input (small-signal and large-signal analyses, respectively). We deal with the latter task here and defer the former to Chapter 5.

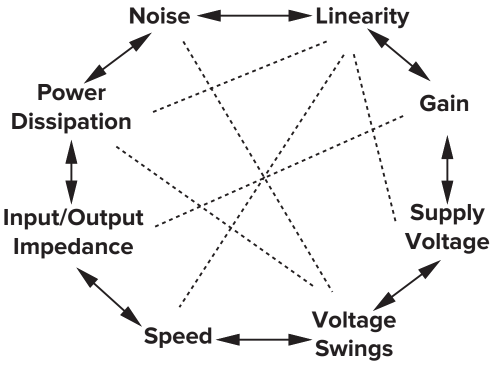


Figure 3.3 Analog design octagon.

Table 3.1 Amplifier categories.

Common-Source Stage	Source Follower	Common-Gate Stage	Cascode
With Resistive Load	With Resistive Bias	With Resistive Load	Telescopic
With Diode-Connected Load	With Current-Source Bias	With Current-Source Load	Folded
With Current-Source Load			
With Active Load			
With Source Degeneration			

3.3 ■ Common-Source Stage

3.3.1 Common-Source Stage with Resistive Load

By virtue of its transconductance, a MOSFET converts changes in its gate-source voltage to a small-signal drain current, which can pass through a resistor to generate an output voltage. Shown in Fig. 3.4(a), the common-source stage performs such an operation.[1] We study both the large-signal and the small-signal behavior of the circuit. Note that the input impedance of the circuit is very high at low frequencies.

If the input voltage increases from zero, M_1 is off and $V_{out} = V_{DD}$ [Fig. 3.4(b)]. As V_{in} approaches V_{TH}, M_1 begins to turn on, drawing current from R_D and lowering V_{out}. Transistor M_1 turns on in saturation regardless of the values of V_{DD} and R_D (why?), and we have

$$V_{out} = V_{DD} - R_D \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{TH})^2 \tag{3.3}$$

where channel-length modulation is neglected. With further increase in V_{in}, V_{out} drops more, and the transistor continues to operate in saturation until V_{in} exceeds V_{out} by V_{TH} [point A in Fig. 3.4(b)]. At this point,

$$V_{in1} - V_{TH} = V_{DD} - R_D \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in1} - V_{TH})^2 \tag{3.4}$$

from which $V_{in1} - V_{TH}$ and hence V_{out} can be calculated.

For $V_{in} > V_{in1}$, M_1 is in the triode region:

$$V_{out} = V_{DD} - R_D \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2(V_{in} - V_{TH})V_{out} - V_{out}^2\right] \tag{3.5}$$

[1] The common-source topology is identified as receiving the input at the gate and producing the output at the drain.

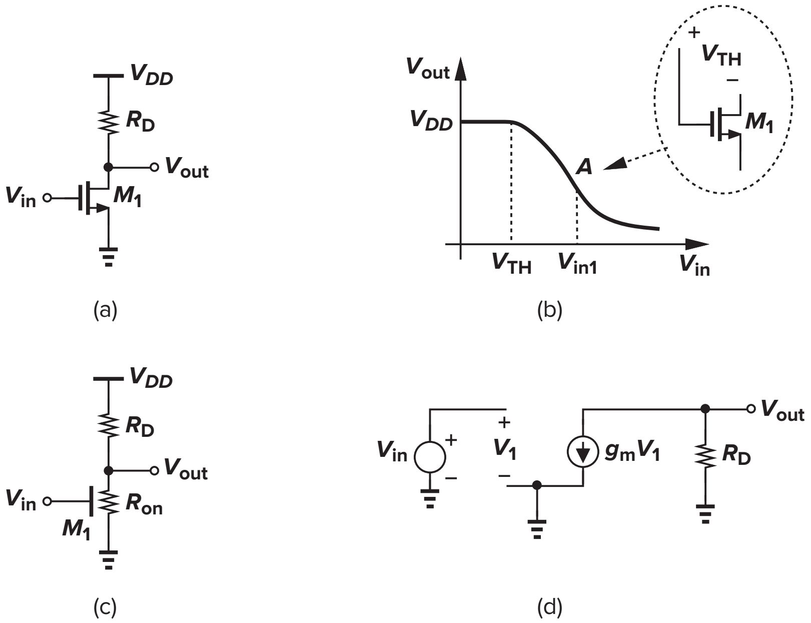


Figure 3.4 (a) Common-source stage, (b) input-output characteristic, (c) equivalent circuit in the deep triode region, and (d) small-signal model for the saturation region.

If V_{in} is high enough to drive M_1 into the deep triode region, $V_{out} \ll 2(V_{in} - V_{TH})$, and, from the equivalent circuit of Fig. 3.4(c),

$$V_{out} = V_{DD}\frac{R_{on}}{R_{on} + R_D} \tag{3.6}$$

$$= \frac{V_{DD}}{1 + \mu_n C_{ox}\frac{W}{L}R_D(V_{in} - V_{TH})} \tag{3.7}$$

Since the transconductance drops in the triode region, we usually ensure that $V_{out} > V_{in} - V_{TH}$, and hence the current operates to the left of point A in Fig. 3.4(b). Using (3.3) as the input-output characteristic and viewing its slope as the small-signal gain, we have

$$A_v = \frac{\partial V_{out}}{\partial V_{in}} \tag{3.8}$$

$$= -R_D\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{TH}) \tag{3.9}$$

$$= -g_m R_D \tag{3.10}$$

This result can be directly derived from the observation that M_1 converts an input voltage change ΔV_{in} to a drain current change $g_m\Delta V_{in}$, and hence an output voltage change $-g_m R_D \Delta V_{in}$. The small-signal model of Fig. 3.4(d) yields the same result: $V_{out} = -g_m V_1 R_D = -g_m V_{in} R_D$. Note that, as mentioned in Chapter 2, V_{in}, V_1, and V_{out} in this figure denote small-signal quantities.

Even though derived for small-signal operation, the equation $A_v = -g_m R_D$ predicts certain effects if the circuit senses a *large* signal swing. Since g_m itself varies with the input signal according to

$g_m = \mu_n C_{ox}(W/L)(V_{GS} - V_{TH})$, the gain of the circuit changes substantially if the signal is large. In other words, if the gain of the circuit *varies* significantly with the signal swing, then the circuit operates in the large-signal mode. The dependence of the gain upon the signal level leads to nonlinearity (Chapter 14), usually an undesirable effect.

A key result here is that to minimize the nonlinearity, the gain equation must be a weak function of signal-dependent parameters such as g_m. We present several examples of this concept in this chapter and in Chapter 14.

▶ Example 3.1

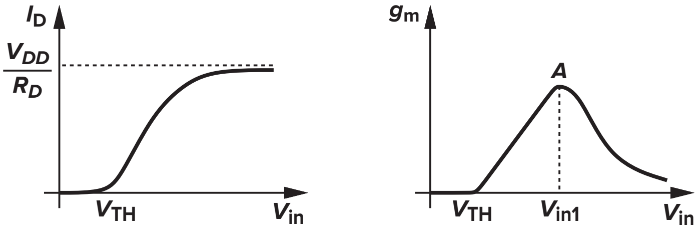
Sketch the drain current and transconductance of M_1 in Fig. 3.4(a) as a function of the input voltage.

Solution

The drain current becomes significant for $V_{in} > V_{TH}$, eventually approaching V_{DD}/R_D if $R_{on1} \ll R_D$ [Fig. 3.5(a)]. Since in saturation, $g_m = \mu_n C_{ox}(W/L)(V_{in} - V_{TH})$, the transconductance begins to rise for $V_{in} > V_{TH}$. In the triode region, $g_m = \mu_n C_{ox}(W/L)V_{DS}$, falling as V_{in} exceeds V_{in1} [Fig. 3.5(b)]. Starting with Eq. (3.5), the reader can show that

$$A_v = \frac{\partial V_{out}}{\partial V_{in}} = \frac{-\mu_n C_{ox}(W/L)R_D V_{out}}{1 + \mu_n C_{ox}(W/L)R_D(V_{in} - V_{TH} - V_{out})} \tag{3.11}$$

which reaches a maximum if $V_{out} = V_{in} - V_{TH}$ (point A).

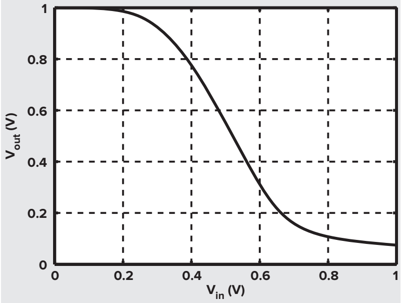


(a) (b)

Figure 3.5

Nanometer Design Notes

How does the CS stage behave in nanometer technologies? The figure plots the simulated input-output characteristic for $W/L = 2\ \mu$m/40 nm, $R_D = 2$ kΩ, and $V_{DD} = 1$ V. We observe that the circuit provides a gain of about 3 in the input range of 0.4 V to 0.6 V. The output swing is limited to about 0.3 V–0.8 V for the gain not to drop significantly.



◀

▶ Example 3.2

A CS stage is driven by a sinusoid, $V_{in} = V_1 \cos \omega_1 t + V_0$, where V_0 is the bias value and V_1 is large enough to drive the transistor into the off and triode regions. Sketch the g_m of the transistor as a function of time.

Solution

Let us first sketch the output voltage (Fig. 3.6), noting that when $V_{in} = V_1 + V_0$, V_{out} is low, M_1 is in the triode region, and g_m assumes a small value. As V_{in} falls and V_{out} and g_m rise, M_1 enters saturation at $t = t_1$ (when $V_{in} - V_{out} = V_{TH}$) and g_m reaches its maximum (why?). As V_{in} falls further, so do I_D and g_m. At $t = t_2$, g_m reaches zero.

We observe that (a) since the voltage gain is approximately equal to $-g_m R_D$, it experiences the same variation as the g_m, and (b) g_m varies periodically.[2]

[2] We even express g_m as a Fourier series in more advanced courses.

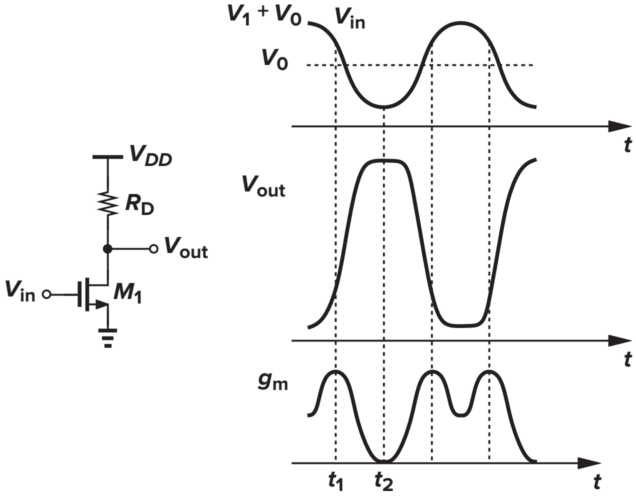


Figure 3.6

How do we maximize the voltage gain of a common-source stage? Writing (3.10) as

$$A_v = -\sqrt{2\mu_n C_{ox}\frac{W}{L}I_D}\frac{V_{RD}}{I_D} \tag{3.12}$$

where V_{RD} denotes the voltage drop across R_D, we have

$$A_v = -\sqrt{2\mu_n C_{ox}\frac{W}{L}}\frac{V_{RD}}{\sqrt{I_D}} \tag{3.13}$$

Thus, the magnitude of A_v can be increased by increasing W/L or V_{RD} or decreasing I_D if other parameters are constant. It is important to understand the trade-offs resulting from this equation. A larger device size leads to greater device capacitances, and a higher V_{RD} limits the maximum voltage swings. For example, if $V_{DD}-V_{RD}=V_{in}-V_{TH}$, then M_1 is at the edge of the triode region, allowing only very small swings at the output (and input). If V_{RD} remains constant and I_D is reduced, then R_D must increase, thereby leading to a greater time constant at the output node. In other words, as noted in the analog design octagon, the circuit exhibits trade-offs between gain, bandwidth, and voltage swings. Lower supply voltages further tighten these trade-offs.

For large values of R_D, the effect of channel-length modulation in M_1 becomes significant. Modifying (3.3) to include this effect,

$$V_{out} = V_{DD} - R_D\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in}-V_{TH})^2(1+\lambda V_{out}) \tag{3.14}$$

we have

$$\begin{aligned}\frac{\partial V_{out}}{\partial V_{in}} &= -R_D\mu_n C_{ox}\frac{W}{L}(V_{in}-V_{TH})(1+\lambda V_{out})\\ &\quad -R_D\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in}-V_{TH})^2\lambda\frac{\partial V_{out}}{\partial V_{in}}\end{aligned} \tag{3.15}$$

We recognize that $(1/2)\mu_n C_{ox}(W/L)(V_{in} - V_{TH})^2\lambda = 1/r_O$ and

$$A_v = -R_D g_m - \frac{R_D}{r_O} A_v \tag{3.16}$$

Thus,

$$A_v = -g_m \frac{r_O R_D}{r_O + R_D} \tag{3.17}$$

The small-signal model of Fig. 3.7 gives the same result with much less effort. That is, since $g_m V_1 (r_O \| R_D) = -V_{out}$ and $V_1 = V_{in}$, we have $V_{out}/V_{in} = -g_m (r_O \| R_D)$.

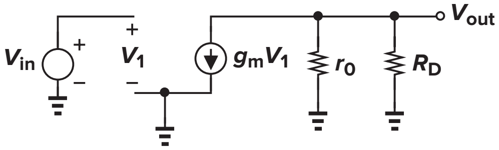


Figure 3.7 Small-signal model of CS stage including the transistor output resistance.

Example 3.3

Assuming that M_1 in Fig. 3.8 is biased in saturation, calculate the small-signal voltage gain of the circuit.

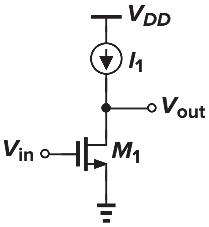


Figure 3.8

Solution

Since I_1 introduces an infinite impedance ($R_D = \infty$), the gain is limited by the output resistance of M_1:

$$A_v = -g_m r_O \tag{3.18}$$

Called the "intrinsic gain" of a transistor, this quantity represents the maximum voltage gain that can be achieved using a single device. In today's CMOS technology, $g_m r_O$ of short-channel devices is between roughly 5 and 10. We usually assume $1/g_m \ll r_O$.

In Fig. 3.8, Kirchhoff's current law (KCL) requires that $I_{D1} = I_1$. Then, how can V_{in} change the current of M_1 if I_1 is constant? Writing the total drain current of M_1 as

$$I_{D1} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{in} - V_{TH})^2 (1 + \lambda V_{out}) \tag{3.19}$$

$$= I_1 \tag{3.20}$$

we note that V_{in} appears in the square term and V_{out} in the linear term. As V_{in} increases, V_{out} must decrease such that the product remains constant. We may nevertheless say "I_{D1} increases as V_{in} increases." This statement simply refers to the quadratic part of the equation.

An important conclusion here is that, to maximize the voltage gain, we must maximize the (small-signal) load impedance. Why can we not replace the load with an open circuit? This is because the circuit still needs a path from V_{DD} to ground for the bias current of M_1.

▶ **Example 3.4**

It is possible to use the bulk (back gate) of a MOSFET as the terminal controlling the channel. Shown in Fig. 3.9 is an example. Determine the voltage gain if $\lambda = 0$.

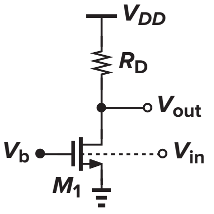


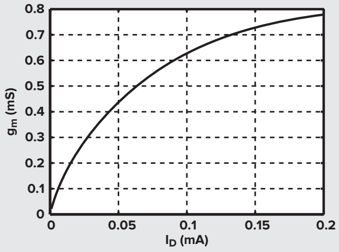
Figure 3.9

Solution

From the small-signal MOS model developed in Chapter 2, we recall that the drain current is given by $g_{mb}V_{in}$. Thus, $A_v = -g_{mb}R_D$. ◀

Nanometer Design Notes

How do we design a CS stage for a given gain and supply voltage? With W/L, I_D, and R_D under our control, we seem to have a wide design space. A good starting point is to choose a small device, $W/L = 0.5\ \mu$m/40 nm, a low bias current, $I_D = 50\ \mu$A, and a sufficiently large load resistance to achieve the required gain. To this end, we use simulations to plot the transconductance of such a device as a function of I_D, obtaining $g_m = 0.45$ mS. Thus, for a voltage gain of , say, 10, R_D must reach 22.2 kΩ if $\lambda = 0$. Is this an acceptable design? The answer depends on the application. In addition to gain, the circuit must also satisfy certain bandwidth, noise, and output swing requirements.



3.3.2 CS Stage with Diode-Connected Load

In some CMOS technologies, it is difficult to fabricate resistors with tightly-controlled values or a reasonable physical size (Chapter 19). Consequently, it is desirable to replace R_D in Fig. 3.4(a) with a MOS transistor.

A MOSFET can operate as a small-signal resistor if its gate and drain are shorted [Fig. 3.10(a)]. Called a "diode-connected" device in analogy with its bipolar counterpart, this configuration exhibits small-signal behavior similar to that of two-terminal resistor. Note that the transistor is always in saturation because the drain and the gate have the same potential. Using the small-signal equivalent shown in Fig. 3.10(b) to obtain the impedance of the device, we write $V_1 = V_X$ and $I_X = V_X/r_O + g_m V_X$. That is, the impedance of the diode is simply equal to $V_X/I_X = (1/g_m)\|r_O \approx 1/g_m$. If body effect exists, we can use the circuit in Fig. 3.11 to write $V_1 = -V_X$, $V_{bs} = -V_X$, and

$$(g_m + g_{mb})V_X + \frac{V_X}{r_O} = I_X \tag{3.21}$$

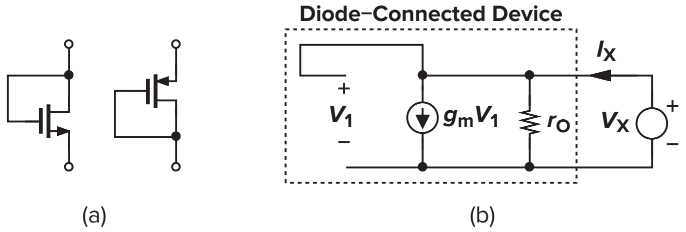


Figure 3.10 (a) Diode-connected NMOS and PMOS devices; (b) small-signal equivalent circuit.

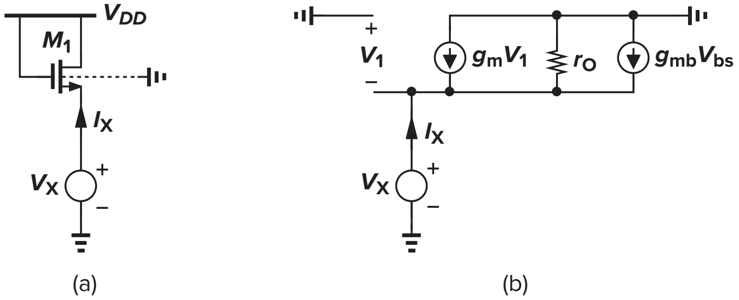


Figure 3.11 (a) Arrangement for measuring the equivalent resistance of a diode-connected MOSFET; (b) small-signal equivalent circuit.

It follows that

$$\frac{V_X}{I_X} = \frac{1}{g_m + g_{mb} + r_O^{-1}} \tag{3.22}$$

$$= \frac{1}{g_m + g_{mb}} \| r_O \tag{3.23}$$

$$\approx \frac{1}{g_m + g_{mb}} \tag{3.24}$$

In the general case, $V_X/I_X = (1/g_m)||r_O||(1/g_{mb})$. Interestingly, the impedance seen at the source of M_1 is *lower* when body effect is included. Intuitive explanation of this effect is left as an exercise for the reader.

From a large-signal point of view, a diode-connected device acts as a "square-root" operator if its current is considered the input and its V_{GS} or $V_{GS} - V_{TH}$ the output (why?). We return to this point later.

▶ Example 3.5

Consider the circuit shown in Fig. 3.12(a). In some cases, we are interested in the impedance seen looking into the source, R_X. Determine R_X if $\lambda = 0$.

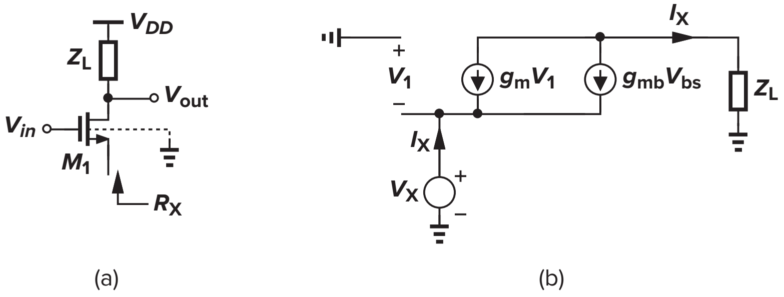


Figure 3.12 Impedance seen at the source with $\lambda = 0$.

Solution

To determine R_X, we set all independent sources to zero, draw the small-signal model, and apply a voltage source as shown in Fig. 3.12(b). Since $V_1 = -V_X$ and $V_{bs} = -V_X$, we have

$$(g_m + g_{mb})V_X = I_X \tag{3.25}$$

and

$$\frac{V_X}{I_X} = \frac{1}{g_m + g_{mb}} \tag{3.26}$$

This result should not be surprising: the topologies in Fig. 3.12(a) and Fig. 3.11(a) are similar except that the drain of M_1 in Fig. 3.12(b) is not at ac ground. This difference does not manifest itself if $\lambda = 0$. We sometimes say, "looking into the source of a MOSFET, we see $1/g_m$," assuming implicitly that $\lambda = \gamma = 0$.

We now study a common-source stage with a diode-connected load (Fig. 3.13). With negligible channel-length modulation, (3.24) can be substituted in (3.10) for the load impedance, yielding

$$A_v = -g_{m1}\frac{1}{g_{m2} + g_{mb2}} \tag{3.27}$$

$$= -\frac{g_{m1}}{g_{m2}}\frac{1}{1+\eta} \tag{3.28}$$

where $\eta = g_{mb2}/g_{m2}$. Expressing g_{m1} and g_{m2} in terms of device dimensions and bias currents, we have

$$A_v = -\frac{\sqrt{2\mu_n C_{ox}(W/L)_1 I_{D1}}}{\sqrt{2\mu_n C_{ox}(W/L)_2 I_{D2}}}\frac{1}{1+\eta} \tag{3.29}$$

and, since $I_{D1} = I_{D2}$,

$$A_v = -\sqrt{\frac{(W/L)_1}{(W/L)_2}}\frac{1}{1+\eta} \tag{3.30}$$

This equation reveals an interesting property: if the variation of η with the output voltage is neglected, the gain is independent of the bias currents and voltages (so long as M_1 stays in saturation). In other words, as the input and output signal levels vary, the gain remains relatively constant, indicating that the input-output characteristic is relatively linear.

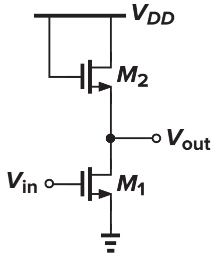


Figure 3.13 CS stage with diode-connected load.

The linear behavior of the circuit can also be confirmed by large-signal analysis. Neglecting channel-length modulation for simplicity, we have in Fig. 3.13

$$\frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_1 (V_{in} - V_{TH1})^2 = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_2 (V_{DD} - V_{out} - V_{TH2})^2 \tag{3.31}$$

and hence

$$\sqrt{\left(\frac{W}{L}\right)_1}(V_{in} - V_{TH1}) = \sqrt{\left(\frac{W}{L}\right)_2}(V_{DD} - V_{out} - V_{TH2}) \tag{3.32}$$

Thus, if the variation of V_{TH2} with V_{out} is small, the circuit exhibits a linear input-output characteristic. In essence, the squaring function performed by M_1 (from the input voltage to its drain current) and the square root function performed by M_2 (from its drain current to its overdrive) act as $f^{-1}(f(x)) = x$.

The small-signal gain can also be computed by differentiating both sides with respect to V_{in}:

$$\sqrt{\left(\frac{W}{L}\right)_1} = \sqrt{\left(\frac{W}{L}\right)_2}\left(-\frac{\partial V_{out}}{\partial V_{in}} - \frac{\partial V_{TH2}}{\partial V_{in}}\right) \tag{3.33}$$

which, upon application of the chain rule $\partial V_{TH2}/\partial V_{in} = (\partial V_{TH2}/\partial V_{out})(\partial V_{out}/\partial V_{in}) = \eta(\partial V_{out}/\partial V_{in})$, reduces to

$$\frac{\partial V_{out}}{\partial V_{in}} = -\sqrt{\frac{(W/L)_1}{(W/L)_2}}\frac{1}{1+\eta} \tag{3.34}$$

It is instructive to study the overall large-signal characteristic of the circuit as well. But let us first consider the circuit shown in Fig. 3.14(a). What is the final value of V_{out} if I_1 drops to zero? As I_1 decreases, so does the overdrive of M_2. Thus, for small I_1, $V_{GS2} \approx V_{TH2}$ and $V_{out} \approx V_{DD} - V_{TH2}$. In reality, the subthreshold conduction in M_2 eventually brings V_{out} to V_{DD} if I_D approaches zero, but at very low current levels, the finite capacitance at the output node slows down the change from $V_{DD} - V_{TH2}$ to V_{DD}. This is illustrated in the time-domain waveforms of Fig. 3.14(b). For this reason, in circuits that have frequent switching activity, we assume that V_{out} remains around $V_{DD} - V_{TH2}$ when I_1 falls to small values.

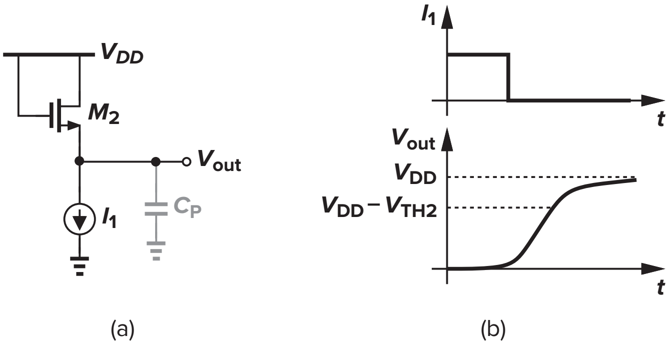


Figure 3.14 (a) Diode-connected device with stepped bias current; (b) variation of source voltage versus time.

Now we return to the circuit of Fig. 3.13. Plotted in Fig. 3.15 versus V_{in}, the output voltage equals $V_{DD} - V_{TH2}$ if $V_{in} < V_{TH1}$. For $V_{in} > V_{TH1}$, Eq. (3.32) holds and V_{out} follows an approximately straight line. As V_{in} exceeds $V_{out} + V_{TH1}$ (beyond point A), M_1 enters the triode region, and the characteristic becomes nonlinear.

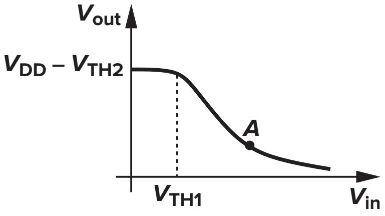


Figure 3.15 Input-output characteristic of a CS stage with diode-connected load.

The diode-connected load of Fig. 3.13 can be implemented with a PMOS device as well. Shown in Fig. 3.16, the circuit is free from body effect, providing a small-signal voltage gain equal to

$$A_v = -\sqrt{\frac{\mu_n (W/L)_1}{\mu_p (W/L)_2}} \tag{3.35}$$

where channel-length modulation is neglected.

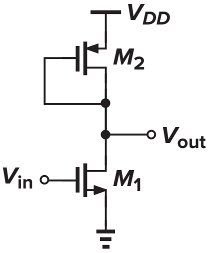


Figure 3.16 CS stage with diode-connected PMOS device.

Equations (3.30) and (3.35) indicate that the gain of a common-source stage with diode-connected load is a relatively weak function of the device dimensions. For example, to achieve a gain of 5, $\mu_n (W/L)_1/[\mu_p (W/L)_2] = 25$, implying that, with $\mu_n \approx 2\mu_p$, we must have $(W/L)_1 \approx 12.5(W/L)_2$. In a sense, a high gain requires a "strong" input device and a "weak" load device. In addition to disproportionately wide or long transistors (and hence a large input or load capacitance), a high gain translates to another important limitation: reduction in allowable voltage swings. Specifically, since in Fig. 3.16, $I_{D1} = |I_{D2}|$,

$$\mu_n \left(\frac{W}{L}\right)_1 (V_{GS1} - V_{TH1})^2 = \mu_p \left(\frac{W}{L}\right)_2 (V_{GS2} - V_{TH2})^2 \tag{3.36}$$

if $\lambda = 0$, revealing that

$$\frac{|V_{GS2} - V_{TH2}|}{V_{GS1} - V_{TH1}} = A_v \tag{3.37}$$

In the above example, the overdrive voltage of M_2 must be 5 times that of M_1. For example, with $V_{GS1} - V_{TH1} = 100$ mV and $|V_{TH2}| = 0.3$ V, we have $|V_{GS2}| = 0.8$ V, severely limiting the output swing. This is another example of the trade-offs suggested by the analog design octagon. Note that, with diode-connected loads, the swing is constrained by both the required overdrive voltage and the threshold voltage. That is, even with a small overdrive, the output level cannot exceed $V_{DD} - |V_{TH}|$.

An interesting paradox arises here if we write $g_m = \mu C_{ox}(W/L)|V_{GS} - V_{TH}|$. The voltage gain of the circuit is then given by

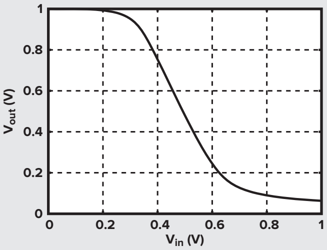
$$|A_v| = \frac{g_{m1}}{g_{m2}} \tag{3.38}$$

$$= \frac{\mu_n C_{ox} (W/L)_1 (V_{GS1} - V_{TH1})}{\mu_p C_{ox} (W/L)_2 |V_{GS2} - V_{TH2}|} \tag{3.39}$$

Equation (3.39) implies that A_v is *inversely* proportional to $|V_{GS2} - V_{TH2}|$. It is left for the reader to resolve the seemingly opposite trends suggested by (3.37) and (3.39).

Nanometer Design Notes

Let us design and simulate a CS stage with a diode-connected PMOS load in 40-nm technology. We select $W/L = 5\ \mu$m/40 nm for the NMOS device and $1\ \mu$m/40 nm for the PMOS device. The large-signal I-V characteristic is shown in the figure. We observe that the circuit provides a small-signal gain of about 1.5 for an input range of roughly 0.4 V to 0.5 V. (In this dc sweep, V_{out} can reach V_{DD} when the input transistor is off.)



▶ Example 3.6

In the circuit of Fig. 3.17, M_1 is biased in saturation with a drain current equal to I_1. The current source $I_S = 0.75I_1$ is added to the circuit. How is (3.37) modified for this case? Assume $\lambda = 0$.

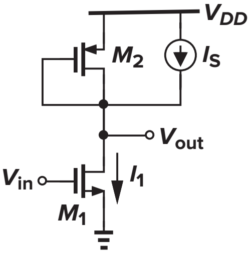


Figure 3.17

Solution

Since $|I_{D2}| = I_1/4$, we have

$$A_v = -\frac{g_{m1}}{g_{m2}} \tag{3.40}$$

$$= -\sqrt{\frac{4\mu_n(W/L)_1}{\mu_p(W/L)_2}} \tag{3.41}$$

Moreover,

$$\mu_n\left(\frac{W}{L}\right)_1 (V_{GS1} - V_{TH1})^2 = 4\mu_p\left(\frac{W}{L}\right)_2 (V_{GS2} - V_{TH2})^2 \tag{3.42}$$

yielding

$$\frac{|V_{GS2} - V_{TH2}|}{V_{GS1} - V_{TH1}} = \frac{A_v}{4} \tag{3.43}$$

Thus, for a gain of 5, the overdrive of M_2 need be only 1.25 times that of M_1. Alternatively, for a given overdrive voltage, this circuit achieves a gain four times that of the stage in Fig. 3.16. Intuitively, this is because for a given $|V_{GS2} - V_{TH2}|$, if the current decreases by a factor of 4, then $(W/L)_2$ must decrease proportionally, and $g_{m2} = \sqrt{2\mu_p C_{ox}(W/L)_2 I_{D2}}$ is lowered by the same factor. ◀

▶ Example 3.7

A student attempts to calculate the voltage gain in the previous example by differentiating both sides of (3.42). Does this approach give a correct result? Why?

Solution

Since $V_{GS2} = V_{out} - V_{DD}$, differentiation and multiplication by C_{ox} yield

$$\mu_n C_{ox}\left(\frac{W}{L}\right)_1 (V_{in} - V_{TH1})^2 = 4\mu_p C_{ox}\left(\frac{W}{L}\right)_2 (V_{out} - V_{DD} - V_{TH2})\frac{\partial V_{out}}{\partial V_{in}} \tag{3.44}$$

It follows that $\partial V_{out}/\partial V_{in} = -g_{m1}/(4g_{m2})$. This incorrect result arises because (3.42) is valid for only *one* value of V_{in}. As V_{in} is perturbed by the signal, I_1 departs from $4|I_{D2}|$ and (3.42) cannot be differentiated. ◀

In today's CMOS technology, channel-length modulation is quite significant and, more important, the behavior of transistors notably departs from the square law. Thus, the gain of the stage in Fig. 3.13 must be expressed as

$$A_v = -g_{m1}\left(\frac{1}{g_{m2}}\|r_{O1}\|r_{O2}\right) \tag{3.45}$$

where g_{m1} and g_{m2} must be obtained as described in Chapter 17.

3.3.3 CS Stage with Current-Source Load

In applications requiring a large voltage gain in a single stage, the relationship $A_v = -g_m R_D$ suggests that we should increase the load impedance of the CS stage. With a resistor or diode-connected load, however, increasing the load resistance translates to a large dc drop across the load, thereby limiting the output voltage swing.

A more practical approach is to replace the load with a device that does not obey Ohm's law, e.g., a current source. Described briefly in Example 3.3, the resulting circuit is shown in Fig. 3.18, where both transistors operate in saturation. Since the total impedance seen at the output node is equal to $r_{O1}\|r_{O2}$, the gain is given by is

$$A_v = -g_{m1}(r_{O1}\|r_{O2}) \tag{3.46}$$

The key point here is that the output impedance and the minimum required $|V_{DS}|$ of M_2 are less strongly coupled than the value and voltage drop of a resistor; the former need not satisfy Ohm's law, but the latter must. The voltage $|V_{DS2,min}| = |V_{GS2} - V_{TH2}|$ can be reduced to less than a hundred millivolts by simply increasing the width of M_2. If r_{O2} is not sufficiently high, the length and width of M_2 can be increased to achieve a smaller λ while maintaining the same overdrive voltage. The penalty is the larger capacitance introduced by M_2 at the output node.

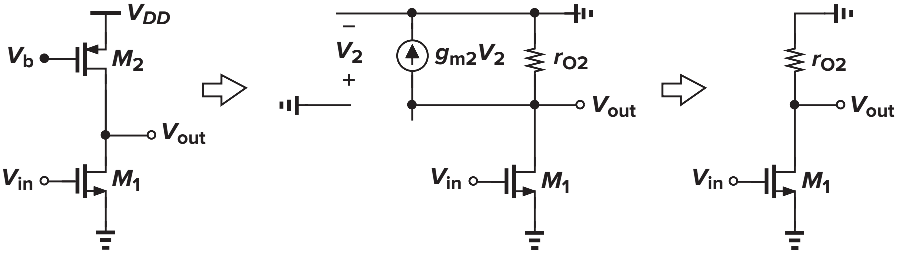


Figure 3.18 CS stage with current-source load.

We should remark that the output bias voltage of the circuit in Fig. 3.18 is not well-defined. Thus, the stage is reliably biased only if a feedback loop forces V_{out} to a known value (Chapter 8). The large-signal analysis of the circuit is left as an exercise for the reader.

As explained in Chapter 2, the output impedance of MOSFETs at a given drain current can be scaled by changing the channel length, i.e., to the first order, $\lambda \propto 1/L$, and hence $r_O \propto L/I_D$. Since the gain of the stage shown in Fig. 3.18 is proportional to $r_{O1}\|r_{O2}$, we may surmise that longer transistors yield a higher voltage gain.

Let us consider M_1 and M_2 separately. If L_1 is scaled up by a factor of α (> 1), then W_1 may need to be scaled proportionally as well. This is because, for a given drain current, $V_{GS1} - V_{TH1} \propto 1/\sqrt{(W/L)_1}$, i.e., if W_1 is not scaled, the overdrive voltage increases, limiting the output voltage swing. Also, since $g_{m1} \propto \sqrt{(W/L)_1}$, scaling up only L_1 lowers g_{m1}.

In applications where these issues are unimportant, W_1 can remain constant while L_1 increases. Thus, the intrinsic gain of M_1 can be written as

$$g_{m1}r_{O1} = \sqrt{2\left(\frac{W}{L}\right)_1 \mu_n C_{ox} I_D}\frac{1}{\lambda I_D} \tag{3.47}$$

indicating that the gain *increases* with L because λ depends more strongly on L than g_m does. Also, note that $g_m r_O$ *decreases* as I_D increases.

Increasing L_2 while keeping W_2 constant increases r_{O2} and hence the voltage gain, but at the cost of a higher $|V_{DS2,min}|$, which is required to maintain M_2 in saturation.

▶ **Example 3.8**

Compare the maximum output voltage swings of CS stages with resistive and current-source loads.

Solution

For the resistively-loaded stage [Fig. 3.19(a)], the maximum output voltage is near V_{DD} (when V_{in} falls to about V_{TH1}). The minimum is the value that places M_1 at the edge of the triode region, $V_{in} - V_{TH1}$.

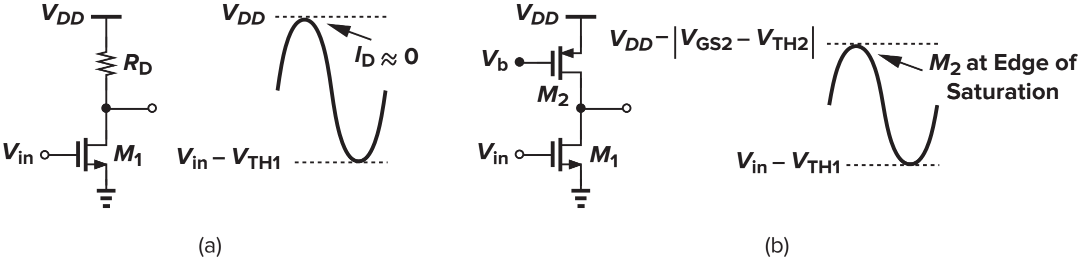


Figure 3.19 Output swing in CS stage with (a) resistive load and (b) current-source load.

For the stage with a current-source load [Fig. 3.19(b)], the maximum output voltage is that which places M_2 at the edge of the triode region, $V_{DD} - |V_{GS2} - V_{TH2}|$. Thus, the latter actually provides *smaller* swings than the former, but can always achieve a *higher* gain if L_1 and L_2 are increased.

◀

3.3.4 CS Stage with Active Load

In the amplifier topology of Fig. 3.19(b), the PMOS device serves as a constant current source. Is it possible for M_2 to operate as an *amplifying* device? Yes; we can apply the input signal to the gate of M_2 as well [Fig. 3.20(a)], converting it to an "active" load. The reader may recognize this topology as a CMOS inverter. Suppose both transistors are in saturation and V_{in} rises by ΔV_0. Two changes now occur: (a) I_{D1} increases, pulling V_{out} lower, and (b) M_2 injects less current into the output node, allowing V_{out} to drop. The two changes thus *enhance* each other, leading to a greater voltage gain. Equivalently, as seen in Fig. 3.20(b), the two transistors operate in parallel and collapse into one as illustrated in Fig. 3.20(c). It follows that $-(g_{m1} + g_{m2})V_{in}(r_{O1}||r_{O2}) = V_{out}$, and hence

$$A_v = -(g_{m1} + g_{m2})(r_{O1}||r_{O2}) \tag{3.48}$$

Compared to the amplifier of Fig. 3.19(b), this circuit exhibits the same output resistance, $r_{O1}||r_{O2}$, but a higher transconductance. This topology is also called a "complementary CS stage."

The amplifier of Fig. 3.20(a) must deal with two critical issues. First, the bias current of the two transistors is a strong function of PVT. In particular, since $V_{GS1} + |V_{GS2}| = V_{DD}$, variations in V_{DD} or the threshold voltages directly translate to changes in the drain currents. Second, the circuit *amplifies*

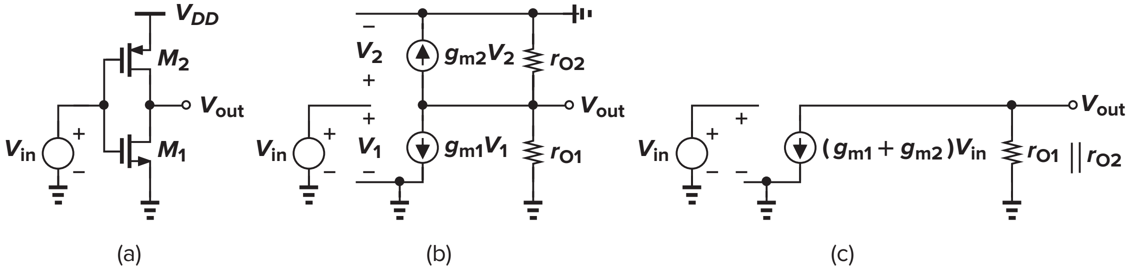


Figure 3.20 (a) CS stage with active load, (b) small-signal model, and (c) simplified model.

supply voltage variations ("supply noise")! To understand this point, consider the arrangement depicted in Fig. 3.21, where V_B is a bias voltage to place M_1 and M_2 in saturation. In Problem 3.31, we prove that the small-signal gain from V_{DD} to V_{out} is given by

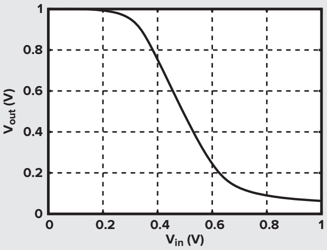
$$\frac{V_{out}}{V_{DD}} = \frac{g_{m2}r_{O2}+1}{r_{O2}+r_{O1}}r_{O1} \tag{3.49}$$

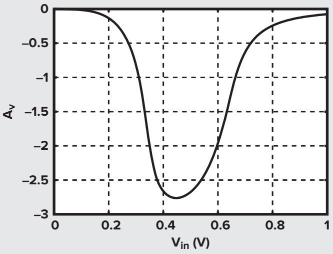
$$= \left(g_{m2}+\frac{1}{r_{O2}}\right)(r_{O1}||r_{O2}) \tag{3.50}$$

about half of the A_v found above. These issues are addressed in Chapter 5.

Nanometer Design Notes

With minimum channel lengths, the CS stage with current-source load provides a low gain. For example, if $(W/L)_{NMOS} = 5$ μm/40 nm and $(W/L)_{PMOS} = 10$ μm/40 nm, we obtain the input-output characteristic shown in the figure, where the maximum gain is about 2.5! If we plot the slope, we also see the useful output voltage range to be about 0.7 V with $V_{DD} = 1$ V. Outside this range, the gain drops considerably.





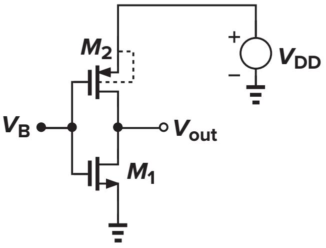


Figure 3.21 Arrangement for studying supply sensitivity of CS stage with active load.

3.3.5 CS Stage with Triode Load

A MOS device operating in the deep triode region behaves as a resistor and can therefore serve as the load in a CS stage. Illustrated in Fig. 3.22, such a circuit biases the gate of M_2 at a sufficiently low level, ensuring that the load is in the deep triode region for all output voltage swings. Since

$$R_{on2} = \frac{1}{\mu_p C_{ox}(W/L)_2(V_{DD}-V_b-|V_{THP}|)} \tag{3.51}$$

the voltage gain can be readily calculated.

The principal drawback of this circuit stems from the dependence of R_{on2} upon $\mu_p C_{ox}$, V_b, and V_{THP}. Since $\mu_p C_{ox}$ and V_{THP} vary with process and temperature, and since generating a precise value for V_b requires additional complexity, this circuit is difficult to use. Triode loads, however, consume less voltage headroom than do diode-connected devices because in Fig. 3.22, $V_{out,max} = V_{DD}$, whereas in Fig. 3.16, $V_{out,max} \approx V_{DD} - |V_{THP}|$.

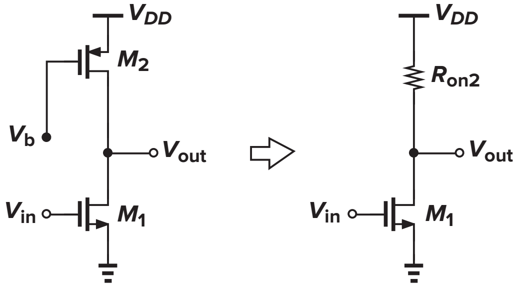


Figure 3.22 CS stage with triode load.

Among the five CS variants studied above, those employing resistive, current-source, or active loads find wider usage than the other two.

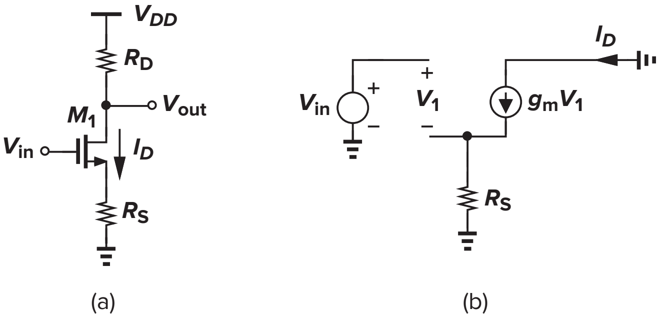


Figure 3.23 CS stage with source degeneration.

3.3.6 CS Stage with Source Degeneration

In some applications, the nonlinear dependence of the drain current upon the overdrive voltage introduces excessive nonlinearity, making it desirable to "soften" the device characteristics. In Sec. 3.3.2, we noted the linear behavior of a CS stage using a diode-connected load, which allows "postcorrection" of the nonlinearity. Alternatively, as depicted in Fig. 3.23(a), this can be accomplished by placing a "degeneration" resistor in series with the source terminal so as to make the input device more linear. Let us neglect channel-length modulation and body effect. Here, as V_{in} increases, so do I_D and the voltage drop across R_S. That is, a fraction of the change in V_{in} appears across the resistor rather than as the gate-source overdrive, thus leading to a smoother variation of I_D. From another perspective, we intend to make the gain equation a weaker function of g_m. Since $V_{out} = V_{DD} - I_D R_D$, the nonlinearity of the circuit arises from the nonlinear dependence of I_D upon V_{in}. We note that $\partial V_{out}/\partial V_{in} = -(\partial I_D/\partial V_{in})R_D$, and define the equivalent transconductance of the circuit as $G_m = \partial I_D/\partial V_{in}$.[3] Now, assuming that $I_D = f(V_{GS})$, we write

$$G_m = \frac{\partial I_D}{\partial V_{in}} \tag{3.52}$$

$$= \frac{\partial f}{\partial V_{GS}} \frac{\partial V_{GS}}{\partial V_{in}} \tag{3.53}$$

[3] As explained later, the output voltage must be kept constant when G_m is calculated.

Since $V_{GS} = V_{in} - I_D R_S$, we have $\partial V_{GS}/\partial V_{in} = 1 - R_S \partial I_D/\partial V_{in}$, obtaining

$$G_m = \left(1 - R_S \frac{\partial I_D}{\partial V_{in}}\right) \frac{\partial f}{\partial V_{GS}} \tag{3.54}$$

But, $\partial f/\partial V_{GS}$ is the transconductance of M_1, and

$$G_m = \frac{g_m}{1 + g_m R_S} \tag{3.55}$$

The small-signal voltage gain is thus equal to

$$A_v = -G_m R_D \tag{3.56}$$

$$= \frac{-g_m R_D}{1 + g_m R_S} \tag{3.57}$$

The same result can be derived using the small-signal model of Fig. 3.23(b) by writing a KVL, $V_{in} = V_1 + I_D R_S$, and noting that $I_D = g_m V_1$. Equation (3.55) implies that as R_S increases, G_m becomes a weaker function of g_m and hence the drain current. In fact, for $R_S \gg 1/g_m$, we have $G_m \approx 1/R_S$, i.e., $\Delta I_D \approx \Delta V_{in}/R_S$, concluding that most of the change in V_{in} appears across R_S. We say that the drain current is a "linearized" function of the input voltage. In Problem 3.30, we examine this effect from a different perspective. The linearization is obtained at the cost of lower gain [and higher noise (Chapter 7)].

For our subsequent calculations, it is useful to determine G_m in the presence of body effect and channel-length modulation. With the aid of the equivalent circuit shown in Fig. 3.24, we recognize that the current through R_S equals I_{out} and, therefore, $V_{in} = V_1 + I_{out} R_S$. Summing the currents at node X, we have

$$I_{out} = g_m V_1 - g_{mb} V_X - \frac{I_{out} R_S}{r_O} \tag{3.58}$$

$$= g_m (V_{in} - I_{out} R_S) + g_{mb}(-I_{out} R_S) - \frac{I_{out} R_S}{r_O} \tag{3.59}$$

It follows that

$$G_m = \frac{I_{out}}{V_{in}} \tag{3.60}$$

$$= \frac{g_m r_O}{R_S + [1 + (g_m + g_{mb}) R_S] r_O} \tag{3.61}$$

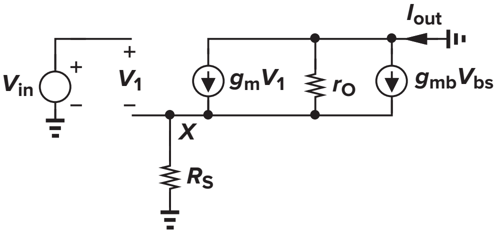
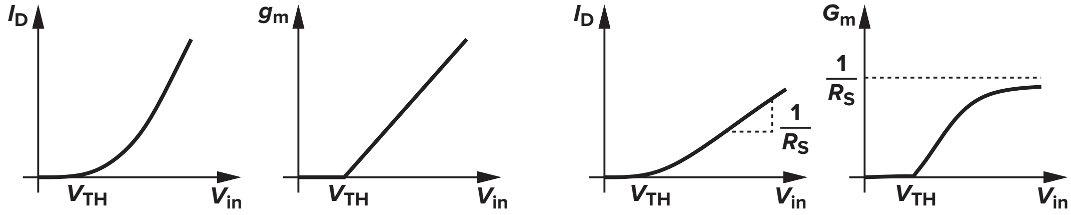


Figure 3.24 Small-signal equivalent circuit of a degenerated CS stage.

Let us now examine the large-signal behavior of the CS stage with $R_S = 0$ and $R_S \neq 0$. For $R_S = 0$, our derivations in Chapter 2 indicate that I_D and g_m vary as shown in Fig. 3.25(a). For $R_S \neq 0$, the turn-on behavior is similar to that in Fig. 3.25(a) because, at low current levels, $1/g_m \gg R_S$, and hence $G_m \approx g_m$ [Fig. 3.25(b)]. As the overdrive and therefore g_m increase, the effect of degeneration, $1 + g_m R_S$ in (3.55), becomes more significant. For large values of V_{in} (if M_1 is still saturated), I_D is approximately a linear function of V_{in} and G_m approaches $1/R_S$.



(a) (b)

Figure 3.25 Drain current and transconductance of a CS device (a) without and (b) with source degeneration.

Example 3.9

Plot the small-signal voltage gain of the circuit in Fig. 3.23 as a function of the input bias voltage, V_{in}.

Solution

Using the results derived above for the equivalent transconductance of M_1 and R_S, we arrive at the plot shown in Fig. 3.26. For V_{in} slightly greater than V_{TH}, $1/g_m \gg R_S$ and $A_v \approx -g_m R_D$. As V_{in} increases, degeneration becomes more significant and $A_v = -g_m R_D/(1+g_m R_S)$. For large values of V_{in}, $G_m \approx 1/R_S$ and $A_v = -R_D/R_S$. However, if $V_{in} > V_{out} + V_{TH}$, that is, if $R_D I_D > V_{TH} + V_{DD} - V_{in}$, M_1 enters the triode region and A_v drops.

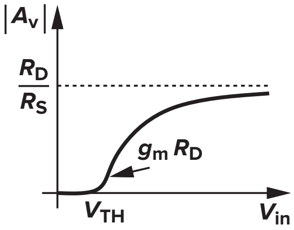


Figure 3.26

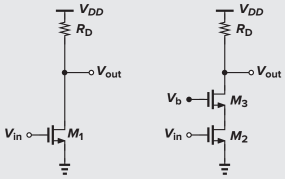
Equation (3.57) can be rewritten as

$$A_v = -\frac{R_D}{\frac{1}{g_m} + R_S} \tag{3.62}$$

This result allows formulating the gain by inspection. First, let us examine the denominator of (3.62). The expression is equal to the *series* combination of the inverse transconductance of the device and the explicit resistance seen from the source to ground. We call the denominator "the resistance seen in the source path" because if, as shown in Fig. 3.27, we disconnect the bottom terminal of R_S from ground and calculate the resistance seen "looking up" (while setting the input to zero), we obtain $R_S + 1/g_m$.

Nanometer Design Notes

A common issue in nanometer technologies is that a MOS transistor experiences ”stress” if its V_{GS}, V_{DS}, or V_{DG} exceeds certain limits. For example, in 40-nm technology, these voltages should remain below 1 V. Interestingly, the cascode structure can avoid device stress even if V_{DD} is greater than allowed. As can be seen from the diagram below, as the drain current decreases and V_{out} approaches V_{DD}, M_1 experiences $V_{DS} = V_{DD}$, whereas M_2 sees $V_{DS} \approx V_b - V_{TH2}$. Similarly, $V_{DS3} < V_{DD}$ (why?).



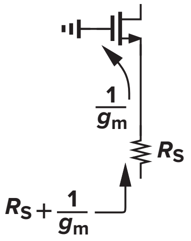


Figure 3.27 Resistance seen in the source path.

Noting that the numerator of (3.62) is the resistance seen at the drain, we view the magnitude of the gain as the resistance seen at the drain node divided by the total resistance in the source path. This method greatly simplifies the analysis of more complex circuits.

▶ **Example 3.10**

Assuming $\lambda = \gamma = 0$, calculate the small-signal gain of the circuit shown in Fig. 3.28(a).

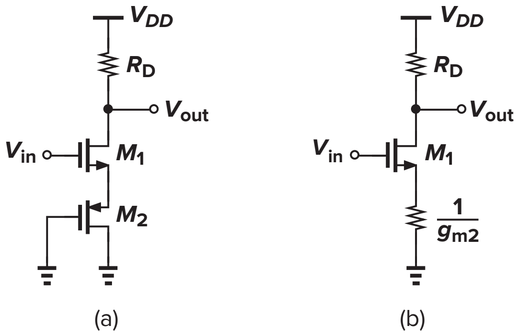


Figure 3.28

Solution

Noting that M_2 is a diode-connected device and simplifying the circuit to that shown in Fig. 3.28(b), we use the above rule to write

$$A_v = -\frac{R_D}{\dfrac{1}{g_{m1}} + \dfrac{1}{g_{m2}}} \tag{3.63}$$

◀

Output Resistance Another important consequence of source degeneration is the increase in the output resistance of the stage. We calculate the output resistance first with the aid of the equivalent circuit shown in Fig. 3.29, where the load resistor, R_D, is excluded for now. Note that body effect is also included to

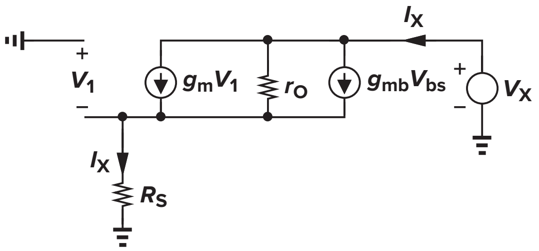


Figure 3.29 Equivalent circuit for calculating the output resistance of a degenerated CS stage.

arrive at a general result. Since the current through R_S is equal to I_X, $V_1 = -I_X R_S$, and the current flowing through r_O is given by $I_X - (g_m + g_{mb})V_1 = I_X + (g_m + g_{mb})R_S I_X$. Adding the voltage drops across r_O and R_S, we obtain

$$r_O[I_X + (g_m + g_{mb})R_S I_X] + I_X R_S = V_X \tag{3.64}$$

It follows that

$$\begin{aligned} R_{out} &= [1 + (g_m + g_{mb})R_S]r_O + R_S && (3.65)\\ &= [1 + (g_m + g_{mb})r_O]R_S + r_O && (3.66)\end{aligned}$$

Equation (3.65) indicates that r_O is "boosted" by a factor of $1 + (g_m + g_{mb})R_S$ and then added to R_S. As an alternative perspective, Eq. (3.66) suggests that R_S is boosted by a factor of $1 + (g_m + g_{mb})r_O$ (a value close to the transistor's intrinsic gain) and then added to r_O. Both views prove useful in analyzing circuits. Note that the overall output resistance is equal to the parallel combination of R_{out} and R_D. If $(g_m + g_{mb})r_O \gg 1$, we have

$$\begin{aligned} R_{out} &\approx (g_m + g_{mb})r_O R_S + r_O && (3.67)\\ &= [1 + (g_m + g_{mb})R_S]r_O && (3.68)\end{aligned}$$

To gain more insight, let us consider the circuit of Fig. 3.29 with $R_S = 0$ and $R_S > 0$. If $R_S = 0$, then $g_m V_1 = g_{mb} V_{bs} = 0$ and $I_X = V_X/r_O$. On the other hand, if $R_S > 0$, we have $I_X R_S > 0$ and $V_1 < 0$, obtaining *negative* $g_m V_1$ and $g_{mb} V_{bs}$. Thus, the current supplied by V_X is *less* than V_X/r_O, and hence the output impedance is greater than r_O.

The relationship in (3.65) can also be derived by inspection. As shown in Fig. 3.30(a), we apply a voltage to the output node, change its value by ΔV, and measure the resulting change, ΔI, in the output current. Since the current through R_S must change by ΔI (why?), we first compute the voltage change across R_S. To this end, we draw the circuit as shown in Fig. 3.30(b) and note that the resistance seen looking into the source of M_1 is equal to $1/(g_m + g_{mb})$ [Eq. (3.24)], thus arriving at the equivalent circuit in Fig. 3.30(c). The voltage change across R_S is therefore equal to

$$\Delta V_{RS} = \Delta V \frac{\dfrac{1}{g_m + g_{mb}} \| R_S}{\dfrac{1}{g_m + g_{mb}} \| R_S + r_O} \tag{3.69}$$

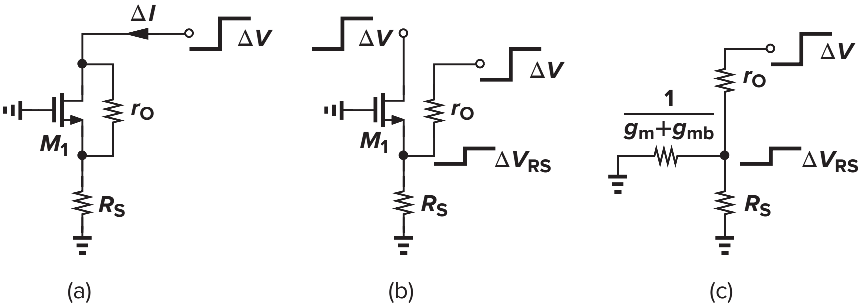


Figure 3.30 (a) Change in drain current in response to change in applied voltage to drain; (b) equivalent of (a); (c) small-signal model.

The change in the current is

$$\Delta I = \frac{\Delta V_{RS}}{R_S} \tag{3.70}$$

$$= \Delta V \frac{1}{[1 + (g_m + g_{mb})]R_S r_O + R_S} \tag{3.71}$$

that is,

$$\frac{\Delta V}{\Delta I} = [1 + (g_m + g_{mb})R_S]r_O + R_S \tag{3.72}$$

With the foregoing developments, we can now compute the gain of a degenerated CS stage in the general case, taking into account both body effect and channel-length modulation. In the equivalent circuit depicted in Fig. 3.31, the current through R_S must equal that through R_D, i.e., $-V_{out}/R_D$. Thus, the source voltage with respect to ground (and the bulk) is equal to $-V_{out}R_S/R_D$, yielding $V_1 = V_{in} + V_{out}R_S/R_D$. The current flowing through r_O from top to bottom can therefore be written as

$$I_{ro} = -\frac{V_{out}}{R_D} - (g_m V_1 + g_{mb} V_{bs}) \tag{3.73}$$

$$= -\frac{V_{out}}{R_D} - \left[g_m\left(V_{in} + V_{out}\frac{R_S}{R_D}\right) + g_{mb}V_{out}\frac{R_S}{R_D}\right] \tag{3.74}$$

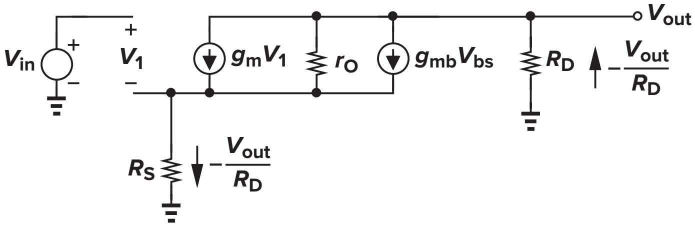


Figure 3.31 Small-signal model of degenerated CS stage with finite output resistance.

Since the voltage drops across r_O and R_S must add up to V_{out}, we have

$$V_{out} = I_{ro}r_O - \frac{V_{out}}{R_D}R_S \tag{3.75}$$

$$= -\frac{V_{out}}{R_D}r_O - \left[g_m\left(V_{in} + V_{out}\frac{R_S}{R_D}\right) + g_{mb}V_{out}\frac{R_S}{R_D}\right]r_O - V_{out}\frac{R_S}{R_D} \tag{3.76}$$

It follows that

$$\frac{V_{out}}{V_{in}} = \frac{-g_m r_O R_D}{R_D + R_S + r_O + (g_m + g_{mb})R_S r_O} \tag{3.77}$$

To gain more insight into this result, we recognize that the last three terms in the denominator, namely, $R_S + r_O + (g_m + g_{mb})R_S r_O$, represent the output resistance of a MOS device degenerated by a resistor R_S, as originally derived in (3.66). Let us now rewrite (3.77) as

$$A_v = \frac{-g_m r_O R_D[R_S + r_O + (g_m + g_{mb})R_S r_O]}{R_D + R_S + r_O + (g_m + g_{mb})R_S r_O} \cdot \frac{1}{R_S + r_O + (g_m + g_{mb})R_S r_O} \tag{3.78}$$

$$= -\frac{g_m r_O}{R_S + r_O + (g_m + g_{mb})R_S r_O} \cdot \frac{R_D[R_S + r_O + (g_m + g_{mb})R_S r_O]}{R_D + R_S + r_O + (g_m + g_{mb})R_S r_O} \tag{3.79}$$

The two fractions in (3.79) represent two important parameters of the circuit: the first is identical to that in (3.61), i.e., the equivalent transconductance of a degenerated MOSFET; and the second denotes the parallel combination of R_D and $R_S + r_O + (g_m + g_{mb})R_S r_O$, i.e., the overall output resistance of the circuit.

The above discussion suggests that in some circuits, it may be easier to calculate the voltage gain by exploiting the following lemma. We recall that the output port of a linear circuit can be represented by a Norton equivalent [Fig. 3.32(a)].

Lemma In a linear circuit, the voltage gain is equal to $-G_m R_{out}$, where G_m denotes the transconductance of the circuit when the output is shorted to ground [Fig. 3.32(b)] and R_{out} represents the output resistance of the circuit when the input voltage is set to zero [Fig. 3.32(c)].

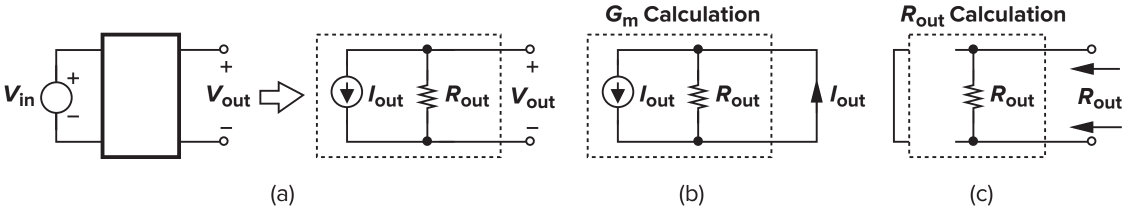


Figure 3.32 (a) Norton equivalent of a linear circuit; (b) G_m calculation; and (c) R_{out} calculation.

The lemma can be proved by noting that the output voltage in Fig. 3.32(a) is equal to $-I_{out}R_{out}$, and I_{out} can be obtained by measuring the short-circuit current at the output. Defining $G_m = I_{out}/V_{in}$, we have $V_{out} = -G_m V_{in} R_{out}$. This lemma proves useful if G_m and R_{out} can be determined by inspection. Note the direction of I_{out}.

▶ **Example 3.11**

Calculate the voltage gain of the circuit shown in Fig. 3.33. Assume that I_0 is ideal.

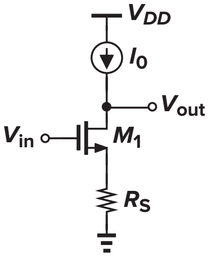


Figure 3.33

Solution

The transconductance and output resistance of the stage are given by Eqs. (3.61) and (3.66), respectively. Thus,

$$A_v = -\frac{g_m r_O}{R_S + [1 + (g_m + g_{mb})R_S]r_O}\{[1 + (g_m + g_{mb})r_O]R_S + r_O\} \tag{3.80}$$

$$= -g_m r_O \tag{3.81}$$

Interestingly, the voltage gain is equal to the intrinsic gain of the transistor and independent of R_S. This is because, if I_0 is ideal, the current through R_S cannot change, and hence the small-signal voltage drop across R_S is zero—as if R_S were zero itself. ◀

3.4 ■ Source Follower

Our analysis of the common-source stage indicates that, to achieve a high voltage gain with limited supply voltage, the load impedance must be as large as possible. If such a stage is to drive a low-impedance load, then a "buffer" must be placed after the amplifier so as to drive the load with negligible reduction in gain. The source follower (also called the "common-drain" stage) can operate as a voltage buffer.

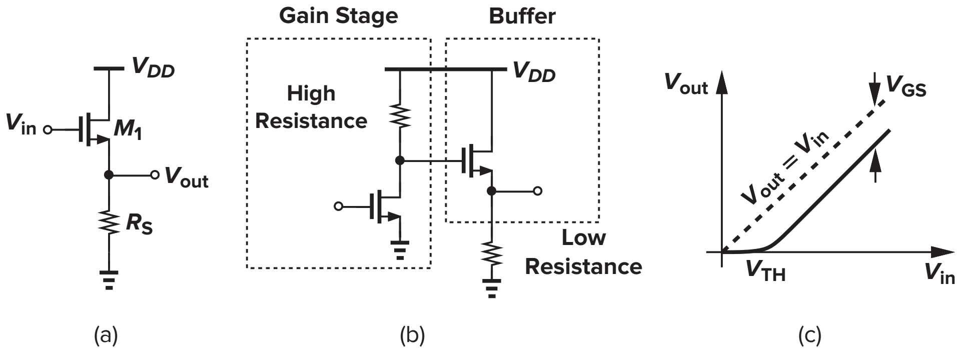


Figure 3.34 (a) Source follower, (b) example of its role as a buffer, and (c) its input-output characteristic.

Illustrated in Fig. 3.34(a), the source follower senses the signal at the gate, while presenting a high input impedance, and drives the load at the source, allowing the source potential to "follow" the gate voltage. Figure 3.34(b) depicts how the circuit can be used to drive a low resistance without degrading the voltage gain of a CS stage. Beginning with the large-signal behavior of the source follower, we note that for $V_{in} < V_{TH}$, M_1 is off and $V_{out} = 0$. As V_{in} exceeds V_{TH}, M_1 turns on in saturation (why?) and I_{D1} flows through R_S [Fig. 3.34(c)]. As V_{in} increases further, V_{out} follows the input with a difference (level shift) equal to V_{GS}. We can express the input-output characteristic as

$$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{TH} - V_{out})^2 R_S = V_{out} \tag{3.82}$$

where channel-length modulation is neglected. Let us calculate the small-signal gain of the circuit by differentiating both sides of (3.82) with respect to V_{in}:

$$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}2(V_{in} - V_{TH} - V_{out})\left(1 - \frac{\partial V_{TH}}{\partial V_{in}} - \frac{\partial V_{out}}{\partial V_{in}}\right) R_S = \frac{\partial V_{out}}{\partial V_{in}} \tag{3.83}$$

Since $\partial V_{TH}/\partial V_{in} = (\partial V_{TH}/\partial V_{SB})(\partial V_{SB}/\partial V_{in}) = \eta \partial V_{out}/\partial V_{in}$,

$$\frac{\partial V_{out}}{\partial V_{in}} = \frac{\mu_n C_{ox}\dfrac{W}{L}(V_{in} - V_{TH} - V_{out})R_S}{1 + \mu_n C_{ox}\dfrac{W}{L}(V_{in} - V_{TH} - V_{out})R_S(1+\eta)} \tag{3.84}$$

Also, note that

$$g_m = \mu_n C_{ox}\frac{W}{L}(V_{in} - V_{TH} - V_{out}) \tag{3.85}$$

Consequently,

$$A_v = \frac{g_m R_S}{1 + (g_m + g_{mb}) R_S} \tag{3.86}$$

The same result is more easily obtained with the aid of a small-signal equivalent circuit. From Fig. 3.35, we have $V_{in} - V_1 = V_{out}$, $V_{bs} = -V_{out}$, and $g_m V_1 - g_{mb} V_{out} = V_{out}/R_S$. Thus, $V_{out}/V_{in} = g_m R_S/[1 + (g_m + g_{mb}) R_S]$.

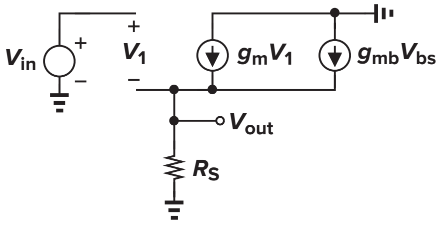


Figure 3.35 Small-signal equivalent circuit of source follower.

Sketched in Fig. 3.36 vs. V_{in}, the voltage gain begins from zero for $V_{in} \approx V_{TH}$ (that is, $g_m \approx 0$) and monotonically increases. As the drain current and g_m increase, A_v approaches $g_m/(g_m + g_{mb}) = 1/(1+\eta)$. Since η itself slowly decreases with V_{out}, A_v would eventually become equal to unity, but for typical allowable source-bulk voltages, η remains greater than roughly 0.2.

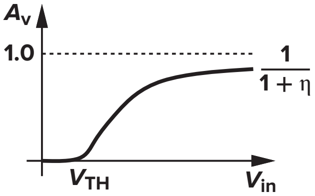


Figure 3.36 Voltage gain of source follower versus input voltage.

An important result of (3.86) is that even if $R_S = \infty$, the voltage gain of a source follower is not equal to one (unless body effect is removed as explained later). We return to this point later. Note that M_1 in Fig. 3.34(a) remains in saturation if V_{in} does not exceed $V_{DD} + V_{TH}$.

In the source follower of Fig. 3.34(a), the drain current of M_1 heavily depends on the input dc level. For example, if V_{in} changes from 0.7 V to 1 V, I_D may increase by a factor of 2, and hence $V_{GS} - V_{TH}$ by $\sqrt{2}$. Even if V_{TH} is relatively constant, the increase in V_{GS} means that V_{out} $(= V_{in} - V_{GS})$ does not follow V_{in} faithfully, thereby incurring nonlinearity. To alleviate this issue, the resistor can be replaced by a constant current source as shown in Fig. 3.37(a). The current source itself is implemented as an NMOS transistor operating in the saturation region [Fig. 3.37(b)].

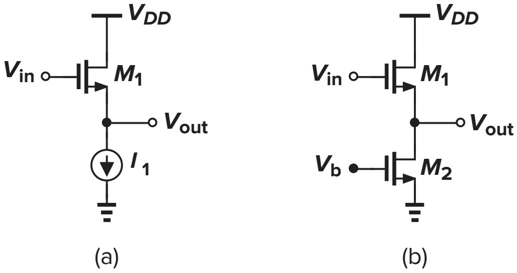


Figure 3.37 Source follower using (a) an ideal current source, and (b) an NMOS transistor as a current source.

▶ Example 3.12

Suppose that in the source follower of Fig. 3.37(a), $(W/L)_1 = 20/0.5$, $I_1 = 200\ \mu$A, $V_{TH0} = 0.6$ V, $2\Phi_F = 0.7$ V, $V_{DD} = 1.2$ V, $\mu_n C_{ox} = 50\ \mu$A/V^2, and $\gamma = 0.4$ V$^{1/2}$.

(a) Calculate V_{out} for $V_{in} = 1.2$ V.

(b) If I_1 is implemented as M_2 in Fig. 3.37(b), find the minimum value of $(W/L)_2$ for which M_2 remains saturated when $V_{in} = 1.2$ V.

Solution

(a) Since the threshold voltage of M_1 depends on V_{out}, we perform a simple iteration. Noting that

$$(V_{in} - V_{TH} - V_{out})^2 = \frac{2I_D}{\mu_n C_{ox} \left(\dfrac{W}{L}\right)_1} \tag{3.87}$$

we first assume that $V_{TH} \approx 0.6$ V, obtaining $V_{out} = 0.153$ V. Now we calculate a new V_{TH} as

$$V_{TH} = V_{TH0} + \gamma(\sqrt{2\Phi_F + V_{SB}} - \sqrt{2\Phi_F}) \tag{3.88}$$

$$= 0.635 \text{ V} \tag{3.89}$$

This indicates that V_{out} is approximately 35 mV less than that calculated above, i.e., $V_{out} \approx 0.118$ V.

(b) Since the drain-source voltage of M_2 is equal to 0.118 V, the device is saturated only if $(V_{GS} - V_{TH})_2 \le 0.118$ V. With $I_D = 200\ \mu$A, this gives $(W/L)_2 \ge 287/0.5$. Note the substantial drain junction and overlap capacitance contributed by M_2 to the output node.

◀

▶ Example 3.13

Explain intuitively why the gain of the source follower in Fig. 3.37(a) is equal to unity if I_1 is ideal and $\lambda = \gamma = 0$.

Solution

In this case, the drain current of M_1 remains exactly constant, and so does V_{GS1}. Since $V_{out} = V_{in} - V_{GS1}$, we observe that a change in V_{in} must equally appear in V_{out}. Alternatively, as shown in Fig. 3.38, we can say that the small-signal drain current cannot flow through any path and must be zero, yielding $V_1 = 0$ and $V_{out} = V_{in}$.

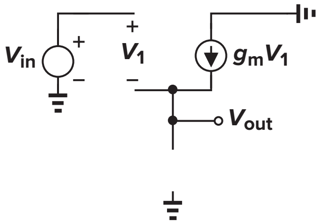


Figure 3.38

◀

To gain a better understanding of source followers, let us calculate the small-signal output resistance of the circuit in Fig. 3.39(a). Using the equivalent circuit of Fig. 3.39(b) and noting that $V_X = -V_{bs}$, we write

$$I_X - g_m V_X - g_{mb} V_X = 0 \tag{3.90}$$

It follows that

$$R_{out} = \frac{1}{g_m + g_{mb}} \tag{3.91}$$

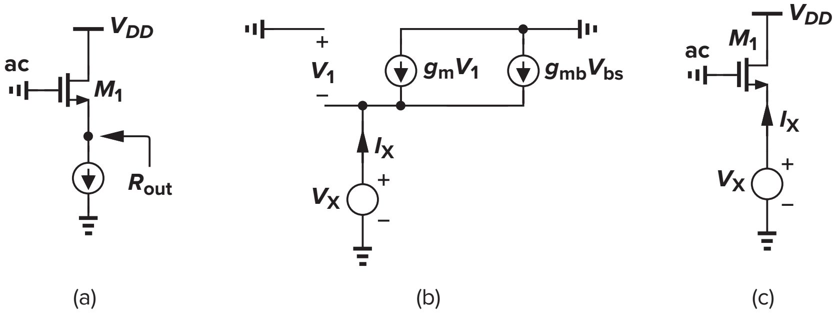


Figure 3.39 Calculation of the output impedance of a source follower.

This result should not come as a surprise: the circuit in Fig. 3.39(b) is similar to that in Fig. 3.11(b). Interestingly, body effect decreases the output resistance of source followers. To understand why, suppose that in Fig. 3.39(c), V_X decreases by ΔV so that the drain current increases. With no body effect, only the gate-source voltage of M_1 would increase by ΔV. With body effect, on the other hand, the threshold voltage of the device decreases as well. Thus, in $(V_{GS} - V_{TH})^2$, the first term increases and the second decreases, resulting in a greater change in the drain current and hence a lower output impedance.

The above phenomenon can also be studied with the aid of the small-signal model shown in Fig. 3.40(a). It is important to note that the magnitude of the current source $g_{mb}V_{bs} = g_{mb}V_X$ is linearly proportional to the voltage across it (because the current source and the voltage source are in parallel). Such a behavior is that of a simple resistor equal to $1/g_{mb}$, yielding the small-signal model shown in Fig. 3.40(b). The equivalent resistor simply appears in parallel with the output, thereby lowering the overall output resistance. Since without $1/g_{mb}$, the output resistance equals $1/g_m$, we conclude that

$$R_{out} = \frac{1}{g_m} \| \frac{1}{g_{mb}} \tag{3.92}$$

$$= \frac{1}{g_m + g_{mb}} \tag{3.93}$$

Modeling the effect of g_{mb} by a resistor—which is valid only for source followers—also helps explain the less-than-unity voltage gain implied by (3.86) for $R_S = \infty$. As shown in the Thevenin equivalent

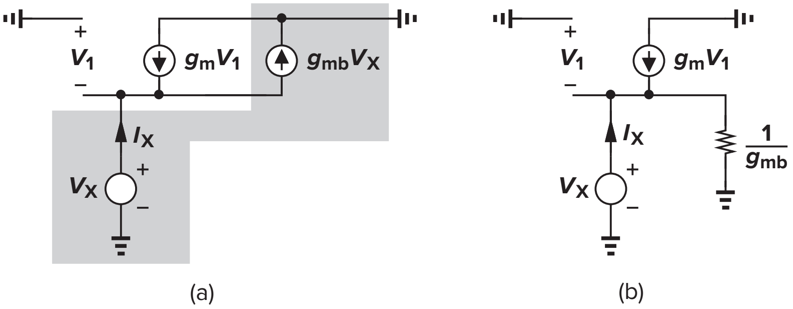


Figure 3.40 Source follower including body effect.

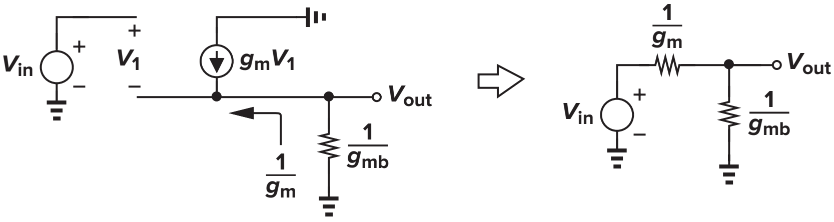


Figure 3.41 Representation of intrinsic source follower by a Thevenin equivalent.

of Fig. 3.41,

$$A_v = \frac{\frac{1}{g_{mb}}}{\frac{1}{g_m} + \frac{1}{g_{mb}}} \tag{3.94}$$

$$= \frac{g_m}{g_m + g_{mb}} \tag{3.95}$$

For completeness, we also study a source follower with a finite load resistance and channel-length modulation [Fig. 3.42(a)]. Noting that $1/g_{mb}$, r_{O1}, r_{O2}, and R_L are in parallel, we can reduce the circuit to that shown in Fig. 3.42(c), where $R_{eq} = (1/g_{mb})||r_{O1}||r_{O2}||R_L$. It follows that

$$A_v = \frac{R_{eq}}{R_{eq} + \frac{1}{g_m}} \tag{3.96}$$

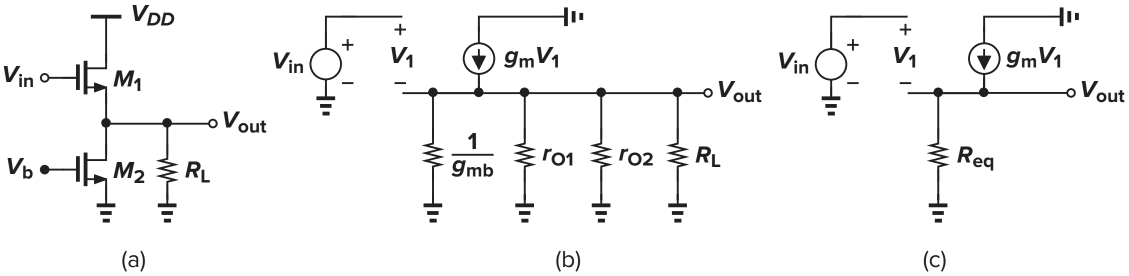


Figure 3.42 (a) Source follower driving load resistance; (b) small-signal equivalent circuit; (c) simplified model.

▶ **Example 3.14**

Calculate the voltage gain of the circuit shown in Fig. 3.43.

Solution

The impedance seen looking into the source of M_2 (a diode-connected device) is equal to $[1/(g_{m2} + g_{mb2})]\|r_{O2}$. The impedance appears in parallel with $1/g_{mb1}$ and r_{O1}. Thus,

$$A_v = \frac{\frac{1}{g_{m2} + g_{mb2}}\|r_{O2}\|r_{O1}\|\frac{1}{g_{mb1}}}{\frac{1}{g_{m2} + g_{mb2}}\|r_{O2}\|r_{O1}\|\frac{1}{g_{mb1}} + \frac{1}{g_{m1}}} \tag{3.97}$$

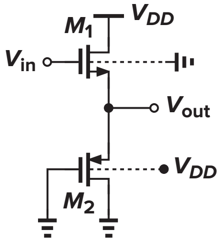


Figure 3.43

Source followers exhibit a high input impedance and a moderate output impedance, but at the cost of two drawbacks: nonlinearity and voltage headroom limitation. We consider these issues in detail.

As mentioned in relation to Fig. 3.34(a), even if a source follower is biased by an ideal current source, its input-output characteristic displays some nonlinearity due to the nonlinear dependence of V_{TH} upon the source potential. In submicron technologies, r_O of the transistor also changes substantially with V_{DS}, thus introducing additional variation in the small-signal gain of the circuit (Chapter 14). For this reason, typical source followers suffer from significant nonlinearity.

The nonlinearity due to body effect can be eliminated if the bulk is tied to the source. This is usually possible only for PFETs because all NFETs share the same substrate. Figure 3.44 shows a PMOS source follower employing two separate n-wells so as to eliminate the body effect of M_1. The lower mobility of PFETs, however, yields a higher output impedance in this case than that available in an NMOS counterpart.

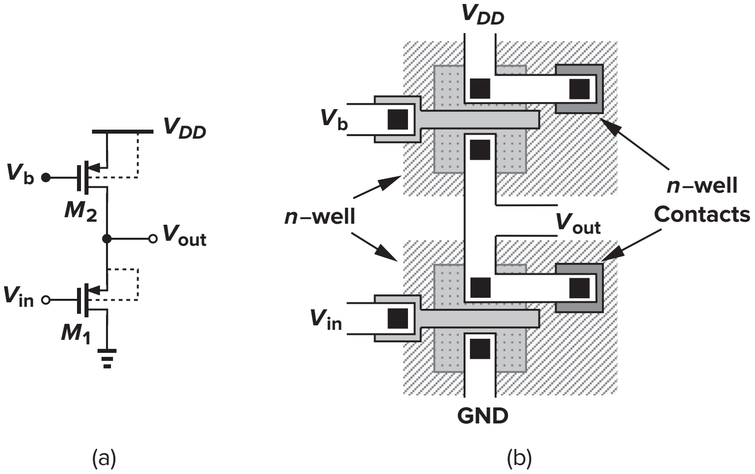


Figure 3.44 (a) PMOS source follower with no body effect; (b) corresponding layout showing separate n-wells.

Source followers also shift the dc level of the signal by V_{GS}, thereby consuming voltage headroom and limiting the voltage swings. To understand this point, consider the example illustrated in Fig. 3.45, a cascade of a common-source stage and a source follower. Without the source follower, the minimum allowable value of V_X would be equal to $V_{GS1} - V_{TH1}$ (for M_1 to remain in saturation). With the source follower, on the other hand, V_X must be greater than $V_{GS2} + (V_{GS3} - V_{TH3})$ so that M_3 is saturated. For comparable overdrive voltages in M_1 and M_3, this means the allowable swing at X is reduced by V_{GS2}, a substantial amount.

It is also instructive to compare the gain of source followers and common-source stages when the load impedance is relatively low. A practical example is the need to drive an external 50-Ω termination in a

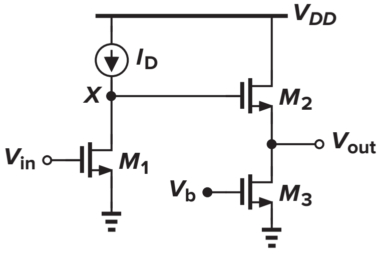


Figure 3.45 Cascade of source follower and CS stage.

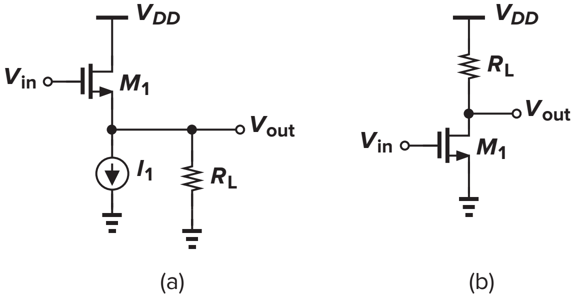


Figure 3.46 (a) Source follower and (b) CS stage driving a load resistance.

high-frequency environment. As shown in Fig. 3.46(a), the load can be driven by a source follower with an overall voltage gain of

$$\frac{V_{out}}{V_{in}}|_{SF} \approx \frac{R_L}{R_L + 1/g_{m1}} \tag{3.98}$$

$$\approx \frac{g_{m1}R_L}{1 + g_{m1}R_L} \tag{3.99}$$

On the other hand, as depicted in Fig. 3.46(b), the load can be included as part of a common-source stage, providing a gain of

$$\frac{V_{out}}{V_{in}}|_{CS} \approx -g_{m1}R_L \tag{3.100}$$

The key difference between these two topologies is the achievable voltage gain for a given bias current. For example, if $1/g_{m1} \approx R_L$, then the source follower exhibits a gain of at most 0.5, whereas the common-source stage provides a gain close to unity. Thus, source followers are not necessarily efficient drivers.

The drawbacks of source followers, namely, nonlinearity due to body effect and voltage headroom consumption due to level shift, limit the use of this topology. As a general rule, we avoid the use of source followers unless they become absolutely necessary. One application of source followers is in performing voltage-level shift, as illustrated by the following example.

▶ **Example 3.15**

(a) In the circuit of Fig. 3.47(a), calculate the voltage gain if C_1 acts as an ac short at the frequency of interest. What is the maximum dc level of the input signal for which M_1 remains saturated?

(b) To accommodate an input dc level close to V_{DD}, the circuit is modified as shown in Fig. 3.47(b). What relationship between the gate-source voltages of M_2 and M_3 guarantees that M_1 is saturated?

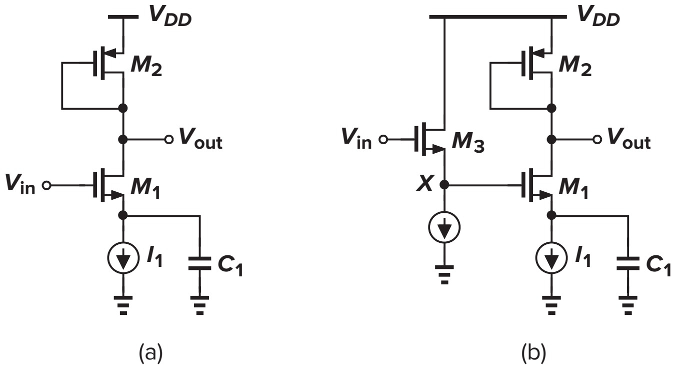


Figure 3.47

Solution

(a) Noting that the source of M_1 is at ac ground, we write the gain as

$$A_v = -g_{m1}[r_{O1} \| r_{O2} \| (1/g_{m2})] \tag{3.101}$$

Since $V_{out} = V_{DD} - |V_{GS2}|$, the maximum allowable dc level of V_{in} is equal to $V_{DD} - |V_{GS2}| + V_{TH1}$.

(b) If $V_{in} = V_{DD}$, then $V_X = V_{DD} - V_{GS3}$. For M_1 to be saturated when $V_{in} = V_{DD}$, we must have $V_{DD} - V_{GS3} - V_{TH1} \leq V_{DD} - |V_{GS2}|$, and hence $V_{GS3} + V_{TH1} \geq |V_{GS2}|$.

As explained in Chapter 7, source followers also introduce substantial noise. For this reason, the circuit of Fig. 3.47(b) is ill-suited to low-noise applications.

3.5 ■ Common-Gate Stage

In common-source amplifiers and source followers, the input signal is applied to the gate of a MOSFET. It is also possible to apply the signal to the source terminal. Shown in Fig. 3.48(a), a common-gate (CG) stage senses the input at the source and produces the output at the drain. The gate is connected to a dc voltage to establish proper operating conditions. Note that the bias current of M_1 flows through the input signal source. Alternatively, as depicted in Fig. 3.48(b), M_1 can be biased by a constant current source, with the signal capacitively coupled to the circuit.

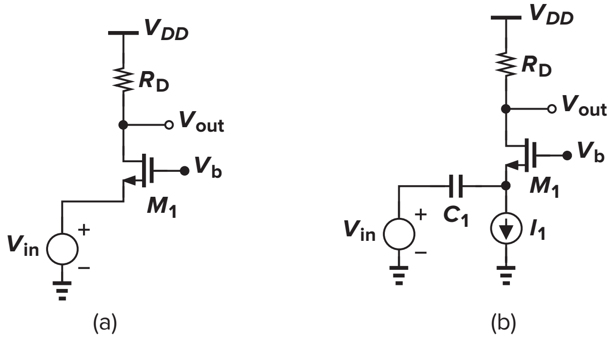


Figure 3.48 (a) Common-gate stage with direct coupling at input; (b) CG stage with capacitive coupling at input.

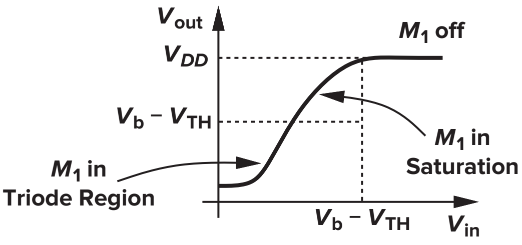


Figure 3.49 Common-gate input-output characteristic.

We first study the large-signal behavior of the circuit in Fig. 3.48(a). For simplicity, let us assume that V_{in} decreases from a large positive value. Also, $\lambda = 0$. For $V_{in} \geq V_b - V_{TH}$, M_1 is off and $V_{out} = V_{DD}$. For lower values of V_{in}, we can write

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{in} - V_{TH})^2 \tag{3.102}$$

if M_1 is in saturation. As V_{in} decreases, so does V_{out}, eventually driving M_1 into the triode region if

$$V_{DD} - \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{in} - V_{TH})^2 R_D = V_b - V_{TH} \tag{3.103}$$

The input-output characteristic is shown in Fig. 3.49, illustrating a case in which M_1 enters the triode region as V_{in} decreases. In the region where M_1 is saturated, we can express the output voltage as

$$V_{out} = V_{DD} - \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{in} - V_{TH})^2 R_D \tag{3.104}$$

obtaining a small-signal gain of

$$\frac{\partial V_{out}}{\partial V_{in}} = -\mu_n C_{ox}\frac{W}{L}(V_b - V_{in} - V_{TH})\left(-1 - \frac{\partial V_{TH}}{\partial V_{in}}\right)R_D \tag{3.105}$$

Since $\partial V_{TH}/\partial V_{in} = \partial V_{TH}/\partial V_{SB} = \eta$, we have

$$\frac{\partial V_{out}}{\partial V_{in}} = \mu_n C_{ox}\frac{W}{L}R_D(V_b - V_{in} - V_{TH})(1+\eta) \tag{3.106}$$

$$= g_m(1+\eta)R_D \tag{3.107}$$

Note that the gain is positive. Interestingly, body effect increases the equivalent transconductance of the stage.

For a given bias current and supply voltage (i.e., a given power budget), how do we maximize the voltage gain of a CG stage? We can increase g_m by widening the input device, eventually reaching subthreshold operation $[g_m \approx I_D/(\xi V_T)]$ (why?), and/or we can increase R_D and, inevitably, the dc drop across it. We must bear in mind that the minimum allowable level of V_{out} in Fig. 3.48(b) is equal to $V_{GS} - V_{TH} + V_{I1}$, where V_{I1} denotes the minimum voltage required by I_1.

▶ **Example 3.16**

(a) Is it possible for M_1 in Fig. 3.48(a) to remain in saturation for the entire range of V_{in} from 0 to V_{DD}?
(b) Is it possible for M_1 to remain in the triode region for the entire range of V_{in} from 0 to V_{DD}?

Solution

(a) Yes, it is possible. To so guarantee, we choose $V_{DD} - R_D I_D > V_b - V_{TH}$, where I_D denotes the drain current at $V_{in} = 0$.
(b) Yes, it is possible. If $V_b > V_{DD} + V_{TH}$, then M_1 turns on at the edge of the triode region at $V_{in} = V_{DD} - V_{TH}$ and goes deeper as V_{in} falls. Of course, this choice of V_b is neither practical nor desirable.

The input impedance of the circuit is also important. We note that for $\lambda = 0$, the impedance seen at the source of M_1 in Fig. 3.48(a) is the same as that at the source of M_1 in Fig. 3.39, namely, $1/(g_m + g_{mb}) = 1/[g_m(1 + \eta)]$. Thus, the body effect decreases the input impedance of the common-gate stage. The relatively low input impedance of the common-gate stage proves useful in some applications.

Example 3.17

In Fig. 3.50, transistor M_1 senses ΔV and delivers a proportional current to a 50-Ω transmission line. The other end of the line is terminated by a 50-Ω resistor in Fig. 3.50(a) and a common-gate stage in Fig. 3.50(b). Assume $\lambda = \gamma = 0$.

(a) Calculate V_{out}/V_{in} at low frequencies for both arrangements.

(b) What condition is necessary to minimize wave reflection at node X?

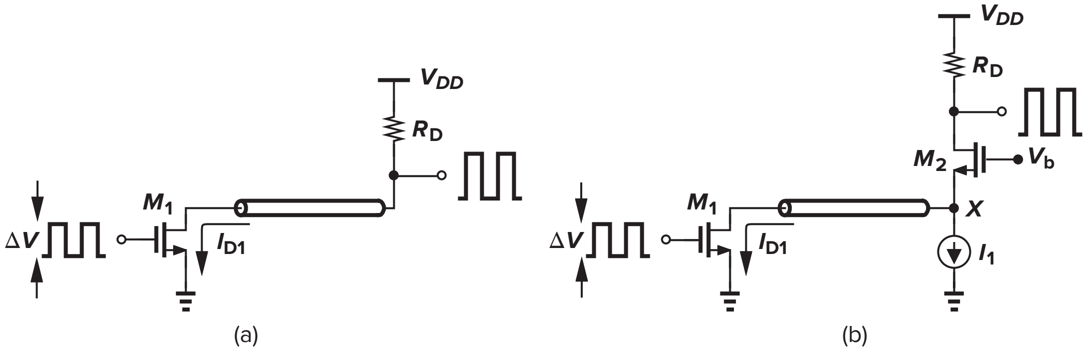


Figure 3.50

Solution

(a) For small signals applied to the gate of M_1, the drain current experiences a change equal to $g_{m1}\Delta V_X$. This current is drawn from R_D in Fig. 3.50(a) and M_2 in Fig. 3.50(b), producing an output voltage swing equal to $-g_{m1}\Delta V_X R_D$. Thus, $A_v = -g_{m1} R_D$ for both cases.

(b) To minimize reflection at node X, the resistance seen at the source of M_2 must equal 50 Ω and the reactance must be small. Thus, $1/(g_{m2} + g_{mb2}) = 50$ Ω, which can be ensured by proper sizing and biasing of M_2. To minimize the capacitances of the transistor, it is desirable to use a small device biased at a large current. (Recall that $g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$.) In addition to higher power dissipation, this remedy also requires a large V_{GS} for M_2.

The key point in this example is that, while the overall voltage gain in both arrangements equals $-g_{m1} R_D$, the value of R_D in Fig. 3.50(b) can be much greater than 50 Ω without introducing reflections at point X. Thus, the common-gate circuit can provide a higher voltage gain than that in Fig. 3.50(a).

Now let us study the common-gate topology in a more general case, taking into account both the output impedance of the transistor and the impedance of the signal source. Depicted in Fig. 3.51(a), the circuit can be analyzed with the aid of its equivalent shown in Fig. 3.51(b). Noting that the current flowing

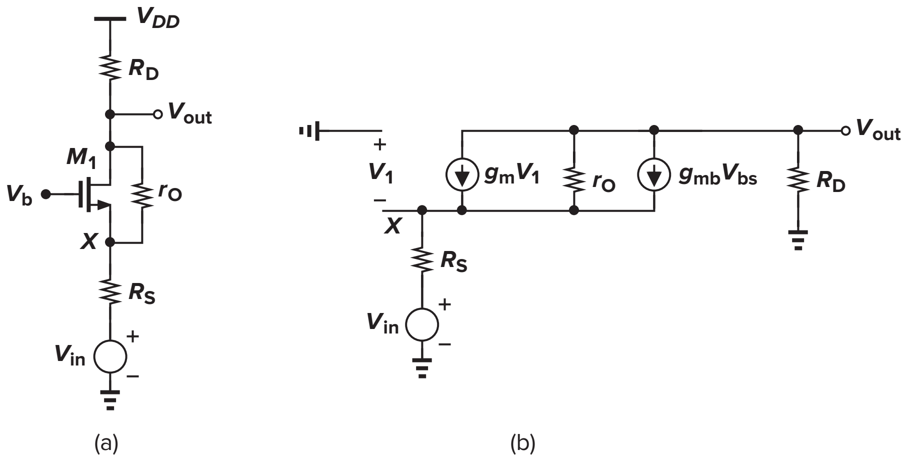


Figure 3.51 (a) CG stage with finite transistor output resistance; (b) small-signal equivalent circuit.

through R_S is equal to $-V_{out}/R_D$, we have

$$V_1 - \frac{V_{out}}{R_D} R_S + V_{in} = 0 \tag{3.108}$$

Moreover, since the current through r_O is equal to $-V_{out}/R_D - g_m V_1 - g_{mb} V_1$, we can write

$$r_O \left(\frac{-V_{out}}{R_D} - g_m V_1 - g_{mb} V_1 \right) - \frac{V_{out}}{R_D} R_S + V_{in} = V_{out} \tag{3.109}$$

Upon substitution for V_1 from (3.108), (3.109) reduces to

$$r_O \left[\frac{-V_{out}}{R_D} - (g_m + g_{mb}) \left(V_{out} \frac{R_S}{R_D} - V_{in} \right) \right] - \frac{V_{out} R_S}{R_D} + V_{in} = V_{out} \tag{3.110}$$

It follows that

$$\frac{V_{out}}{V_{in}} = \frac{(g_m + g_{mb}) r_O + 1}{r_O + (g_m + g_{mb}) r_O R_S + R_S + R_D} R_D \tag{3.111}$$

Note the similarity between (3.111) and (3.77). The gain of the common-gate stage is slightly higher due to body effect.

▶ Example 3.18

Calculate the voltage gain of the circuit shown in Fig. 3.52(a) if $\lambda \neq 0$ and $\gamma \neq 0$.

Solution

We first find the Thevenin equivalent of M_1 and V_{in}. As shown in Fig. 3.52(b), M_1 operates as a source follower, the equivalent Thevenin voltage is given by

$$V_{in,eq} = \frac{r_{O1} \left\| \dfrac{1}{g_{mb1}} \right.}{r_{O1} \left\| \dfrac{1}{g_{mb1}} \right. + \dfrac{1}{g_{m1}}} V_{in} \tag{3.112}$$

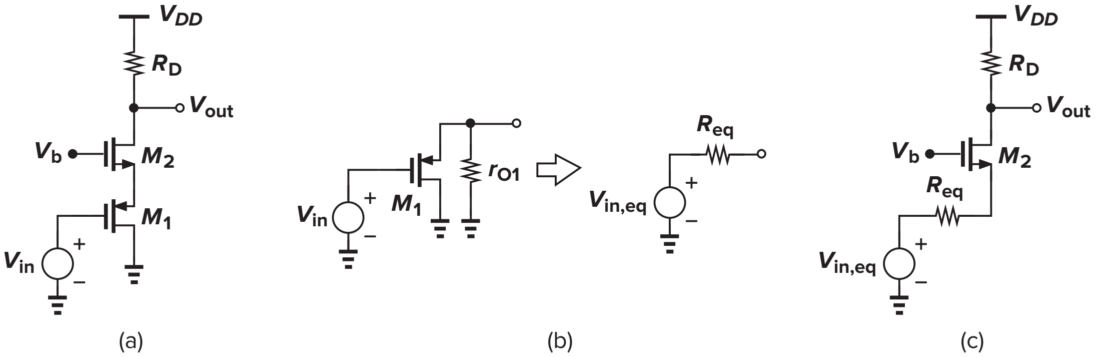


Figure 3.52

and the equivalent Thevenin resistance is

$$R_{eq} = r_{O1} \left\| \frac{1}{g_{mb1}} \right\| \frac{1}{g_{m1}} \tag{3.113}$$

Redrawing the circuit as in Fig. 3.52(c), we use (3.111) to write

$$\frac{V_{out}}{V_{in}} = \frac{(g_{m2}+g_{mb2})r_{O2}+1}{r_{O2}+[1+(g_{m2}+g_{mb2})r_{O2}]\left(r_{O1}\left\|\frac{1}{g_{mb1}}\right\|\frac{1}{g_{m1}}\right)+R_D} R_D \frac{r_{O1}\left\|\frac{1}{g_{mb1}}\right.}{r_{O1}\left\|\frac{1}{g_{mb1}}\right. + \frac{1}{g_{m1}}} \tag{3.114}$$

This example demonstrates the ease with which a circuit can be analyzed by inspection—while relying on previously-derived results—rather than by blindly writing KVLs and KCLs.

The input and output impedances of the common-gate topology are also of interest. To obtain the impedance seen at the source [Fig. 3.53(a)], we use the equivalent circuit in Fig. 3.53(b). Since $V_1 = -V_X$ and the current through r_O is equal to $I_X + g_m V_1 + g_{mb} V_1 = I_X - (g_m + g_{mb})V_X$, we can add up the voltages across r_O and R_D and equate the result to

$$R_D I_X + r_O[I_X - (g_m + g_{mb})V_X] = V_X \tag{3.115}$$

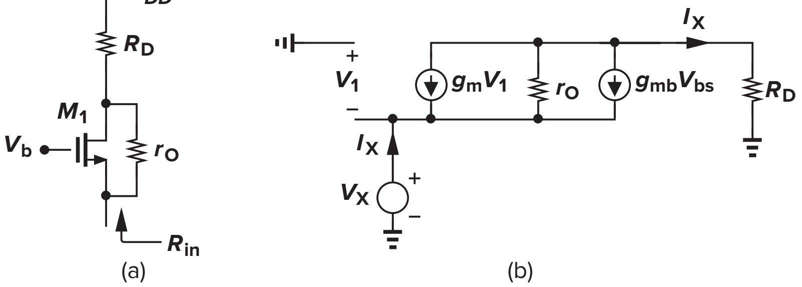


Figure 3.53 (a) Input resistance of a CG stage; (b) small-signal equivalent circuit.

Thus,

$$\frac{V_X}{I_X} = \frac{R_D + r_O}{1 + (g_m + g_{mb})r_O} \tag{3.116}$$

$$\approx \frac{R_D}{(g_m + g_{mb})r_O} + \frac{1}{g_m + g_{mb}} \tag{3.117}$$

if $(g_m + g_{mb})r_O \gg 1$. This result reveals that the drain impedance is divided by $(g_m + g_{mb})r_O$ when seen at the source. This is particularly important in short-channel devices because of their low intrinsic gain. Two special cases of (3.116) are worth studying. First, suppose $R_D = 0$. Then,

$$\frac{V_X}{I_X} = \frac{r_O}{1 + (g_m + g_{mb})r_O} \tag{3.118}$$

$$= \frac{1}{\dfrac{1}{r_O} + g_m + g_{mb}} \tag{3.119}$$

which is simply the impedance seen at the source of a source follower, a predictable result because if $R_D = 0$, the circuit configuration is the same as in Fig. 3.39(a).

Second, let us replace R_D with an ideal current source. Equation (3.117) predicts that the input impedance approaches *infinity*. While somewhat surprising, this result can be explained with the aid of Fig. 3.54. Since the total current through the transistor is fixed and equal to I_1, a change in the source potential cannot change the device current, and hence $I_X = 0$. In other words, the input impedance of a common-gate stage is relatively low *only* if the load impedance connected to the drain is small.

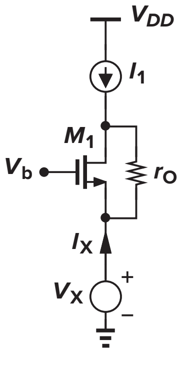


Figure 3.54 Input resistance of a CG stage with ideal current-source load.

▶ **Example 3.19**

Calculate the voltage gain of a common-gate stage with a current-source load [Fig. 3.55(a)].

Solution

Letting R_D approach infinity in (3.111), we have

$$A_v = (g_m + g_{mb})r_O + 1 \tag{3.120}$$

Interestingly, the gain does not depend on R_S. From our foregoing discussion, we recognize that if $R_D \to \infty$, so does the impedance seen at the source of M_1, and the small-signal voltage at node X becomes *equal* to V_{in}. We can therefore simplify the circuit as shown in Fig. 3.55(b), readily arriving at (3.120).

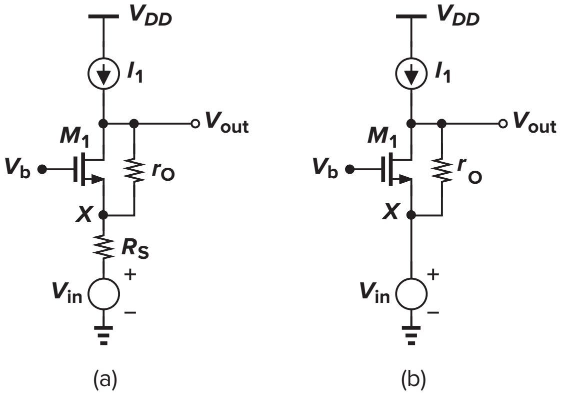
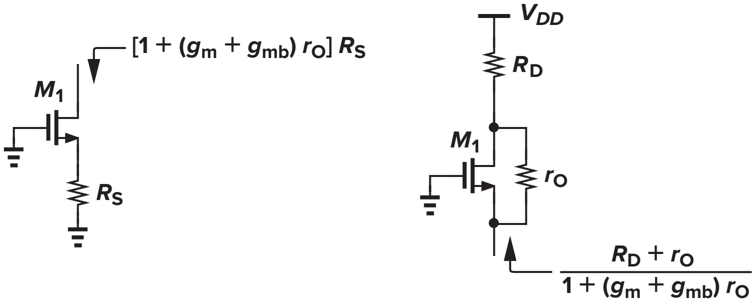


Figure 3.55

Our analysis of the degenerated CS stage and the CG stage gives another interesting insight. As illustrated in Fig. 3.56, we loosely say that a transistor transforms its source resistance *up* and its drain resistance *down* (when seen at the appropriate terminal).



$[1 + (g_m + g_{mb})\, r_O]\, R_S$
M_1
R_S
V_{DD}
R_D
M_1
r_O
$\dfrac{R_D + r_O}{1 + (g_m + g_{mb})\, r_O}$

Figure 3.56 Impedance transformation by a MOSFET.

In order to calculate the output impedance of the common-gate stage, we use the circuit in Fig. 3.57. We note that the result is similar to that in Fig. 3.29, and hence

$$R_{out} = \{[1 + (g_m + g_{mb})r_O]R_S + r_O\} \| R_D \tag{3.121}$$

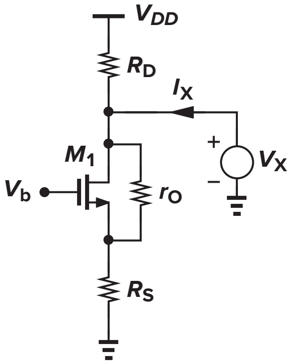


Figure 3.57 Calculation of output resistance of a CG stage.

Example 3.20

As seen in Example 3.17, the input signal of a common-gate stage may be a current rather than a voltage. Shown in Fig. 3.58 is such an arrangement. Calculate V_{out}/I_{in} and the output impedance of the circuit if the input current source exhibits an output impedance equal to R_P.

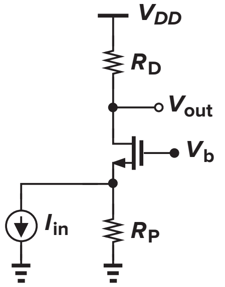


Figure 3.58

Solution

To find V_{out}/I_{in}, we replace I_{in} and R_P with a Thevenin equivalent and use (3.111) to write

$$\frac{V_{out}}{I_{in}} = \frac{(g_m + g_{mb})r_O + 1}{r_O + (g_m + g_{mb})r_O R_P + R_P + R_D} R_D R_P \tag{3.122}$$

The output impedance is simply equal to

$$R_{out} = \{[1 + (g_m + g_{mb})r_O]R_P + r_O\} \| R_D \tag{3.123}$$

3.6 ■ Cascode Stage

As mentioned in Example 3.17, the input signal of a common-gate stage may be a current. We also know that a transistor in a common-source arrangement converts a voltage signal to a current signal. The cascade of a CS stage and a CG stage is called a "cascode"[4] topology, providing many useful properties. Figure 3.59 shows the basic configuration: M_1 generates a small-signal drain current proportional to the small-signal input voltage, V_{in}, and M_2 simply routes the current to R_D. We call M_1 the input device and M_2 the cascode device. Note that in this example, M_1 and M_2 carry equal bias and signal currents. As we describe the attributes of the circuit in this section, many advantages of the cascode topology over a simple common-source stage become evident. This circuit is also known as the "telescopic" cascode.

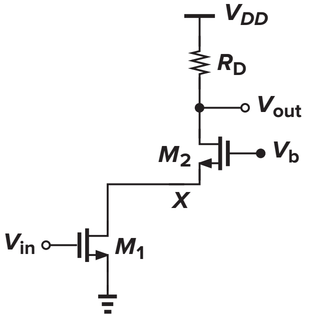


Figure 3.59 Cascode stage.

[4] The term cascode is believed to be the acronym for "cascaded triodes," possibly invented in vacuum tube days.

Before delving into our analysis, it is instructive to explore the circuit qualitatively. We wish to know what happens if the value of V_{in} or V_b changes by a small amount. Assume that both transistors are in saturation and $\lambda = \gamma = 0$. If V_{in} rises by ΔV, then I_{D1} increases by $g_{m1}\Delta V$. This change in current flows through the impedance seen at X, i.e., the impedance seen at the source of M_2, which is equal to $1/g_{m2}$. Thus, V_X falls by an amount given by $g_{m1}\Delta V \cdot (1/g_{m2})$ [Fig. 3.60(a)]. The change in I_{D1} also flows through R_D, producing a drop of $g_{m1}\Delta V R_D$ in V_{out}—just as in a simple CS stage.

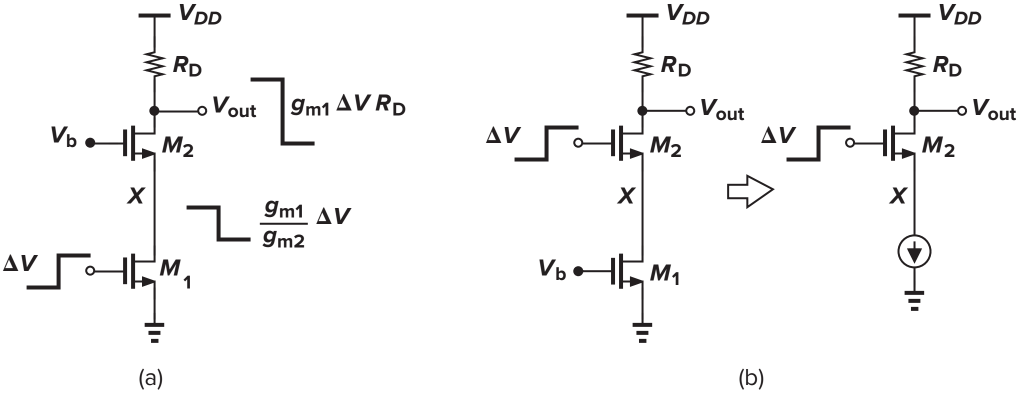


Figure 3.60 Cascode stage sensing a signal at the gate of (a) an input device and (b) a cascode device.

Now, consider the case where V_{in} is fixed and V_b increases by ΔV. Since V_{GS1} is constant and $r_{O1} = \infty$, we simplify the circuit as shown in Fig. 3.60(b). How do V_X and V_{out} change here? As far as node X is concerned, M_2 operates as a *source follower* because it senses an input, ΔV, at its gate and generates an output at X. With $\lambda = \gamma = 0$, the small-signal voltage gain of the follower is equal to unity, regardless of the value of R_D (why?). Thus, V_X rises by ΔV. On the other hand, V_{out} does *not* change because I_{D2} is equal to I_{D1} and hence remains *constant*. We say that the voltage gain from V_b to V_{out} is zero in this case.

Let us now study the bias conditions of the cascode, still assuming that $\lambda = \gamma = 0$. For M_1 to operate in saturation, we must have $V_X \geq V_{in} - V_{TH1}$. If M_1 and M_2 are both in saturation, M_2 operates as a source follower and V_X is determined primarily by V_b: $V_X = V_b - V_{GS2}$. Thus, $V_b - V_{GS2} \geq V_{in} - V_{TH1}$, and hence $V_b > V_{in} + V_{GS2} - V_{TH1}$ (Fig. 3.61). For M_2 to be saturated, $V_{out} \geq V_b - V_{TH2}$; that is,

$$V_{out} \geq V_{in} - V_{TH1} + V_{GS2} - V_{TH2} \tag{3.124}$$

$$= (V_{GS1} - V_{TH1}) + (V_{GS2} - V_{TH2}) \tag{3.125}$$

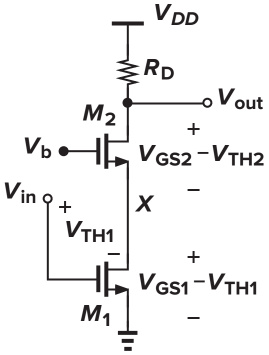


Figure 3.61 Allowable voltages in cascode stage.

if V_b is chosen to place M_1 at the edge of saturation. Consequently, the minimum output level for which both transistors operate in saturation is equal to the overdrive voltage of M_1 plus that of M_2. In other words, addition of M_2 to the circuit reduces the output voltage swing by at least the overdrive voltage of M_2. We say that M_2 is "stacked" on top of M_1. We also loosely say that the minimum output voltage is equal to two overdrives or $2V_{D,sat}$.

We now analyze the large-signal behavior of the cascode stage shown in Fig. 3.59 as V_{in} goes from zero to V_{DD}. For $V_{in} \leq V_{TH1}$, M_1 and M_2 are off, $V_{out} = V_{DD}$, and $V_X \approx V_b - V_{TH2}$ (if subthreshold conduction is neglected) (Fig. 3.62). As V_{in} exceeds V_{TH1}, M_1 begins to draw current, and V_{out} drops. Since I_{D2} increases, V_{GS2} must increase as well, causing V_X to fall. As V_{in} assumes sufficiently large values, two effects can occur: (1) V_X drops below V_{in} by V_{TH1}, forcing M_1 into the triode region; (2) V_{out} drops below V_b by V_{TH2}, driving M_2 into the triode region. Depending on the device dimensions and the values of R_D and V_b, one effect may occur before the other. For example, if V_b is relatively low, M_1 may enter the triode region first. Note that if M_2 goes into the deep triode region, V_X and V_{out} become nearly equal.

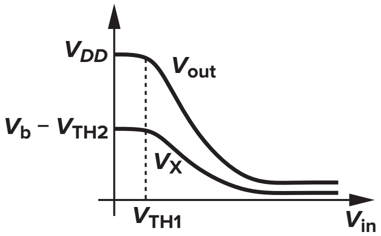


Figure 3.62 Input-output characteristic of a cascode stage.

Let us now consider the small-signal characteristics of a cascode stage, assuming that both transistors operate in saturation. If $\lambda = 0$, the voltage gain is equal to that of a common-source stage because the drain current produced by the input device must flow through the cascode device. Illustrated in the equivalent circuit of Fig. 3.63, this result is independent of the transconductance and body effect of M_2. It can also be verified using $A_v = -G_m R_{out}$.

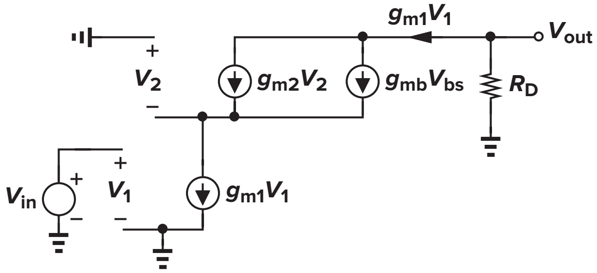


Figure 3.63 Small-signal equivalent circuit of cascode stage.

▶ **Example 3.21**

Calculate the voltage gain of the circuit shown in Fig. 3.64 if $\lambda = 0$.

Solution

The small-signal drain current of M_1, $g_{m1}V_{in}$, is divided between R_P and the impedance seen looking into the source of M_2, $1/(g_{m2} + g_{mb2})$. Thus, the current flowing through M_2 is

$$I_{D2} = g_{m1}V_{in}\frac{(g_{m2} + g_{mb2})R_P}{1 + (g_{m2} + g_{mb2})R_P} \tag{3.126}$$

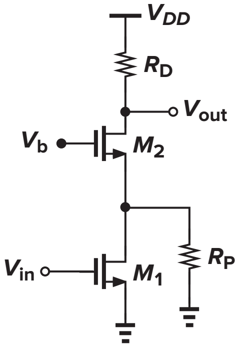


Figure 3.64

The voltage gain is therefore given by

$$A_v = -\frac{g_{m1}(g_{m2} + g_{mb2})R_P R_D}{1 + (g_{m2} + g_{mb2})R_P} \tag{3.127}$$

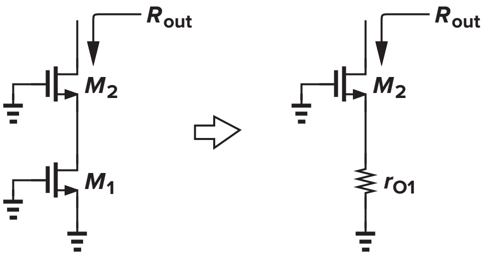


Figure 3.65 Calculation of output resistance of cascode stage.

Output Resistance An important property of the cascode structure is its high output impedance. As illustrated in Fig. 3.65, for calculation of R_{out}, the circuit can be viewed as a common-source stage with a degeneration resistor equal to r_{O1}. Thus, from (3.66),

$$R_{out} = [1 + (g_{m2} + g_{mb2})r_{O2}]r_{O1} + r_{O2} \tag{3.128}$$

Assuming $g_m r_O \gg 1$, we have $R_{out} \approx (g_{m2} + g_{mb2})r_{O2}r_{O1}$. That is, M_2 boosts the output impedance of M_1 by a factor of $(g_{m2} + g_{mb2})r_{O2}$. As shown in Fig. 3.66, cascoding can be extended to three or more stacked devices to achieve a higher output impedance, but the required additional voltage

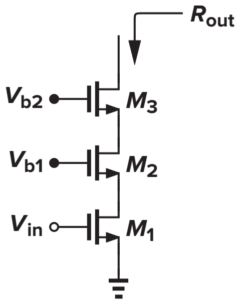
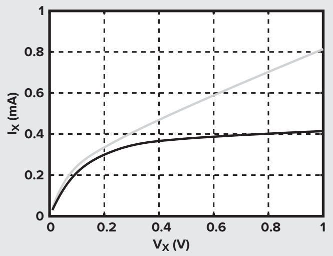


Figure 3.66 Triple cascode.

Nanometer Design Notes

With a limited voltage headroom, nanometer cascode current sources are only moderately better than single transistors. The figure shows the I-V characteristic of an NMOS current source before and after cascoding (gray and black curves, respectively). Here, W/L = 5 μm/40 nm for both devices. We observe that for V_X < 0.2 V, the cascode has only a slightly higher output impedance.



headroom makes such configurations less attractive. For example, the minimum output voltage of a triple cascode is equal to the sum of three overdrive voltages.

To appreciate the utility of a high output impedance, recall from the lemma in Sec. 3.3.3 that the voltage gain can be written as $-G_m R_{out}$. Since G_m is typically determined by the transconductance of a transistor, e.g., M_1 in Fig. 3.59, and hence bears trade-offs with the bias current and device capacitances, it is desirable to increase the voltage gain by maximizing R_{out}. Shown in Fig. 3.67 is an example. If both M_1 and M_2 operate in saturation, then $G_m \approx g_{m1}$ and $R_{out} \approx (g_{m2} + g_{mb2})r_{O2}r_{O1}$, yielding $A_v = (g_{m2} + g_{mb2})r_{O2}g_{m1}r_{O1}$. Thus, the maximum voltage gain is roughly equal to the *square* of the intrinsic gain of the transistors.

▶ **Example 3.22**

Calculate the exact voltage gain of the circuit shown in Fig. 3.67.

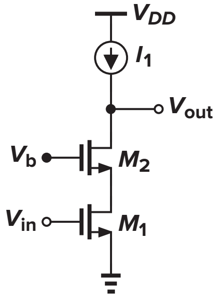


Figure 3.67 Cascode stage with current-source load.

Solution

The actual G_m of the stage is slightly less than g_{m1} because a fraction of the small-signal current produced by M_1 is shunted to ground by r_{O1}. As depicted in Fig. 3.68(a), we short the output node to ac ground and recognize that the impedance seen looking into the source of M_2 is equal to $[1/(g_{m2} + g_{mb2})]||r_{O2}$. Thus,

$$I_{out} = g_{m1}V_{in}\frac{r_{O1}}{r_{O1} + \dfrac{1}{g_{m2} + g_{mb2}}\bigg|\bigg| r_{O2}} \tag{3.129}$$

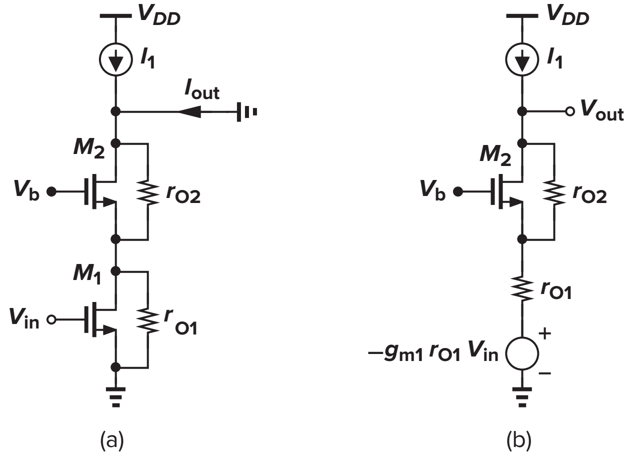


Figure 3.68

It follows that the overall transconductance is equal to

$$G_m = \frac{g_{m1} r_{O1}[r_{O2}(g_{m2} + g_{mb2}) + 1]}{r_{O1} r_{O2}(g_{m2} + g_{mb2}) + r_{O1} + r_{O2}} \tag{3.130}$$

and hence the voltage gain is given by

$$|A_v| = G_m R_{out} \tag{3.131}$$

$$= g_{m1} r_{O1}[(g_{m2} + g_{mb2}) r_{O2} + 1] \tag{3.132}$$

If we had assumed that $G_m \approx g_{m1}$, then $|A_v| \approx g_{m1}\{[1 + (g_{m2} + g_{mb2})r_{O2}]r_{O1} + r_{O2}\}$.

Another approach to calculating the voltage gain is to replace V_{in} and M_1 by a Thevenin equivalent, reducing the circuit to a common-gate stage. Illustrated in Fig. 3.68(b), this method in conjunction with (3.111) gives the same result as (3.132).

It is also interesting to compare the increase in the output impedance due to cascoding with that due to increasing the length of the input transistor for a given bias current (Fig. 3.69). Suppose, for example, that the length of the input transistor of a CS stage is quadrupled while the width remains constant. Then, since $I_D = (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$, the overdrive voltage is doubled, and the transistor consumes the same amount of voltage headroom as does a cascode stage. That is, the circuits of Figs. 3.69(b) and (c) impose equal voltage swing constraints.

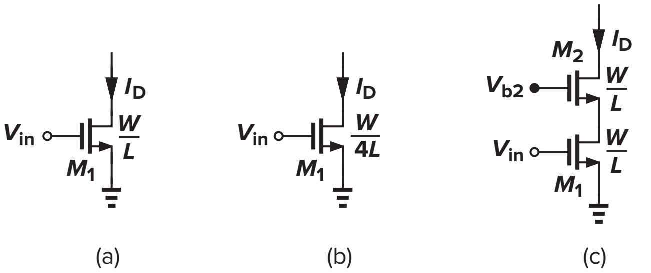


Figure 3.69 Increasing output impedance by increasing the device length or cascoding.

Now consider the output impedance achieved in each case. Since

$$g_m r_O = \sqrt{2\mu_n C_{ox}\frac{W}{L} I_D}\,\frac{1}{\lambda I_D} \tag{3.133}$$

and $\lambda \propto 1/L$, quadrupling L only doubles the value of $g_m r_O$ while cascoding results in an output impedance of roughly $g_m r_O^2$. Note that the transconductance of M_1 in Fig. 3.69(b) is half that in Fig. 3.69(c), degrading the performance. In other words, for a given voltage headroom, the cascode structure provides a higher output impedance.

A cascode structure need not operate as an amplifier. Another popular application of this topology is in building constant current sources. The high output impedance yields a current source closer to the ideal, but at the cost of voltage headroom. For example, current source I_1 in Fig. 3.67 can be implemented as a PMOS cascode (Fig. 3.70), exhibiting an impedance equal to $[1 + (g_{m3} + g_{mb3})r_{O3}]r_{O4} + r_{O3}$.

We calculate the voltage gain with the aid of the lemma illustrated in Fig. 3.32. Writing $G_m \approx g_{m1}$, we note that R_{out} is now equal to the parallel combination of the NMOS cascode output impedance and the PMOS cascode output impedance:

$$R_{out} = \{[1 + (g_{m2} + g_{mb2})r_{O2}]r_{O1} + r_{O2}\} \| \{[1 + (g_{m3} + g_{mb3})r_{O3}]r_{O4} + r_{O3}\} \tag{3.134}$$

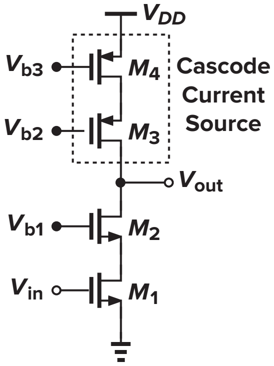


Figure 3.70 NMOS cascode amplifier with PMOS cascode load.

The gain is given by $|A_v| \approx g_{m1} R_{out}$. For typical values, we approximate the voltage gain as

$$|A_v| \approx g_{m1}[(g_{m2}r_{O2}r_{O1})\|(g_{m3}r_{O3}r_{O4})] \tag{3.135}$$

▶ **Example 3.23**

How much voltage swing can the cascode amplifier of Fig. 3.70 support at the output?

Solution

Recall from Fig. 3.61 that V_{b1} can be chosen low enough to place M_1 at the edge of saturation, $V_{b1} = V_{GS2} + (V_{GS1} - V_{TH1})$, allowing a *minimum* value of $(V_{GS2} - V_{TH2}) + (V_{GS1} - V_{TH1})$ for V_{out}. Similarly, V_{b2} can be chosen high enough to bias M_4 at the edge of saturation: $V_{b2} + |V_{GS3}| = V_{DD} - |V_{GS4} - V_{TH4}|$. This choice allows a *maximum* value of $V_{DD} - |V_{GS4} - V_{TH4}| - |V_{GS3} - V_{TH3}|$ for V_{out}. Thus, the total allowable voltage swing at the output is equal to

$$V_{out,max} - V_{out,min} = V_{DD} - (V_{GS1} - V_{TH1}) - (V_{GS2} - V_{TH2}) - |V_{GS3} - V_{TH3}| - |V_{GS4} - V_{TH4}| \tag{3.136}$$

We loosely say that the output swing is equal to V_{DD} minus four overdrives or $4V_{D,sat}$.

◀

We should caution the reader that the dc value at the output of the cascode amplifier shown in Fig. 3.70 is poorly defined because two possibly unequal high-impedance current sources are placed in series. (What happens if two unequal ideal current sources appear in series?) For this reason, the circuit must be biased in a negative-feedback loop.

Poor Man's Cascode A "minimalist" cascode current source omits the bias voltage necessary for the cascode device. Shown in Fig. 3.71, this "poor man's cascode" places M_2 in the triode region because $V_{GS1} > V_{TH1}$ and $V_{DS2} = V_{GS2} - V_{GS1} < V_{GS2} - V_{TH2}$. In fact, if M_1 and M_2 have identical dimensions, it can be proved that the structure is equivalent to a single transistor having twice the length—not really a cascode.

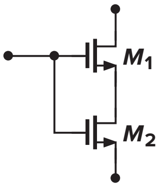


Figure 3.71 Poor man's cascode.

In modern CMOS technologies, however, transistors with different threshold voltages are available, allowing M_2 to operate in saturation if M_1 has a lower threshold than M_2. For example, if $V_{TH2} - V_{TH1} =$ 150 mV and if $V_{GS1} - V_{TH1} <$ 100 mV, then M_2 is saturated and the circuit acts as a true cascode.

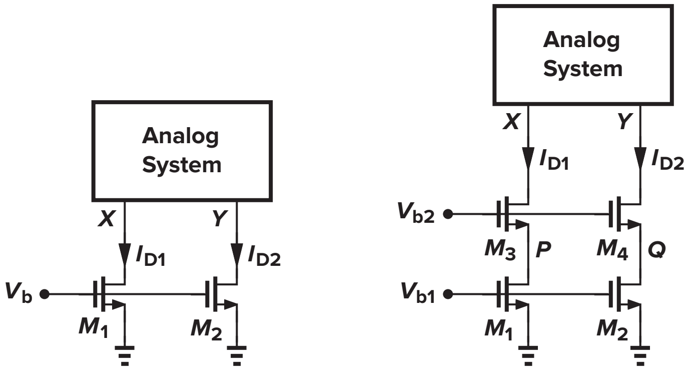
Shielding Property Recall from Fig. 3.30 that the high output impedance arises from the fact that if the output-node voltage is changed by ΔV, the resulting change at the source of the cascode device is much less. In a sense, the cascode transistor "shields" the input device from voltage variations at the output. The shielding property of cascodes proves useful in many circuits.

▶ Example 3.24

Two identical NMOS transistors are used as constant current sources in a system [Fig. 3.72(a)]. However, due to the internal circuitry of the system, V_X is higher than V_Y by ΔV.

(a) Calculate the resulting difference between I_{D1} and I_{D2} if $\lambda \neq 0$.

(b) Add cascode devices to M_1 and M_2 and repeat part (a).



(a) (b)

Figure 3.72

Solution

(a) We have

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{TH})^2(\lambda V_{DS1} - \lambda V_{DS2}) \tag{3.137}$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{TH})^2(\lambda \Delta V) \tag{3.138}$$

(b) As shown in Fig. 3.72(b), cascoding reduces the effect of V_X and V_Y upon I_{D1} and I_{D2}, respectively. As depicted in Fig. 3.30 and implied by Eq. (3.69), a difference ΔV between V_X and V_Y translates to a difference ΔV_{PQ} between P and Q equal to

$$\Delta V_{PQ} = \Delta V \frac{r_{O1}}{[1 + (g_{m3} + g_{mb3})r_{O3}]r_{O1} + r_{O3}} \tag{3.139}$$

$$\approx \frac{\Delta V}{(g_{m3} + g_{mb3})r_{O3}} \tag{3.140}$$

Thus,

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_b - V_{TH})^2 \frac{\lambda \Delta V}{(g_{m3} + g_{mb3})r_{O3}} \tag{3.141}$$

In other words, cascoding reduces the mismatch between I_{D1} and I_{D2} by a factor of $(g_{m3} + g_{mb3})r_{O3}$.

The shielding property of cascodes diminishes if the cascode device enters the triode region. To understand why, let us consider the circuit in Fig. 3.73, assuming that V_X decreases from a large positive value. As V_X falls below $V_{b2} - V_{TH2}$, M_2 enters the triode region and requires a greater gate-source overdrive so as to sustain the current drawn by M_1. We can write

$$I_{D2} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_2 [2(V_{b2} - V_P - V_{TH2})(V_X - V_P) - (V_X - V_P)^2] \tag{3.142}$$

concluding that as V_X decreases, V_P also drops, so that I_{D2} remains constant. In other words, variation of V_X is less attenuated as it appears at P. If V_X falls sufficiently, V_P goes below $V_{b1} - V_{TH1}$, driving M_1 into the triode region.

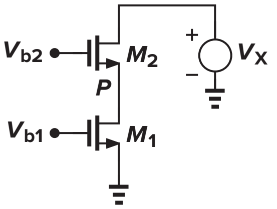


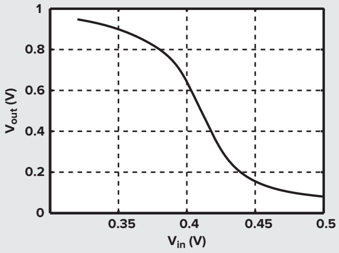
Figure 3.73 Output swing of cascode stage.

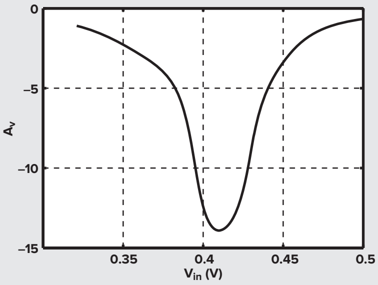
3.6.1 Folded Cascode

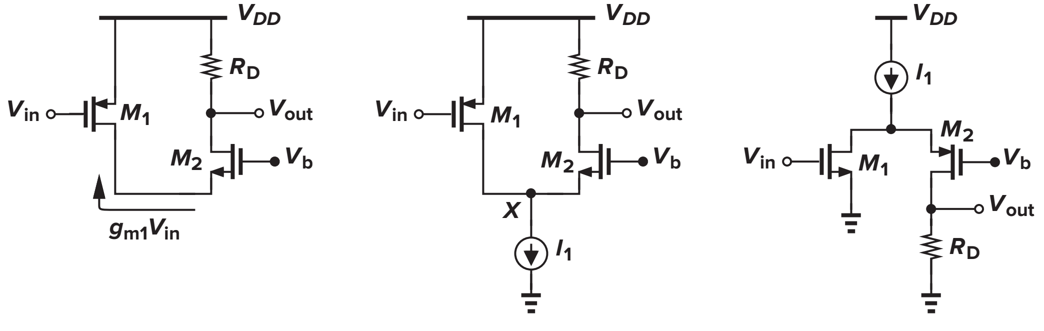
The idea behind the cascode structure is to convert the input voltage to a current and apply the result to a common-gate stage. However, the input device and the cascode device need not be of the same type. For example, as depicted in Fig. 3.74(a), a PMOS-NMOS combination performs the same function. In order to bias M_1 and M_2, a current source must be added as in Fig. 3.74(b). Note that $|I_{D1}| + I_{D2}$ is equal to I_1 and hence constant. The small-signal operation is as follows. If V_{in} becomes more positive, $|I_{D1}|$ decreases, forcing I_{D2} to increase and hence V_{out} to drop. The voltage gain and output impedance of the circuit can be obtained as calculated for the NMOS-NMOS cascode of Fig. 3.59. Shown in Fig. 3.74(c) is an NMOS-PMOS cascode. The advantages and disadvantages of these types will be explained later.

Nanometer Design Notes

Let us implement a cascode stage with $(W/L)_{NMOS} = 10\ \mu$m/40 nm, $(W/L)_{PMOS} = 20\ \mu$m/40 nm, and $I_D = 0.3$ mA. The figure plots the input-output characteristic of the circuit, displaying nonlinearity at the extremes of V_{out}. How do we quantify this nonlinearity? We can say that the output swing should not cause, say, more than a 20% drop in the small-signal voltage gain. Plotting the derivative of the characteristic in the figure, we observe that the allowable single-ended peak-to-peak output swing is about 0.5 V if the gain must remain greater than 10. We also note that the gain is fairly low, dictating longer PMOS devices if a higher gain is necessary.







(a) (b) (c)

Figure 3.74 (a) Simple folded cascode; (b) folded cascode with proper biasing; (c) folded cascode with NMOS input.

The structures of Figs. 3.74(b) and (c) are called "folded cascode" stages because the small-signal current is "folded" up [in Fig. 3.74(b)] or down [in Fig. 3.74(c)]. We should mention as a point of contrast that the bias current of M_1 in Fig. 3.70 flows through M_2, i.e., it is "reused," whereas those of M_1 and M_2 in Fig. 3.74(b) add up to I_1. Thus, the total bias current in this case must be higher than that in Fig. 3.70 to achieve a comparable performance.

It is instructive to examine the large-signal behavior of a folded-cascode stage. Suppose that in Fig. 3.74(b), V_{in} decreases from V_{DD} to zero. For $V_{in} > V_{DD} - |V_{TH1}|$, M_1 is off and M_2 carries all of I_1,[5] yielding $V_{out} = V_{DD} - I_1 R_D$. For $V_{in} < V_{DD} - |V_{TH1}|$, M_1 turns on in saturation, giving

$$I_{D2} = I_1 - \frac{1}{2}\mu_p C_{ox}\left(\frac{W}{L}\right)_1 (V_{DD} - V_{in} - |V_{TH1}|)^2 \tag{3.143}$$

As V_{in} drops, I_{D2} decreases further, falling to zero if $I_{D1} = I_1$. This occurs at $V_{in} = V_{in1}$ if

$$\frac{1}{2}\mu_p C_{ox}\left(\frac{W}{L}\right)_1 (V_{DD} - V_{in1} - |V_{TH1}|)^2 = I_1 \tag{3.144}$$

Thus,

$$V_{in1} = V_{DD} - \sqrt{\frac{2I_1}{\mu_p C_{ox}(W/L)_1}} - |V_{TH1}| \tag{3.145}$$

If V_{in} falls below this level, I_{D1} tends to be greater than I_1, and M_1 enters the triode region so as to ensure $I_{D1} = I_1$. The result is plotted in Fig. 3.75. The reader is encouraged to determine the input voltage at which $|I_{D1}| = I_{D2}$.

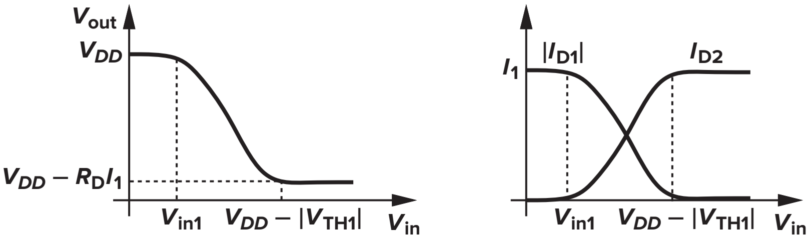


Figure 3.75 Large-signal characteristics of folded cascode.

What happens to V_X in the above test? As I_{D2} drops, V_X rises, reaching $V_b - V_{TH2}$ for $I_{D2} = 0$. As M_1 enters the triode region, V_X approaches V_{DD}.

▶ Example 3.25

Calculate the output impedance of the folded cascode shown in Fig. 3.76(a), where M_3 operates as the bias current source.

Solution

Using the simplified model in Fig. 3.76(b) and Eq. (3.66), we have

$$R_{out} = [1 + (g_{m2} + g_{mb2})r_{O2}](r_{O1} \| r_{O3}) + r_{O2} \tag{3.146}$$

[5] If I_1 is excessively large, M_2 may enter the deep triode region, possibly driving I_1 into the triode region as well.

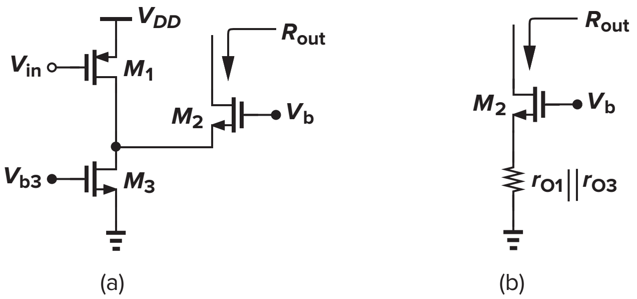


Figure 3.76

Thus, the circuit exhibits an output impedance lower than that of a nonfolded (also called "telescopic") cascode.

In order to achieve a high voltage gain, the load of a folded cascode can be implemented as a cascode itself (Fig. 3.77). This structure is studied more extensively in Chapter 9.

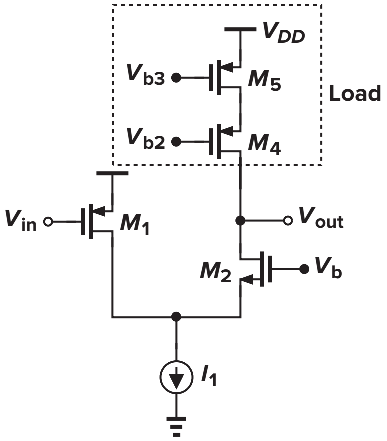


Figure 3.77 Folded cascode with cascode load.

Throughout this chapter, we have attempted to *increase* the output resistance of voltage amplifiers so as to obtain a high gain. This may seem to make the speed of the circuit quite susceptible to the load capacitance. However, as explained in Chapter 8, a high output impedance per se does not pose a serious issue if the amplifier is placed in a proper feedback loop.

3.7 ■ Choice of Device Models

In this chapter, we have developed various expressions for the properties of single-stage amplifiers. For example, the voltage gain of a degenerated common-source stage can be as simple as $-R_D/(R_S+g_m^{-1})$ or as complex as Eq. (3.77). How does one choose a sufficiently accurate device model or expression?

The proper choice is not always straightforward, and making it is a skill gained through practice, experience, and intuition. However, some general principles in choosing the model for each transistor can be followed. First, break the circuit down into a number of familiar topologies. Next, concentrate on each subcircuit and use the simplest transistor model (a single voltage-dependent current source for

FETs operating in saturation) for all transistors. If the drain of a device is connected to a high impedance (e.g., the drain of another), then add r_O to its model. At this point, the basic properties of most circuits can be determined by inspection. In a second, more accurate iteration, the body effect of devices whose source or bulk is not at ac ground can be included as well.

For bias calculations, it is usually adequate to neglect channel-length modulation and body effect in the first pass. These effects do introduce some error, but they can be included in the next iteration step—after the basic properties are understood.

In today's analog design, simulation of circuits is essential because the behavior of short-channel MOSFETs cannot be predicted accurately by hand calculations. Nonetheless, if the designer avoids a simple and intuitive analysis of the circuit and hence skips the task of gaining insight, then he/she cannot interpret the simulation results intelligently. For this reason, we say, "Don't let the computer think for you." Some say, "Don't be a SPICE monkey."

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. All device dimensions are effective values and in microns.

3.1. For the circuit of Fig. 3.13, calculate the small-signal voltage gain if $(W/L)_1 = 50/0.5$, $(W/L)_2 = 10/0.5$, and $I_{D1} = I_{D2} = 0.5$ mA. What is the gain if M_2 is implemented as a diode-connected PMOS device (Fig. 3.16)?

3.2. In the circuit of Fig. 3.18, assume that $(W/L)_1 = 50/0.5$, $(W/L)_2 = 50/2$, and $I_{D1} = I_{D2} = 0.5$ mA when both devices are in saturation. Recall that $\lambda \propto 1/L$.

(a) Calculate the small-signal voltage gain.

(b) Calculate the maximum output voltage swing while both devices are saturated.

3.3. In the circuit of Fig. 3.4(a), assume that $(W/L)_1 = 50/0.5$, $R_D = 2$ kΩ, and $\lambda = 0$.

(a) What is the small-signal gain if M_1 is in saturation and $I_D = 1$ mA?

(b) What input voltage places M_1 at the edge of the triode region? What is the small-signal gain under this condition?

(c) What input voltage drives M_1 into the triode region by 50 mV? What is the small-signal gain under this condition?

3.4. Suppose the common-source stage of Fig. 3.4(a) is to provide an output swing from 1 V to 2.5 V. Assume that $(W/L)_1 = 50/0.5$, $R_D = 2$ kΩ, and $\lambda = 0$.

(a) Calculate the input voltages that yield $V_{out} = 1$ V and $V_{out} = 2.5$ V.

(b) Calculate the drain current and the transconductance of M_1 for both cases.

(c) How much does the small-signal gain, $g_m R_D$, vary as the output goes from 1 V to 2.5 V? (Variation of small-signal gain can be viewed as nonlinearity.)

3.5. Calculate the intrinsic gain of an NMOS device and a PMOS device operating in saturation with $W/L = 50/0.5$ and $|I_D| = 0.5$ mA. Repeat these calculations if $W/L = 100/1$.

3.6. Assuming a constant L, plot the intrinsic gain of a satuated device versus the gate-source voltage if **(a)** the drain current is constant, **(b)** W is constant.

3.7. Assuming a constant L, plot the intrinsic gain of a saturated device versus W/L if **(a)** the gate-source voltage is constant, **(b)** the drain current is constant.

3.8. An NMOS transistor with $W/L = 50/0.5$ is biased with $V_G = +1.2$ V and $V_S = 0$. The drain voltage is varied from 0 to 3 V.

(a) Assuming the bulk voltage is zero, plot the intrinsic gain versus V_{DS}.

(b) Repeat part **(a)** for a bulk voltage of -1 V.

3.9. For an NMOS device operating in saturation, plot g_m, r_O, and $g_m r_O$ as the bulk voltage goes from 0 to $-\infty$ while other terminal voltages remain constant.

3.10. Consider the circuit of Fig. 3.13 with $(W/L)_1 = 50/0.5$ and $(W/L)_2 = 10/0.5$. Assume that $\lambda = \gamma = 0$.

(a) At what input voltage is M_1 at the edge of the triode region? What is the small-signal gain under this condition?

(b) What input voltage drives M_1 into the triode region by 50 mV? What is the small-signal gain under this condition?

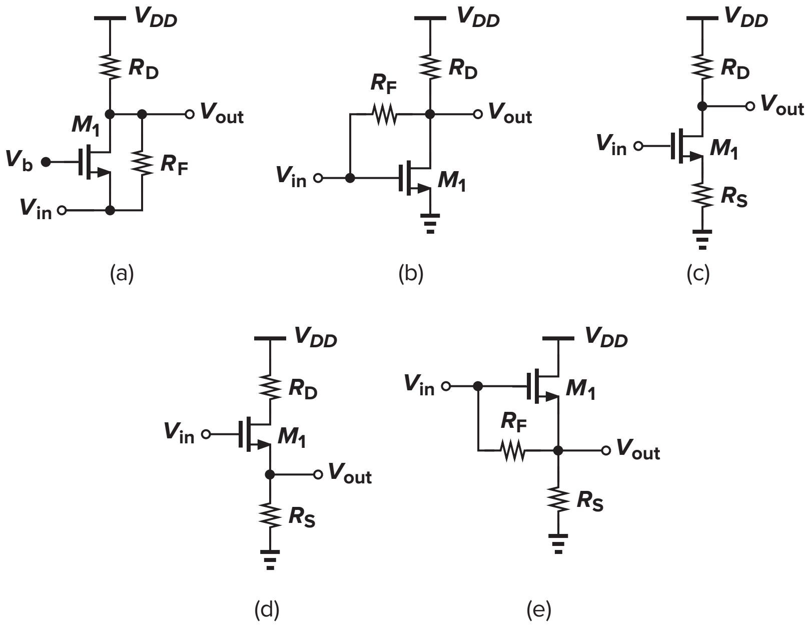
3.11. Repeat Problem 3.10 if body effect is not neglected.

3.12. In the circuit of Fig. 3.17, $(W/L)_1 = 20/0.5$, $I_1 = 1$ mA, and $I_S = 0.75$ mA. Assuming $\lambda = 0$, calculate $(W/L)_2$ such that M_1 is at the edge of the triode region. What is the small-signal voltage gain under this condition?

3.13. Plot the small-signal gain of the circuit shown in Fig. 3.17 as I_S goes from 0 to $0.75I_1$. Assume that M_1 is always saturated, and neglect channel-length modulation and body effect.

3.14. The circuit of Fig. 3.18 is designed to provide an output voltage swing of 2.2 V with a bias current of 1 mA and a small-signal voltage gain of 100. Calculate the dimensions of M_1 and M_2.

3.15. Sketch V_{out} versus V_{in} for the circuits of Fig. 3.78 as V_{in} varies from 0 to V_{DD}. Identify important transition points.



(a) (b) (c)

(d) (e)

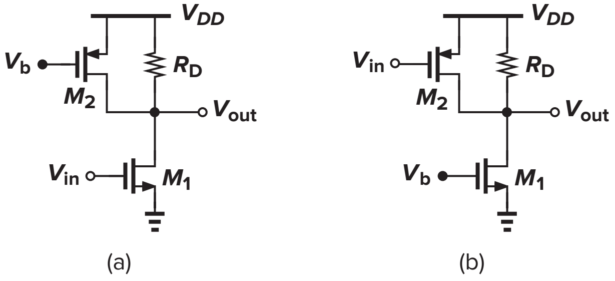
Figure 3.78

3.16. Sketch V_{out} versus V_{in} for the circuits of Fig. 3.79 as V_{in} varies from 0 to V_{DD}. Identify important transition points.

3.17. Sketch V_{out} versus V_{in} for the circuits of Fig. 3.80 as V_{in} varies from 0 to V_{DD}. Identify important transition points.

3.18. Sketch I_X versus V_X for the circuits of Fig. 3.81 as V_X varies from 0 to V_{DD}. Identify important transition points.

3.19. Sketch I_X versus V_X for the circuits of Fig. 3.82 as V_X varies from 0 to V_{DD}. Identify important transition points.



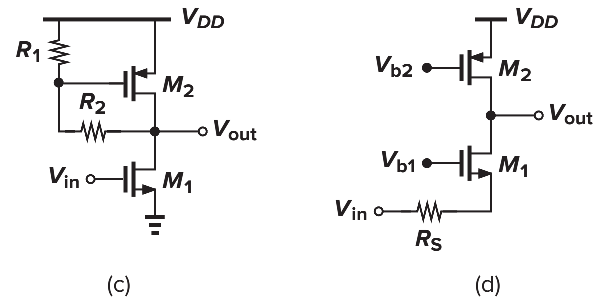


Figure 3.79

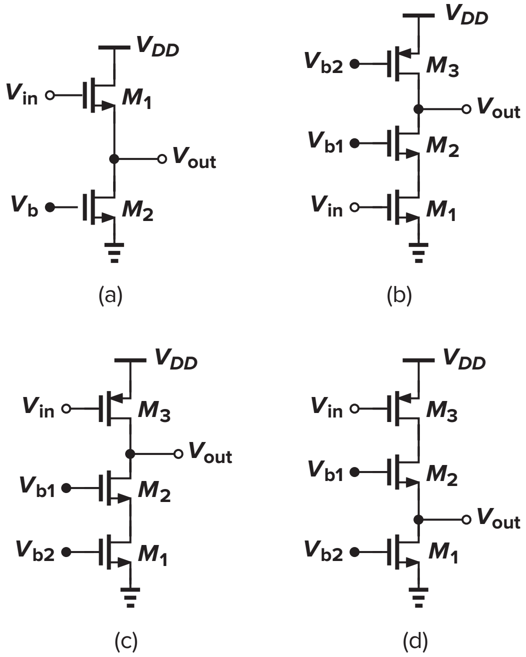
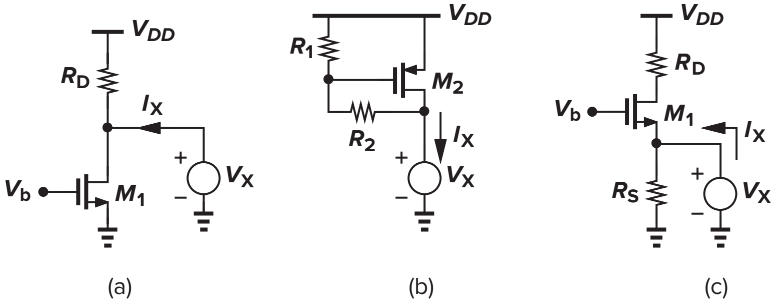
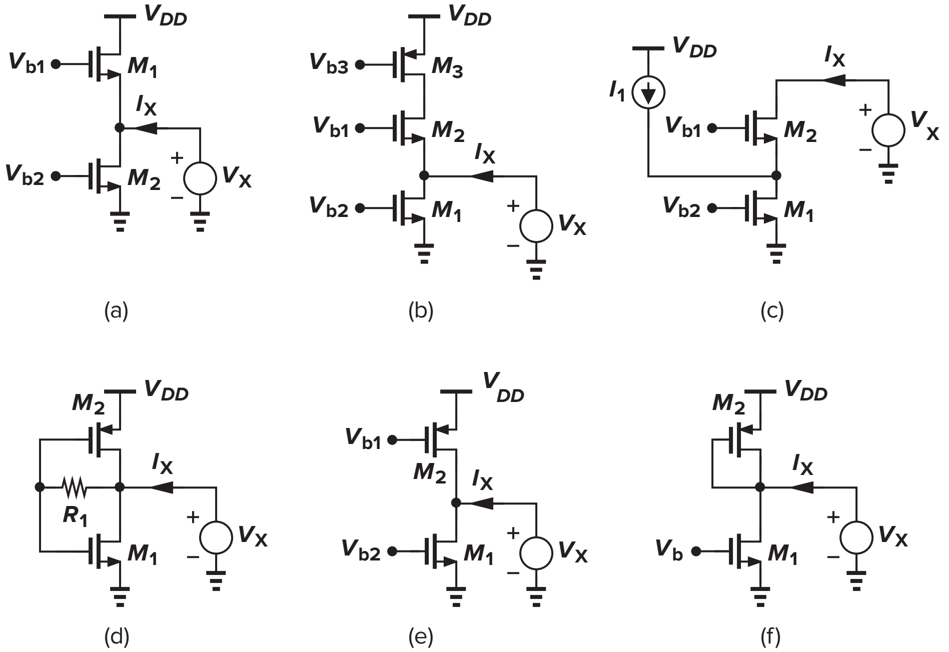


Figure 3.80



(a) (b) (c)

Figure 3.81



(a) (b) (c)

(d) (e) (f)

Figure 3.82

3.20. Assuming all MOSFETs are in saturation, calculate the small-signal voltage gain of each circuit in Fig. 3.83 ($\lambda \neq 0, \gamma = 0$).

3.21. Assuming all MOSFETs are in saturation, calculate the small-signal voltage gain of each circuit in Fig. 3.84 ($\lambda \neq 0, \gamma = 0$).

3.22. Sketch V_X and V_Y as a function of time for each circuit in Fig. 3.85. The initial voltage across C_1 is equal to V_{DD}.

3.23. In the cascode stage of Fig. 3.59, assume that $(W/L)_1 = 50/0.5$, $(W/L)_2 = 10/0.5$, $I_{D1} = I_{D2} = 0.5$ mA, and $R_D = 1$ kΩ.

(a) Choose V_b such that M_1 is 50 mV away from the triode region.

(b) Calculate the small-signal voltage gain.

(c) Using the value of V_b found in part (a), calculate the maximum output voltage swing. Which device enters the triode region first as V_{out} falls?

(d) Calculate the swing at node X for the maximum output swing obtained above.

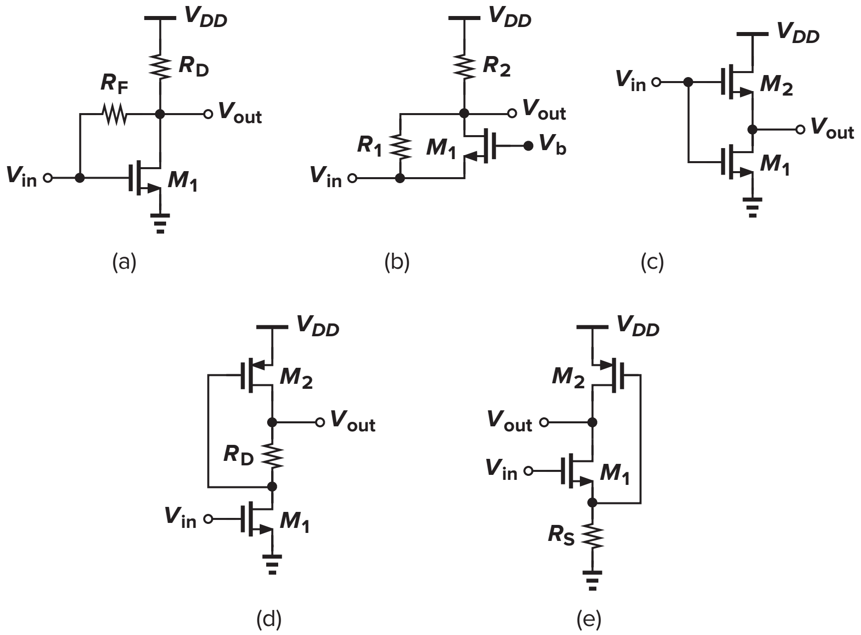


Figure 3.83

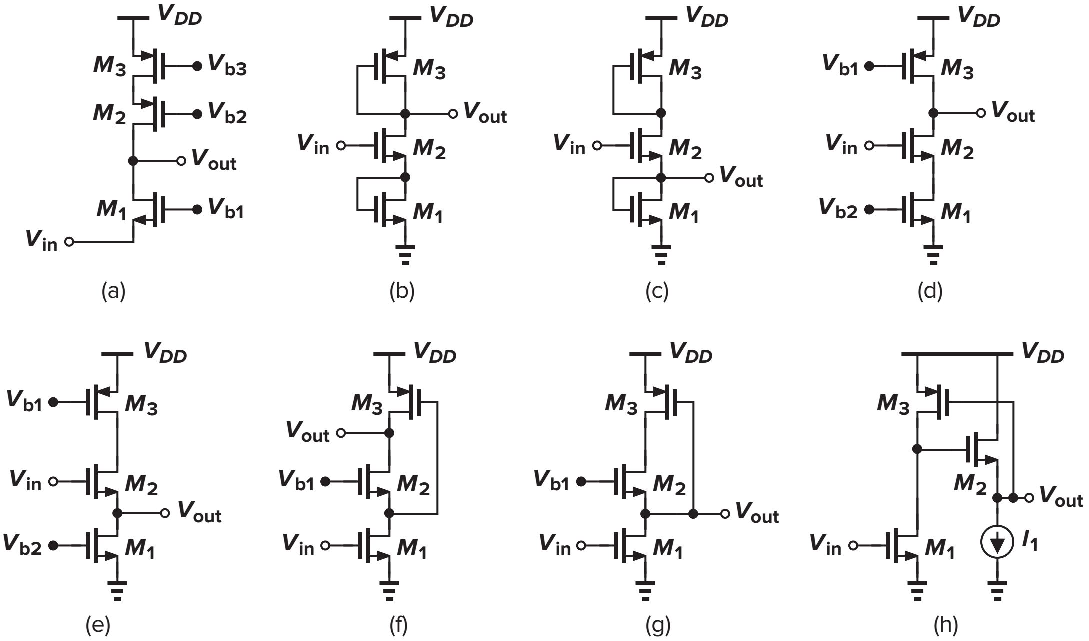


Figure 3.84

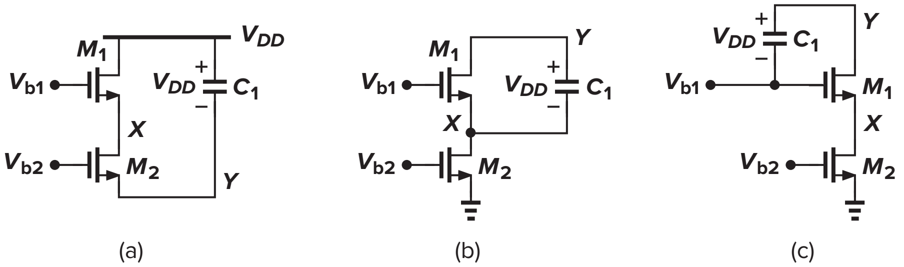


Figure 3.85

3.24. Consider the circuit of Fig. 3.23 with $(W/L)_1 = 50/0.5$, $R_D = 2\ \text{k}\Omega$, and $R_S = 200\ \Omega$.

(a) Calculate the small-signal voltage gain if $I_D = 0.5$ mA.

(b) Assuming $\lambda = \gamma = 0$, calculate the input voltage that places M_1 at the edge of the triode region. What is the gain under this condition?

3.25. Suppose the circuit of Fig. 3.22 is designed for a voltage gain of 5. If $(W/L)_1 = 20/0.5$, $I_{D1} = 0.5$ mA, and $V_b = 0$ V:

(a) Calculate the aspect ratio of M_2.

(b) What input level places M_1 at the edge of the triode region? What is the small-signal gain under this condition?

(c) What input level places M_2 at the edge of the saturation region? What is the small-signal gain under this condition?

3.26. Sketch the small-signal voltage gain of the circuit shown in Fig. 3.22 as V_b varies from 0 to V_{DD}. Consider two cases:

(a) M_1 enters the triode region before M_2 is saturated;

(b) M_1 enters the triode region after M_2 is saturated.

3.27. A source follower can operate as a level shifter. Suppose the circuit of Fig. 3.37(b) is designed to shift the voltage level by 1 V, i.e., $V_{in} - V_{out} = 1$ V.

(a) Calculate the dimensions of M_1 and M_2 if $I_{D1} = I_{D2} = 0.5$ mA, $V_{GS2} - V_{GS1} = 0.5$ V, and $\lambda = \gamma = 0$.

(b) Repeat part (a) if $\gamma = 0.45\ \text{V}^{-1}$ and $V_{in} = 2.5$ V. What is the minimum input voltage for which M_2 remains saturated?

3.28. Sketch the small-signal gain, V_{out}/V_{in}, of the cascode stage shown in Fig. 3.59 as V_b goes from 0 to V_{DD}. Assume that $\lambda = \gamma = 0$.

3.29. The cascode of Fig. 3.70 is designed to provide an output swing of 1.9 V with a bias current of 0.5 mA. If $\gamma = 0$ and $(W/L)_{1-4} = W/L$, calculate V_{b1}, V_{b2}, and W/L. What is the voltage gain if $L = 0.5\ \mu$m?

3.30. Consider the gate-source voltage of M_1 in Fig. 3.23(a): $V_{GS} = V_{in} - I_D R_S$. Determine ΔV_{GS} in response to a change in V_{in} and show that it *decreases* as $g_m R_S$ increases. How does this trend show that the circuit becomes more linear?

3.31. Prove that the voltage gain from V_{DD} to V_{out} in Fig. 3.21 is given by

$$\frac{V_{out}}{V_{in}} = \frac{g_{m2} r_{O2} + 1}{r_{O2} + r_{O1}} r_{O1} \tag{3.147}$$

3.32. In the circuit shown in Fig. 3.86, prove that

$$\frac{V_{out1}}{V_{out2}} = \frac{-R_D}{R_S} \tag{3.148}$$

where V_{out1} and V_{out2} are small-signal quantities and $\lambda,\ \gamma > 0$.

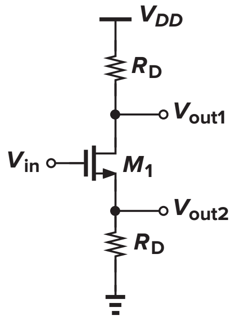


Figure 3.86

3.33. The CG stage of Fig. 3.51(a) is designed such that its input resistance (seen at node X) matches the signal source resistance, R_S. If $\lambda,\ \gamma > 0$, prove that

$$\frac{V_{out}}{V_{in}} = \frac{1 + (g_m + g_{mb})r_O}{2 + \left(1 + \dfrac{r_O}{R_D}\right)} \tag{3.149}$$

Also, prove that

$$\frac{V_{out}}{V_{in}} = \frac{R_D}{2R_S} \tag{3.150}$$

3.34. Calculate the voltage gain of a source follower using the lemma $A_v = -G_m R_{out}$. Assume that the circuit drives a load resistance of R_L and $\lambda,\ \gamma > 0$.

3.35. Calculate the voltage gain of a common-gate stage using the lemma $A_v = -G_m R_{out}$. Assume a source resistance of R_S and $\lambda,\ \gamma > 0$.

3.36. How many amplifier topologies can you create using each of the structures shown in Fig. 3.87 and no other transistors? (The source and drain terminals can be swapped.)

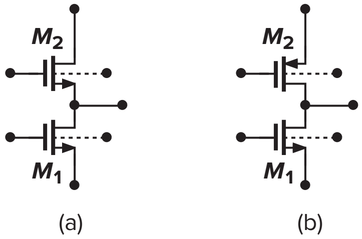


Figure 3.87

CHAPTER

4

Differential Amplifiers

The differential amplifier is among the most important circuit inventions, dating back to the vacuum tube era. Offering many useful properties, differential operation has become the de facto choice in today's high-performance analog and mixed-signal circuits.

This chapter deals with the analysis and design of CMOS differential amplifiers. Following a review of single-ended and differential operation, we describe the basic differential pair and analyze both the large-signal and the small-signal behavior. Next, we introduce the concept of common-mode rejection and formulate it for differential amplifiers. We then study differential pairs with diode-connected and current-source loads as well as differential cascode stages. Finally, we describe the Gilbert cell.

4.1 ■ Single-Ended and Differential Operation

A "single-ended" signal is defined as one that is measured with respect to a fixed potential, usually the ground [Fig. 4.1(a)]. A differential signal is defined as one that is measured between two nodes that have *equal* and *opposite* signal excursions around a fixed potential [Fig. 4.1(b)]. In the strict sense, the two nodes must also exhibit equal impedances to that potential. The "center" potential in differential signaling is called the "common-mode" (CM) level. It is helpful to view the CM level as the bias value of the voltages, i.e., the value in the absence of signals.

The specification of signal swings in a differential system can be confusing. Suppose each single-ended output in Fig. 4.1(b) has a peak amplitude of V_0. Then, the single-ended peak-to-peak swing is $2V_0$ and the differential peak-to-peak swing is $4V_0$. For example, if the voltage at X (with respect to ground) is $V_0 \cos \omega t + V_{CM}$ and that at Y is $-V_0 \cos \omega t + V_{CM}$, then the peak-to-peak swing of $V_X - V_Y$ ($=2V_0 \cos \omega t$) is $4V_0$. It is therefore not surprising that a circuit with a supply voltage of 1 V can deliver a peak-to-peak differential swing of 1.6 V.

An important advantage of differential operation over single-ended signaling is higher immunity to "environmental" noise. Consider the example depicted in Fig. 4.2, where two adjacent lines in a circuit carry a small, sensitive signal and a large clock waveform. Due to capacitive coupling between the lines, transitions on line L_2 corrupt the signal on line L_1. Now suppose, as shown in Fig. 4.2(b), the sensitive signal is distributed as two equal and opposite phases. If the clock line is placed midway between the two, the transitions disturb the differential phases by equal amounts, leaving the *difference* intact. Since the common-mode level of the two phases is disturbed, but the differential output is not corrupted, we say that this arrangement "rejects" common-mode noise.[1]

[1] It is also possible to place a "shield" line between the sensitive line and the clock line (Chapter 19).

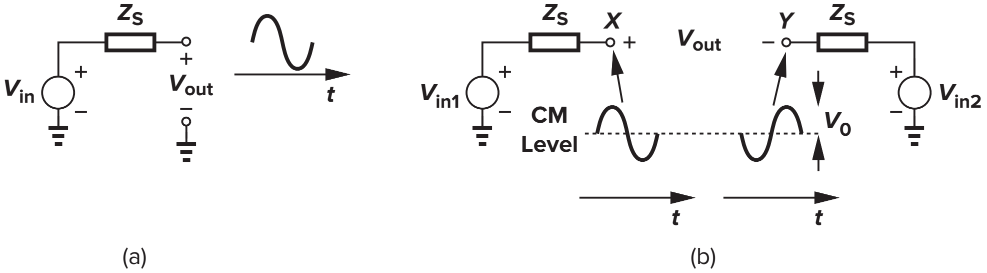


Figure 4.1 (a) Single-ended and (b) differential signals.

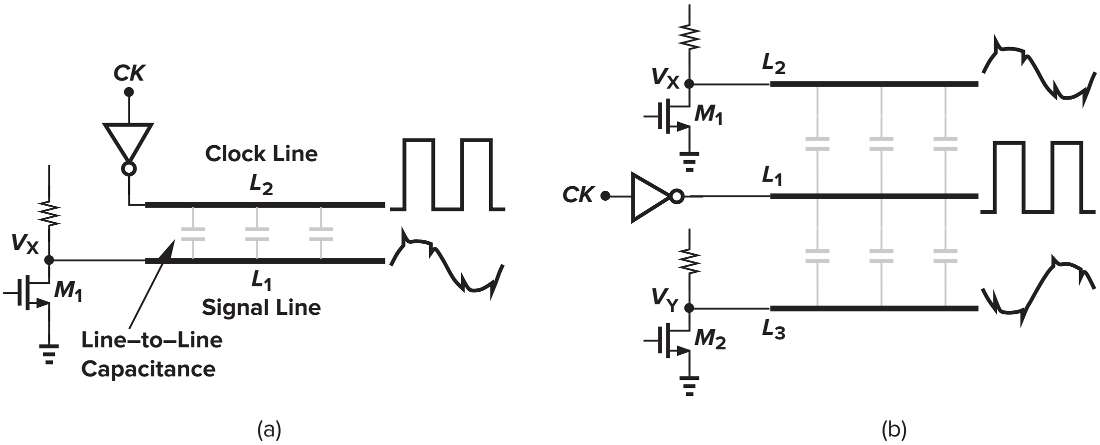


Figure 4.2 (a) Corruption of a signal due to coupling; (b) reduction of coupling by differential operation.

Another example of common-mode rejection occurs with noisy supply voltages. In the CS stage of Fig. 4.3(a), if V_{DD} varies by ΔV, then V_{out} changes by approximately the same amount, i.e., the output is quite susceptible to noise on V_{DD}. Now consider the circuit in Fig. 4.3(b). Here, if the circuit is symmetric, noise on V_{DD} affects V_X and V_Y, but not $V_X - V_Y = V_{out}$. Thus, the circuit of Fig. 4.3(b) is much more robust in dealing with supply noise.

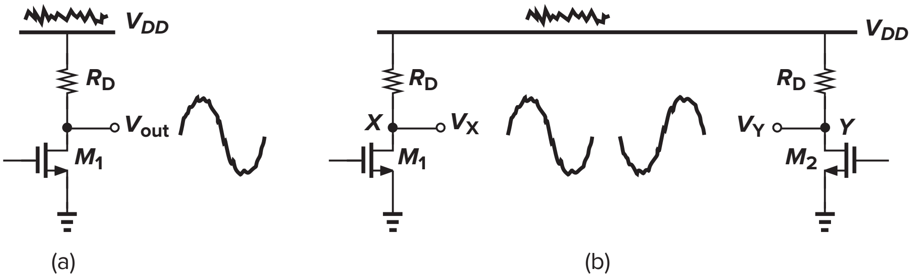


Figure 4.3 Effect of supply noise on (a) a single-ended circuit and (b) a differential circuit.

Thus far, we have seen the importance of employing differential paths for sensitive signals ("victims"). It is also beneficial to employ differential distribution for *noisy lines* ("aggressors"). For example, suppose the clock signal of Fig. 4.2 is distributed in differential form on two lines (Fig. 4.4). Then, with perfect symmetry, the components coupled from CK and $\overline{CK}$ to the signal line cancel each other.

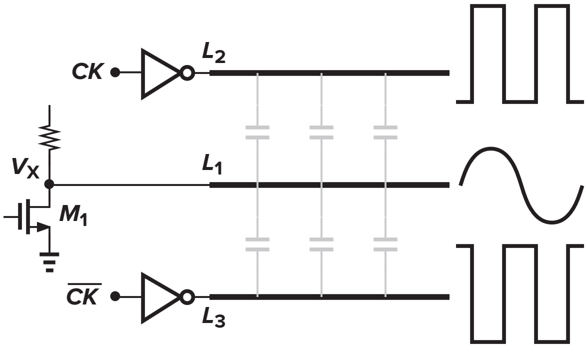


Figure 4.4 Reduction of coupled noise by differential operation.

▶ **Example 4.1**

If differential victims or differential aggressors improve the overall noise immunity, can we choose differential phases for *both* victims and aggressors?

Solution

Yes, we can. Let us consider the arrangement shown in Fig. 4.5(a), where the differential victims are surrounded by the differential aggressors. Unfortunately, in this case, $V_{out}^{+} - V_{out}^{-}$ is corrupted because V_{out}^{+} and V_{out}^{-} experience *opposite* jumps.

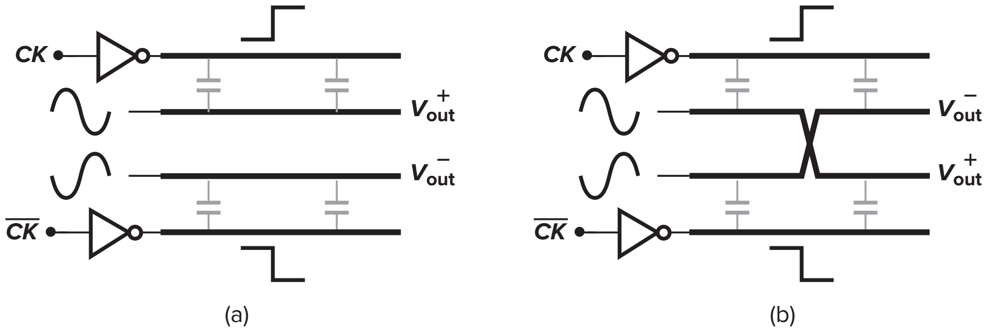


Figure 4.5

Now, suppose we modify the routing as depicted in Fig. 4.5(b), where V_{out}^{+} (V_{out}^{-}) is adjacent to CK ($\overline{CK}$) for half of the distance and to $\overline{CK}$ (CK) for the other half. In this case, the couplings from CK and $\overline{CK}$ cancel each other. Interestingly, V_{out}^{+} and V_{out}^{-} are free from the coupling—and so is their difference. This geometry is an example of "twisted pairs."

◀

Another useful property of differential signaling is the increase in maximum achievable voltage swings. In the circuit of Fig. 4.3, for example, the maximum output swing at X or Y is equal to $V_{DD}-(V_{GS}-V_{TH})$, whereas for $V_X - V_Y$, the peak-to-peak swing is equal to $2[V_{DD} - (V_{GS} - V_{TH})]$. Other advantages of differential circuits over their single-ended counterparts include simpler biasing and higher linearity (Chapter 14).

While differential circuits may occupy about twice as much area as single-ended alternatives, in practice this is a minor drawback. The numerous advantages of differential operation by far outweigh the possible increase in the area.

4.2 ■ Basic Differential Pair

How do we amplify a differential signal? As suggested by the observations in the previous section, we may incorporate two identical single-ended signal paths to process the two phases [Fig. 4.6(a)]. Here, two differential inputs, V_{in1} and V_{in2}, having a certain CM level, $V_{in,CM}$, are applied to the gates. The outputs are also differential and swing around the output CM level, $V_{out,CM}$. Such a circuit indeed offers some of the advantages of differential signaling: high rejection of supply noise, higher output swings, etc. But what happens if V_{in1} and V_{in2} experience a large common-mode disturbance or simply do not have a well-defined common-mode dc level? As the input CM level, $V_{in,CM}$, changes, so do the bias currents of M_1 and M_2, thus varying both the transconductance of the devices and the output CM level. The variation of the transconductance, in turn, leads to a change in the small-signal gain, while the departure of the output CM level from its ideal value lowers the maximum allowable output swings. For example, as shown in Fig. 4.6(b), if the input CM level is excessively low, the minimum values of V_{in1} and V_{in2} may in fact turn off M_1 and M_2, leading to severe clipping at the output. Thus, it is important that the bias currents of the devices have minimal dependence on the input CM level.

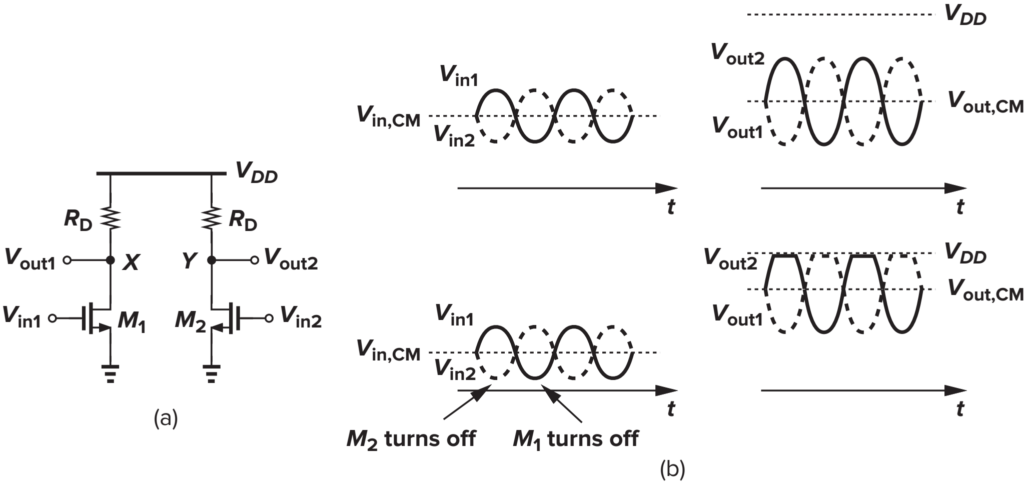


Figure 4.6 (a) Simple differential circuit; (b) illustration of sensitivity to the input common-mode level.

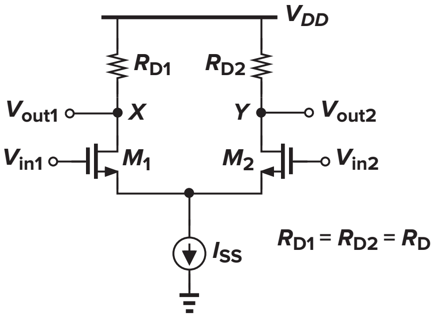


Figure 4.7 Basic differential pair.

A simple modification can resolve the above issue. Shown in Fig. 4.7, the "differential pair"[2] employs a current source I_{SS} to make $I_{D1} + I_{D2}$ independent of $V_{in,CM}$. Thus, if $V_{in1} = V_{in2}$, the bias current of

[2] Also called a "source-coupled" pair or (in the British literature) a "long-tailed" pair.

each transistor equals $I_{SS}/2$ and the output common-mode level is $V_{DD} - R_D I_{SS}/2$. It is instructive to study the large-signal behavior of the circuit for both differential and common-mode input variations. In the large-signal study, we neglect channel-length modulation and body effect.

4.2.1 Qualitative Analysis

Let us assume that in Fig. 4.7, $V_{in1} - V_{in2}$ varies from $-\infty$ to $+\infty$. If V_{in1} is much more negative than V_{in2}, M_1 is off, M_2 is on, and $I_{D2} = I_{SS}$. Thus, $V_{out1} = V_{DD}$ and $V_{out2} = V_{DD} - R_D I_{SS}$. As V_{in1} is brought closer to V_{in2}, M_1 gradually turns on, drawing a fraction of I_{SS} from R_{D1} and hence lowering V_{out1}. Since $I_{D1} + I_{D2} = I_{SS}$, the drain current of M_2 falls and V_{out2} rises. As shown in Fig. 4.8(a), for $V_{in1} = V_{in2}$, we have $V_{out1} = V_{out2} = V_{DD} - R_D I_{SS}/2$, which is the output CM level. As V_{in1} becomes more positive than V_{in2}, M_1 carries a greater current than does M_2 and V_{out1} drops below V_{out2}. For sufficiently large $V_{in1} - V_{in2}$, M_1 "hogs" all of I_{SS}, turning M_2 off. As a result, $V_{out1} = V_{DD} - R_D I_{SS}$ and $V_{out2} = V_{DD}$. Figure 4.8 also plots $V_{out1} - V_{out2}$ versus $V_{in1} - V_{in2}$. Note that the circuit contains three differential quantities: $V_{in1} - V_{in2}$, $V_{out1} - V_{out2}$, and $I_{D1} - I_{D2}$.

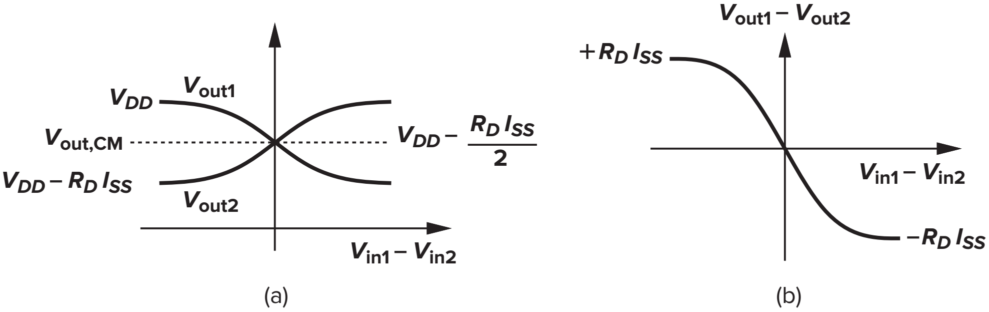


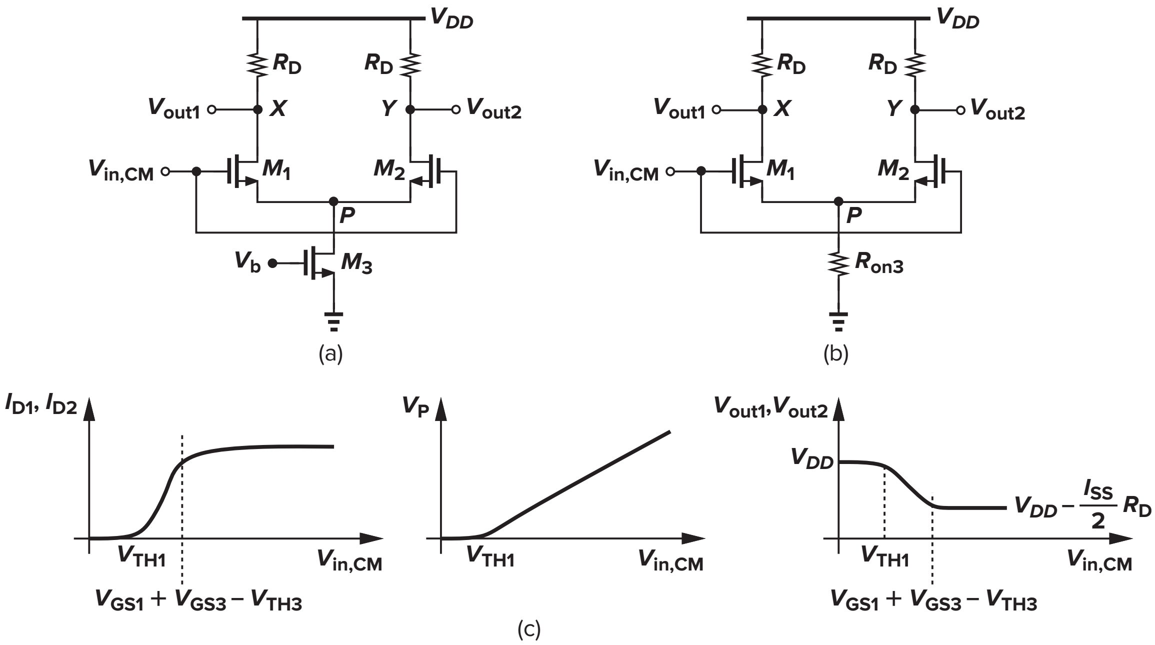
Figure 4.8 Differential input-output characteristics of a differential pair.

The foregoing analysis reveals two important attributes of the differential pair. First, the maximum and minimum levels at the output are well-defined (V_{DD} and $V_{DD} - R_D I_{SS}$, respectively) and independent of the input CM level. Second, as proved later, the small-signal gain (the slope of $V_{out1} - V_{out2}$ versus $V_{in1} - V_{in2}$) is maximum for $V_{in1} = V_{in2}$, gradually falling to zero as $|V_{in1} - V_{in2}|$ increases. In other words, the circuit becomes more nonlinear as the input voltage swing increases. For $V_{in1} = V_{in2}$, we say that the circuit is in "equilibrium."

Now let us consider the common-mode behavior of the circuit. As mentioned earlier, the role of the tail current source is to suppress the effect of input CM level variations on the operation of M_1 and M_2 and the output level. Does this mean that $V_{in,CM}$ can assume arbitrarily low or high values? To answer this question, we set $V_{in1} = V_{in2} = V_{in,CM}$ and vary $V_{in,CM}$ from 0 to V_{DD}. Figure 4.9(a) shows the circuit with I_{SS} implemented by an NFET. Note that the symmetry of the pair requires that $V_{out1} = V_{out2}$.

What happens if $V_{in,CM} = 0$? Since the gate potential of M_1 and M_2 is not more positive than their source potential, both devices are off, yielding $I_{D3} = 0$. This indicates that M_3 operates in the deep triode region because V_b is high enough to create an inversion layer in the transistor. With $I_{D1} = I_{D2} = 0$, the circuit is incapable of signal amplification, $V_{out1} = V_{out2} = V_{DD}$, and $V_P = 0$.

Now suppose $V_{in,CM}$ becomes more positive. Modeling M_3 by a resistor as in Fig. 4.9(b), we note that M_1 and M_2 turn on if $V_{in,CM} \geq V_{TH}$. Beyond this point, I_{D1} and I_{D2} continue to increase, and V_P also rises [Fig. 4.9(c)]. In a sense, M_1 and M_2 constitute a source follower, forcing V_P to track $V_{in,CM}$. For a sufficiently high $V_{in,CM}$, the drain-source voltage of M_3 exceeds $V_{GS3} - V_{TH3}$, allowing the device to operate in saturation. The total current through M_1 and M_2 then remains constant. We conclude that for proper operation, $V_{in,CM} \geq V_{GS1} + (V_{GS3} - V_{TH3})$.



(a)

(b)

(c)

Figure 4.9 (a) Differential pair sensing an input common-mode change; (b) equivalent circuit if M_3 operates in the deep triode region; (c) common-mode input-output characteristics.

What happens if $V_{in,CM}$ rises further? Since V_{out1} and V_{out2} are relatively constant, we expect that M_1 and M_2 enter the triode region if $V_{in,CM} > V_{out1} + V_{TH} = V_{DD} - R_D I_{SS}/2 + V_{TH}$. This sets an upper limit on the input CM level. In summary, the allowable value of $V_{in,CM}$ is bounded as follows:

$$V_{GS1} + (V_{GS3} - V_{TH3}) \leq V_{in,CM} \leq \min\left[V_{DD} - R_D\frac{I_{SS}}{2} + V_{TH},\ V_{DD}\right] \tag{4.1}$$

Beyond the upper bound, the CM characteristics of Fig. 4.9(c) do not change, but the differential gain drops.[3]

▶ Example 4.2

Sketch the small-signal differential gain of a differential pair as a function of the input CM level.

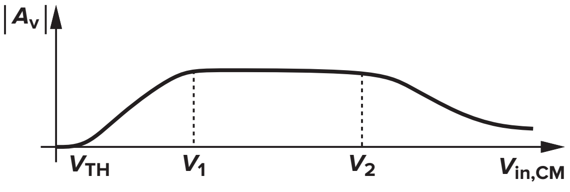


Figure 4.10

Solution

As shown in Fig. 4.10, the gain begins to increase as $V_{in,CM}$ exceeds V_{TH}. After the tail current source enters saturation ($V_{in,CM} = V_1$), the gain remains relatively constant. Finally, if $V_{in,CM}$ is so high that the input transistors enter the triode region ($V_{in,CM} = V_2$), the gain begins to fall.

[3] This bound assumes small differential swings at the input and the output. This point become clear later.

With our understanding of differential and common-mode behavior of the differential pair, we can now answer another important question: How large can the output voltage swings of a differential pair be? Suppose the circuit is biased with input and output bias levels $V_{in,CM}$ and $V_{out,CM}$, respectively, and $V_{in,CM} < V_{out,CM}$. Also, assume that the voltage gain is high, that is, the input swing is much less than the output swing. As illustrated in Fig. 4.11, for M_1 and M_2 to be saturated, each output can go as high as V_{DD} but as low as approximately $V_{in,CM} - V_{TH}$. In other words, the higher the input CM level, the smaller the allowable output swings. For this reason, it is desirable to choose a relatively low $V_{in,CM}$, but, of course, no less than $V_{GS1} + (V_{GS3} - V_{TH3})$. Such a choice affords a single-ended peak-to-peak output swing of $V_{DD} - (V_{GS1} - V_{TH1}) - (V_{GS3} - V_{TH3})$ (why?). The reader is encouraged to repeat this analysis if the voltage gain is around unity.

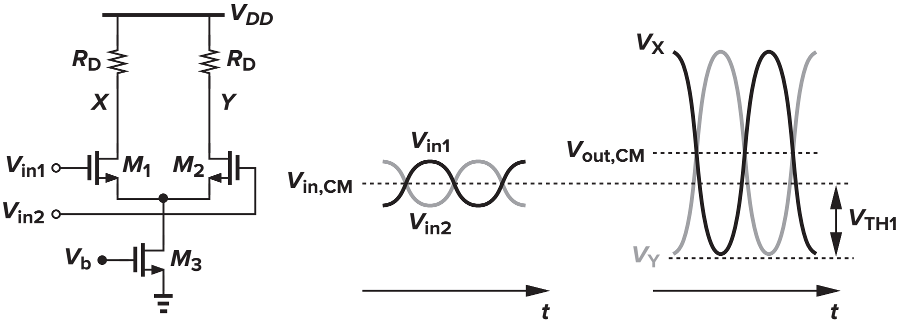


Figure 4.11 Maximum allowable output swings in a differential pair.

▶ **Example 4.3**

Compare the maximum output voltage swings provided by a CS stage and a differential pair.

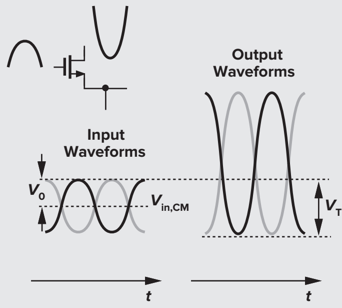
Solution

Recall from Chapter 3 that a CS stage (with resistive load) allows an output swing of V_{DD} minus one overdrive ($V_{DD} - V_{D,sat}$). As seen above, with proper choice of the input CM level, a differential pair provides a maximum output swing of V_{DD} minus two overdrives (single-ended) or $2V_{DD}$ minus four overdrives (differential) ($2V_{DD} - 4V_{D,sat}$), which is typically quite a lot larger than $V_{DD} - V_{D,sat}$.

◀

Nanometer Design Notes

Owing to both severe channel-length modulation and limited supply voltages, the voltage gain of nanometer differential pairs hardly exceeds 5. In this case, the peak *input* swing also limits the output swing. As shown below, for a peak input amplitude of V_0, the minimum allowable output is equal to $V_{in,CM} + V_0 - V_{TH}$. This issue arises in any circuit that has a *negative* gain.



4.2.2 Quantitative Analysis

In this section, we quantify both large-signal and small-signal characteristics of MOS differential pairs. We begin with large-signal analysis to arrive at expressions for the plots shown in Fig. 4.8.

Large-Signal Behavior Consider the differential pair shown in Fig. 4.12. Our objective is to determine $V_{out1} - V_{out2}$ as a function of $V_{in1} - V_{in2}$. We have $V_{out1} = V_{DD} - R_{D1}I_{D1}$ and $V_{out2} = V_{DD} - R_{D2}I_{D2}$, that is, $V_{out1} - V_{out2} = R_{D2}I_{D2} - R_{D1}I_{D1} = R_D(I_{D2} - I_{D1})$ if $R_{D1} = R_{D2} = R_D$. Thus, we simply calculate I_{D1} and I_{D2} in terms of V_{in1} and V_{in2}, assuming the circuit is symmetric, M_1 and M_2 are saturated, and $\lambda = 0$. Since the voltage at node P is equal to $V_{in1} - V_{GS1}$ and $V_{in2} - V_{GS2}$,

$$V_{in1} - V_{in2} = V_{GS1} - V_{GS2} \tag{4.2}$$

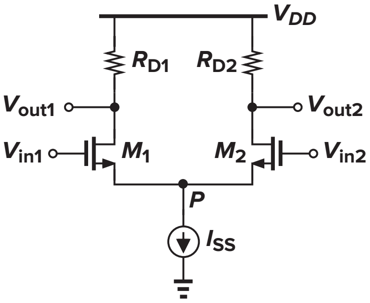


Figure 4.12 Differential pair.

For a square-law device, we have

$$(V_{GS} - V_{TH})^2 = \frac{I_D}{\frac{1}{2}\mu_n C_{ox}\frac{W}{L}} \tag{4.3}$$

and, therefore,

$$V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox}\frac{W}{L}}} + V_{TH} \tag{4.4}$$

It follows from (4.2) and (4.4) that

$$V_{in1} - V_{in2} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\frac{W}{L}}} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\frac{W}{L}}} \tag{4.5}$$

We wish to calculate the differential output current, $I_{D1} - I_{D2}$. Squaring the two sides of (4.5) and recognizing that $I_{D1} + I_{D2} = I_{SS}$, we obtain

$$(V_{in1} - V_{in2})^2 = \frac{2}{\mu_n C_{ox}\frac{W}{L}}(I_{SS} - 2\sqrt{I_{D1}I_{D2}}) \tag{4.6}$$

That is,

$$\frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in1} - V_{in2})^2 - I_{SS} = -2\sqrt{I_{D1}I_{D2}} \tag{4.7}$$

Squaring the two sides again and noting that $4I_{D1}I_{D2} = (I_{D1}+I_{D2})^2-(I_{D1}-I_{D2})^2 = I_{SS}^2-(I_{D1}-I_{D2})^2$, we arrive at

$$(I_{D1} - I_{D2})^2 = -\frac{1}{4}\left(\mu_n C_{ox}\frac{W}{L}\right)^2 (V_{in1} - V_{in2})^4 + I_{SS}\mu_n C_{ox}\frac{W}{L}(V_{in1} - V_{in2})^2 \tag{4.8}$$

Thus,

$$I_{D1} - I_{D2} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in1} - V_{in2})\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - (V_{in1} - V_{in2})^2} \tag{4.9}$$

$$= \sqrt{\mu_n C_{ox}\frac{W}{L}I_{SS}}(V_{in1} - V_{in2})\sqrt{1 - \frac{\mu_n C_{ox}(W/L)}{4I_{SS}}(V_{in1} - V_{in2})^2} \tag{4.10}$$

We can say that M_1, M_2, and the tail operate as a voltage-dependent current source producing $I_{D1} - I_{D2}$ according to the above large-signal characteristics. As expected, $I_{D1}-I_{D2}$ is an odd function of $V_{in1}-V_{in2}$, falling to zero for $V_{in1} = V_{in2}$. As $|V_{in1} - V_{in2}|$ increases from zero, $|I_{D1} - I_{D2}|$ increases because the factor preceding the square root rises more rapidly than the argument in the square root drops.[4]

Before examining (4.9) further, it is instructive to calculate the slope of the characteristic, i.e., the equivalent G_m of M_1 and M_2. Denoting the differential quantities $I_{D1} - I_{D2}$ and $V_{in1} - V_{in2}$ by ΔI_D and ΔV_{in}, respectively, the reader can show that

$$\frac{\partial \Delta I_D}{\partial \Delta V_{in}} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\frac{\dfrac{4I_{SS}}{\mu_n C_{ox} W/L} - 2\Delta V_{in}^2}{\sqrt{\dfrac{4I_{SS}}{\mu_n C_{ox} W/L} - \Delta V_{in}^2}} \tag{4.11}$$

For $\Delta V_{in} = 0$, G_m is maximum (why?) and equal to $\sqrt{\mu_n C_{ox}(W/L)I_{SS}}$. Moreover, since $V_{out1} - V_{out2} = R_D \Delta I = -R_D G_m \Delta V_{in}$, we can write the small-signal differential voltage gain of the circuit in the equilibrium condition as

$$|A_v| = \sqrt{\mu_n C_{ox}\frac{W}{L} I_{SS}} R_D \tag{4.12}$$

Since each transistor carries a bias current of $I_{SS}/2$ in this condition, the factor $\sqrt{\mu_m C_{ox}(W/L)I_{SS}}$ is in fact the same as the transconductance of each device, that is, $|A_v| = g_m R_D$. Equation (4.11) also suggests that G_m falls to zero for $\Delta V_{in} = \sqrt{2I_{SS}/(\mu_n C_{ox} W/L)}$. As we will see below, this value of ΔV_{in} plays an important role in the operation of the circuit.

Let us now examine Eq. (4.9) more closely. If $(V_{in1} - V_{in2})^2 \ll 4I_{SS}/[\mu_n C_{ox}(W/L)]$, then

$$I_{D1} - I_{D2} = \sqrt{\mu_n C_{ox}\frac{W}{L} I_{SS}}(V_{in1} - V_{in2}) \tag{4.13}$$

which yields the same equilibrium G_m as that obtained above.

But what happens for larger values of $|V_{in1} - V_{in2}|$? It appears that the argument in the square root drops to zero for $\Delta V_{in} = \sqrt{4I_{SS}/(\mu_n C_{ox} W/L)}$ and ΔI_D crosses zero at *two* different values of ΔV_{in}, an effect not predicted by our qualitative analysis in Fig. 4.8. This conclusion, however, is incorrect. To understand why, recall that (4.9) was derived with the assumption that both M_1 and M_2 are on. In reality, as ΔV_{in} exceeds a limit, one transistor carries the entire I_{SS}, turning off the other.[5] Denoting this value by ΔV_{in1}, we have $I_{D1} = I_{SS}$ and $\Delta V_{in1} = V_{GS1} - V_{TH}$ because M_2 is nearly off. It follows that

$$\Delta V_{in1} = \sqrt{\frac{2I_{SS}}{\mu_n C_{ox}\dfrac{W}{L}}} \tag{4.14}$$

For $\Delta V_{in} > \Delta V_{in1}$, M_2 is off and (4.9) and (4.10) do not hold. As mentioned above, G_m falls to zero for $\Delta V_{in} = \Delta V_{in1}$. Figure 4.13 plots the behavior.

▶ **Example 4.4**

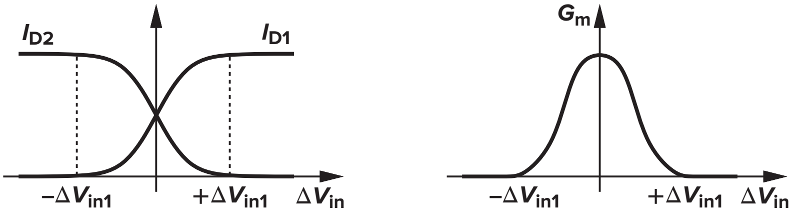
Plot the output currents of a differential pair versus ΔV_{in} as the device width and the tail current vary.

Solution

Consider the characteristic shown in Fig. 4.14(a). As W/L increases, ΔV_{in1} decreases, narrowing the input range across which both devices are on [Fig. 4.14(b)]. As I_{SS} increases, both the input range and the output current swing increase [Fig. 4.14(c)]. Intuitively, we expect the circuit to become more linear as I_{SS} increases or W/L decreases. ◀

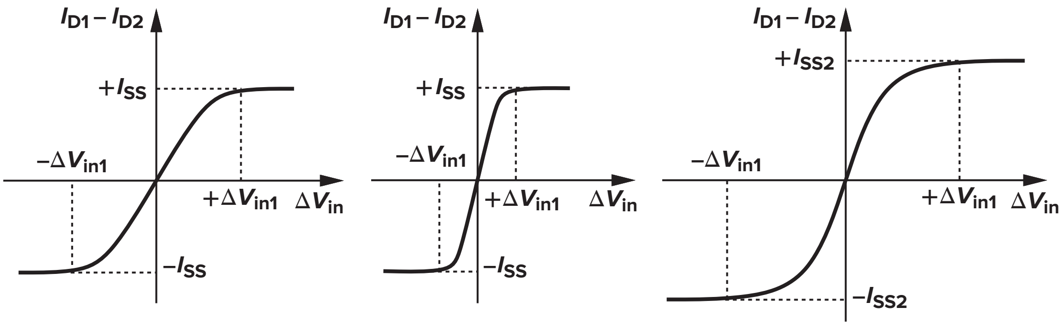
[4] It is interesting to note that, even though I_{D1} and I_{D2} are *square* functions of their respective gate-source voltages, $I_{D1} - I_{D2}$ is an odd function of $V_{in1} - V_{in2}$. This effect is studied in Chapter 14.

[5] We neglect subthreshold conduction here.



(a) (b)

Figure 4.13 Variation of drain currents and overall transconductance of a differential pair versus input voltage.

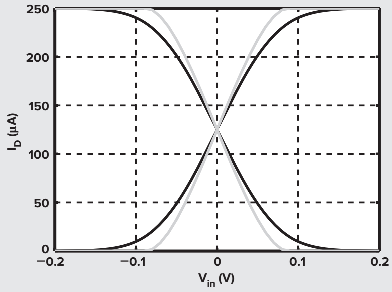


(a) (b) (c)

Figure 4.14

Nanometer Design Notes

Nanometer differential pairs exhibit a similar relation between the equilibrium overdrive and the differential voltage necessary to turn one side off. Plotted below are the output currents of a differential pair with $W/L = 5\ \mu$m/40 nm and $I_{SS} = 0.25$ mA using actual models (black curve) and a square-law model (gray curve). If we define cut-off as when one transistor carries 90% of the tail current, then the nanometer design also displays approximately a factor of $\sqrt{2}$ between the equilibrium and cut-off voltages.



The value of ΔV_{in1} given by (4.14) in essence represents the maximum differential input that the circuit can "handle." It is possible to relate ΔV_{in1} to the overdrive voltage of M_1 and M_2 in equilibrium. For a zero differential input, $I_{D1} = I_{D2} = I_{SS}/2$, yielding

$$(V_{GS} - V_{TH})_{1,2} = \sqrt{\frac{I_{SS}}{\mu_n C_{ox}\dfrac{W}{L}}} \tag{4.15}$$

Thus, ΔV_{in1} is equal to $\sqrt{2}$ times the equilibrium overdrive. The point is that increasing ΔV_{in1} to make the circuit more linear inevitably increases the overdrive voltage of M_1 and M_2. For a given I_{SS}, this is accomplished only by reducing W/L and hence the transconductance of the transistors, trading small-signal gain for linearity. Alternatively, we can increase I_{SS}, but at the cost of power. (What happens to the gain if I_{SS} is increased but $I_{SS}R_D$ is kept constant due to headroom constraints?)

▶ Example 4.5

Due to a manufacturing defect, the differential signals applied to a differential pair have unequal dc levels (Fig. 4.15). If the peak swing, V_0, is small and the imbalance, V_{OS}, happens to be equal to $\Delta V_{in1}/2 = (1/2)\sqrt{2I_{SS}/(\mu_n C_{ox} W/L)}$, sketch the output voltage waveforms and determine the small-signal voltage gain.

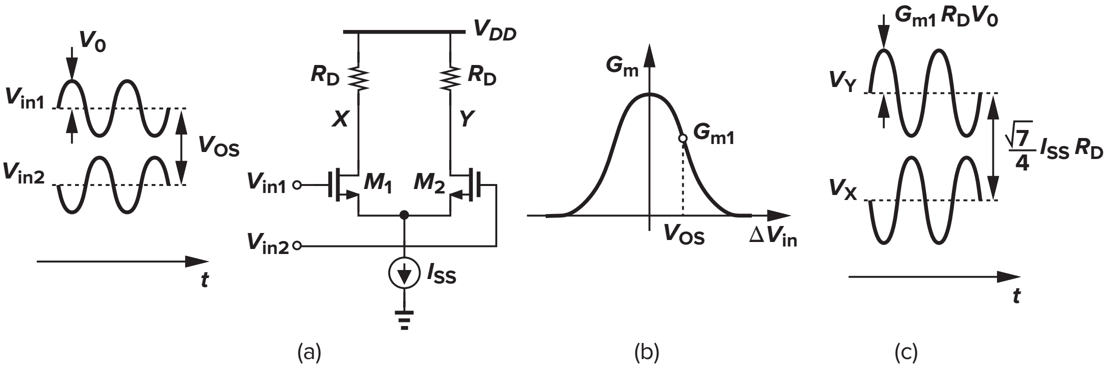


Figure 4.15

Solution

Let us first study the circuit with only dc inputs that differ by V_{OS}. The differential pair senses an imbalance of $V_{in1} - V_{in2} = V_{OS}$ and, from Eq. (4.10), generates

$$I_{D1} - I_{D2} = \frac{\sqrt{7}}{4} I_{SS} \tag{4.16}$$

That is, $I_{D1} \approx 0.83 I_{SS}$, $I_{D2} \approx 0.17 I_{SS}$ and $V_X - V_Y = -(\sqrt{7}/4) I_{SS} R_D$.

Now, we recognize from Fig. 4.15(b) that the input dc imbalance biases the transistors away from the highest transconductance, yielding from Eq. (4.11)

$$G_{m1} = \frac{3}{\sqrt{14}} \sqrt{\mu_n C_{ox} \frac{W}{L} I_{SS}} \tag{4.17}$$

This value is about 20% less than that at equilibrium. The output waveforms are shown in Fig. 4.15(c).

Small-Signal Analysis We now study the small-signal behavior of differential pairs. As depicted in Fig. 4.16, we apply small signals V_{in1} and V_{in2} and assume that M_1 and M_2 are saturated. What is the differential voltage gain, $(V_{out1} - V_{out2})/(V_{in1} - V_{in2})$? Recall from Eq. (4.12) that this quantity equals $\sqrt{\mu_n C_{ox} I_{SS} W/L} R_D$. Since in the vicinity of equilibrium, each transistor carries approximately $I_{SS}/2$, this expression reduces to $g_m R_D$, where g_m denotes the transconductance of M_1 and M_2. To arrive at the same result by small-signal analysis, we employ two different methods, each providing insight into the circuit's operation. We assume that $R_{D1} = R_{D2} = R_D$.

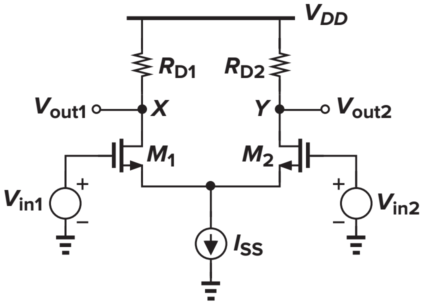


Figure 4.16 Differential pair with small-signal inputs.

Method I The circuit of Fig. 4.16 is driven by two independent signals. Thus, the output can be computed by superposition. (The voltages in this section are small-signal quantities.)

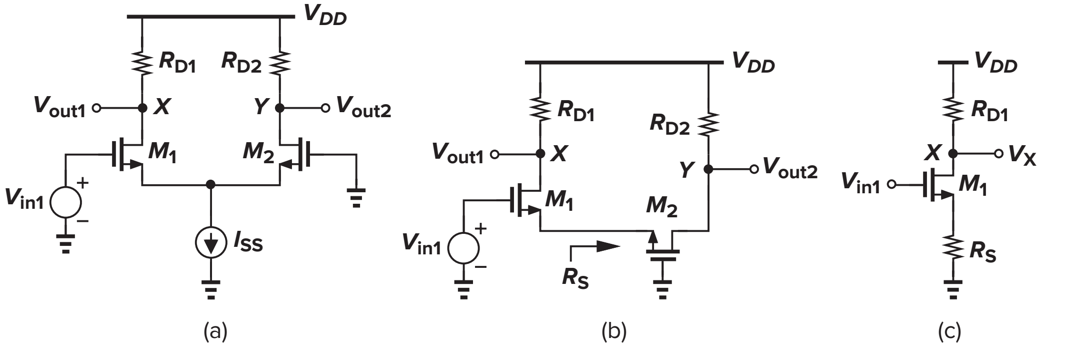


Figure 4.17 (a) Differential pair sensing one input signal; (b) circuit of (a) viewed as a CS stage degenerated by M_2; (c) equivalent circuit of (b).

Let us set V_{in2} to zero and find the effect of V_{in1} at X and Y [Fig. 4.17(a)]. To obtain V_X, we note that M_1 forms a common-source stage with a degeneration resistance equal to the impedance seen looking into the source of M_2 [Fig. 4.17(b)]. Neglecting channel-length modulation and body effect, we have $R_S = 1/g_{m2}$ [Fig. 4.17(c)] and

$$\frac{V_X}{V_{in1}} = \frac{-R_D}{\dfrac{1}{g_{m1}} + \dfrac{1}{g_{m2}}} \tag{4.18}$$

To calculate V_Y, we note that M_1 drives M_2 as a source follower and replace V_{in1} and M_1 by a Thevenin equivalent (Fig. 4.18): the Thevenin voltage $V_T = V_{in1}$ and the resistance $R_T = 1/g_{m1}$. Here, M_2 operates as a common-gate stage, exhibiting a gain equal to

$$\frac{V_Y}{V_{in1}} = \frac{R_D}{\dfrac{1}{g_{m2}} + \dfrac{1}{g_{m1}}} \tag{4.19}$$

It follows from (4.18) and (4.19) that the overall voltage gain for V_{in1} is

$$(V_X - V_Y)|_{\text{Due to } Vin1} = \frac{-2R_D}{\dfrac{1}{g_{m1}} + \dfrac{1}{g_{m2}}} V_{in1} \tag{4.20}$$

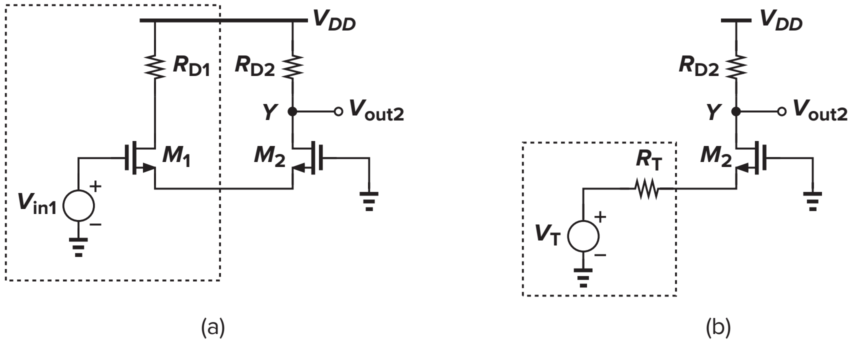


Figure 4.18 Replacing M_1 by a Thevenin equivalent.

which, for $g_{m1} = g_{m2} = g_m$, reduces to

$$(V_X - V_Y)|_{\text{Due to } V_{in1}} = -g_m R_D V_{in1} \tag{4.21}$$

By virtue of symmetry, the effect of V_{in2} at X and Y is identical to that of V_{in1} except for a change in the polarities:

$$(V_X - V_Y)|_{\text{Due to } V_{in2}} = g_m R_D V_{in2} \tag{4.22}$$

Adding the two sides of (4.21) and (4.22) to perform superposition, we have

$$\frac{(V_X - V_Y)_{tot}}{V_{in1} - V_{in2}} = -g_m R_D \tag{4.23}$$

Comparison of (4.21), (4.22), and (4.23) indicates that the magnitude of the differential gain is equal to $g_m R_D$ regardless of how the inputs are applied: in Figs. 4.17 and 4.18, the input is applied to only one side, whereas in Fig. 4.16 the input is the *difference* between two sources. It is also important to recognize that if the output is single-ended, i.e., it is sensed between X or Y and ground, the gain is halved.

▶ Example 4.6

Due to a manufacturing error, in the circuit of Fig. 4.19, M_2 is twice as wide as M_1. Calculate the small-signal gain if the dc levels of V_{in1} and V_{in2} are equal.

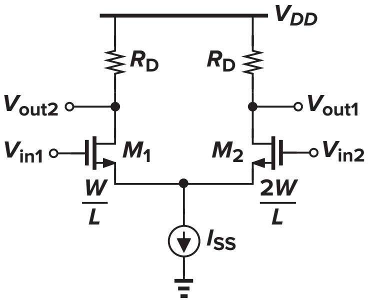


Figure 4.19

Solution

If the gates of M_1 and M_2 are at the same dc potential, then $V_{GS1} = V_{GS2}$ and $I_{D2} = 2I_{D1} = 2I_{SS}/3$. Thus, $g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)I_{SS}/3}$ and $g_{m2} = \sqrt{2\mu_n C_{ox}(2W/L)(2I_{SS})/3} = 2g_{m1}$. Following the same procedure as above, the reader can show that

$$|A_v| = \frac{2R_D}{\dfrac{1}{g_{m1}} + \dfrac{1}{2g_{m1}}} \tag{4.24}$$

$$= \frac{4}{3} g_{m1} R_D \tag{4.25}$$

Note that, for a given I_{SS}, this value is lower than the gain of a symmetric differential pair [Eq. (4.23)] because g_{m1} is smaller. The reader can show that the characteristics of Fig. 4.13 are shifted horizontally, and hence the circuit exhibits an "offset." We utilize this idea in Chapter 14 to linearize differential pairs. ◀

How does the transconductance of a differential pair compare with that of a common-source stage? For a given *total* bias current, the value of g_m in (4.23) is $1/\sqrt{2}$ times that of a single transistor biased at I_{SS} with the same dimensions. Thus, the total G_m is proportionally less.

Method II If a fully-symmetric differential pair senses differential inputs (i.e., the two inputs change by equal and opposite amounts from the equilibrium condition), then the concept of "half circuit" can be applied. We first prove a lemma.

Lemma Consider the symmetric circuit shown in Fig. 4.20(a), where D_1 and D_2 represent any three-terminal active device. Suppose V_{in1} and V_{in2} change differentially, the former from V_0 to $V_0 + \Delta V_{in}$ and the latter from V_0 to $V_0 - \Delta V_{in}$ [Fig. 4.20(b)]. Then, if the circuit remains linear, V_P does not change. Assume $\lambda = 0$.

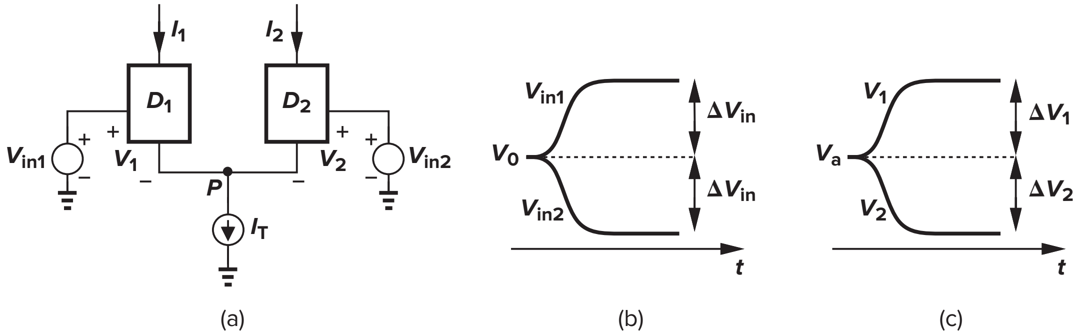


Figure 4.20 Illustration of why node P is a virtual ground.

Proof. The lemma can be proved by invoking symmetry. As long as the operation remains linear, so that the difference between the bias currents of D_1 and D_2 is negligible, the circuit is symmetric. Thus, V_P cannot "favor" the change at one input and "ignore" the other.

From another point of view, the effect of D_1 and D_2 at node P can be represented by Thevenin equivalents (Fig. 4.21). If V_{T1} and V_{T2} change by equal and opposite amounts and R_{T1} and R_{T2} are equal, then V_P remains constant. We emphasize that this is valid if the changes are small enough that we can assume $R_{T1} = R_{T2}$ (e.g., $1/g_{m1} = 1/g_{m2}$).[6] This perspective suggests the lemma's validity even if the tail current source is not ideal. ❑

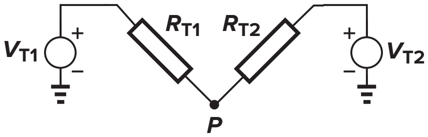


Figure 4.21 Replacing each half of a differential pair by a Thevenin equivalent.

We now offer a more formal proof. Let us assume that V_1 and V_2 have an equilibrium value of V_a and change by ΔV_1 and ΔV_2, respectively [Fig. 4.20(c)]. The output currents therefore change by $g_m \Delta V_1$ and $g_m \Delta V_2$. Since $I_1 + I_2 = I_T$, we have $g_m \Delta V_1 + g_m \Delta V_2 = 0$, i.e., $\Delta V_1 = -\Delta V_2$. We also know that $V_{in1} - V_1 = V_{in2} - V_2$, and hence $V_0 + \Delta V_{in} - (V_a + \Delta V_1) = V_0 - \Delta V_{in} - (V_a + \Delta V_2)$. Consequently, $2\Delta V_{in} = \Delta V_1 - \Delta V_2 = 2\Delta V_1$. In other words, if V_{in1} and V_{in2} change by $+\Delta V_{in}$ and $-\Delta V_{in}$, respectively, then V_1 and V_2 change by the same values, i.e., a differential change in the inputs is simply "absorbed" by V_1 and V_2. In fact, since $V_P = V_{in1} - V_1$, and since V_1 exhibits the same change as V_{in1}, V_P does not change.

[6] It is also possible to derive an expression for the large-signal behavior of V_P and prove that for small $V_{in1} - V_{in2}$, V_P remains constant. We defer this calculation to Chapter 15.

The above lemma greatly simplifies the small-signal analysis of differential amplifiers. As shown in Fig. 4.22, since V_P experiences no change, node P can be considered "ac ground" (or a "virtual ground"), and the circuit can be decomposed into two separate halves. We say that we have applied the "half-circuit concept" [1]. We can write $V_X/V_{in1} = -g_m R_D$ and $V_Y/(-V_{in1}) = -g_m R_D$, where V_{in1} and $-V_{in1}$ denote the voltage *change* on each side. Thus, $(V_X - V_Y)/(2V_{in1}) = -g_m R_D$.

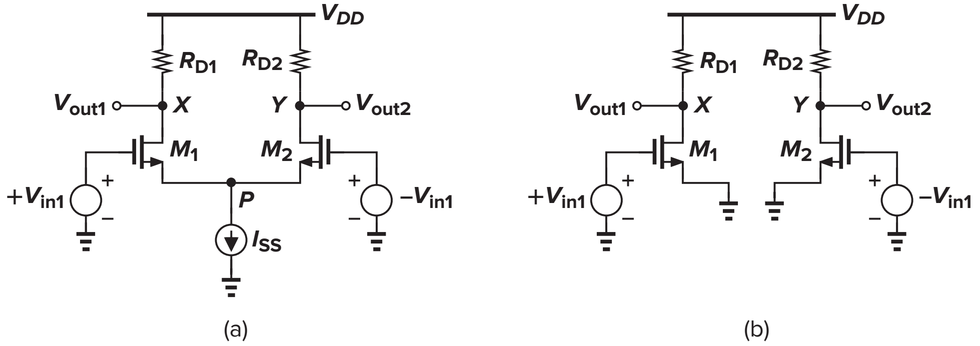


Figure 4.22 Application of the half-circuit concept.

▶ **Example 4.7**

Calculate the differential gain of the circuit of Fig. 4.22(a) if $\lambda \neq 0$.

Solution

Applying the half-circuit concept as illustrated in Fig. 4.23, we have $V_X/V_{in1} = -g_m(R_D \| r_{O1})$ and $V_Y/(-V_{in1}) = -g_m(R_D \| r_{O2})$, thus arriving at $(V_X - V_Y)/(2V_{in1}) = -g_m(R_D \| r_O)$, where $r_O = r_{O1} = r_{O2}$. Note that Method I would require lengthy calculations here.

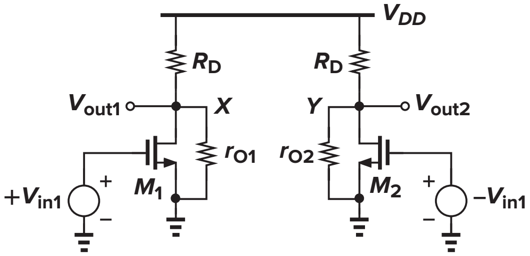


Figure 4.23

The half-circuit concept provides a powerful technique for analyzing symmetric differential pairs with fully differential inputs. But what happens if the two inputs are not fully differential [Fig. 4.24(a)]? As depicted in Figs. 4.24(b) and (c), the two inputs V_{in1} and V_{in2} can be viewed as

$$V_{in1} = \frac{V_{in1} - V_{in2}}{2} + \frac{V_{in1} + V_{in2}}{2} \tag{4.26}$$

$$V_{in2} = \frac{V_{in2} - V_{in1}}{2} + \frac{V_{in1} + V_{in2}}{2} \tag{4.27}$$

Since the second term is common to both inputs, we obtain the equivalent circuit in Fig. 4.24(d), recognizing that the circuit senses a combination of a differential input and a common-mode variation. Therefore, as illustrated in Fig. 4.25, the effect of each type of input can be computed by superposition, with the half-circuit concept applied to the differential-mode operation. We deal with CM analysis in Sec. 4.3.

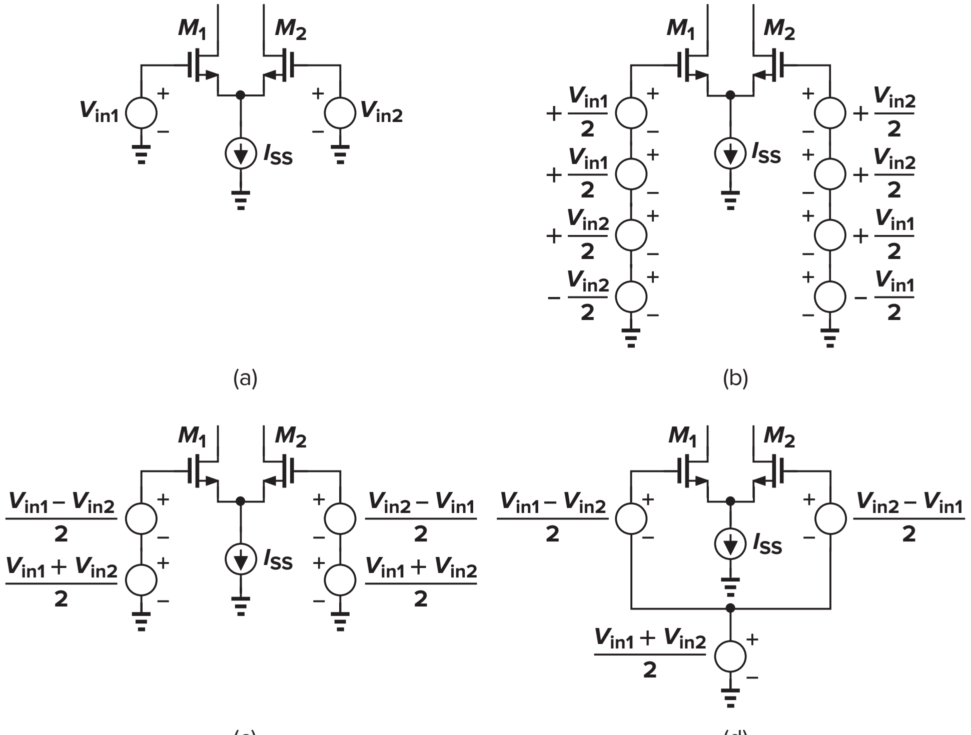


Figure 4.24 Conversion of arbitrary inputs to differential and common-mode components.

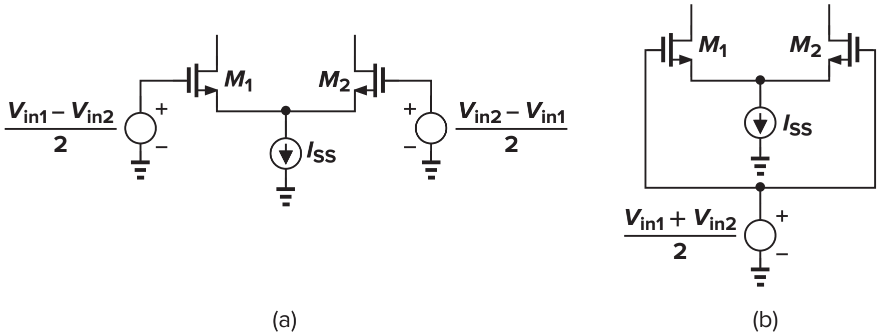

M_1
M_2
$\frac{V_{in1}-V_{in2}}{2}$
$\frac{V_{in2}-V_{in1}}{2}$
I_{SS}
$\frac{V_{in1}+V_{in2}}{2}$
(a)
(b)

Figure 4.25 Superposition for (a) differential and (b) common-mode signals.

▶ Example 4.8

In the circuit of Fig. 4.22(a), calculate V_X and V_Y if $V_{in1} \neq -V_{in2}$ and $\lambda \neq 0$.

Solution

For differential-mode operation, we have from Fig. 4.26(a)

$$V_X = -g_m(R_D \| r_{O1})\frac{V_{in1}-V_{in2}}{2} \tag{4.28}$$

$$V_Y = -g_m(R_D \| r_{O2})\frac{V_{in2}-V_{in1}}{2} \tag{4.29}$$

That is,

$$V_X - V_Y = -g_m(R_D \| r_O)(V_{in1} - V_{in2}) \tag{4.30}$$

which is to be expected.

For common-mode operation, the circuit reduces to that in Fig. 4.26(b). How much do V_X and V_Y change as $V_{in,CM}$ changes? If the circuit is fully symmetric and I_{SS} an ideal current source, the currents drawn by M_1 and M_2 from R_{D1} and R_{D2} are exactly equal to $I_{SS}/2$ and independent of $V_{in,CM}$. Thus, V_X and V_Y remain equal to $V_{DD} - R_D(I_{SS}/2)$ and experience no change as $V_{in,CM}$ varies. Interestingly, the circuit simply amplifies the difference between V_{in1} and V_{in2} while eliminating the effect of $V_{in,CM}$.

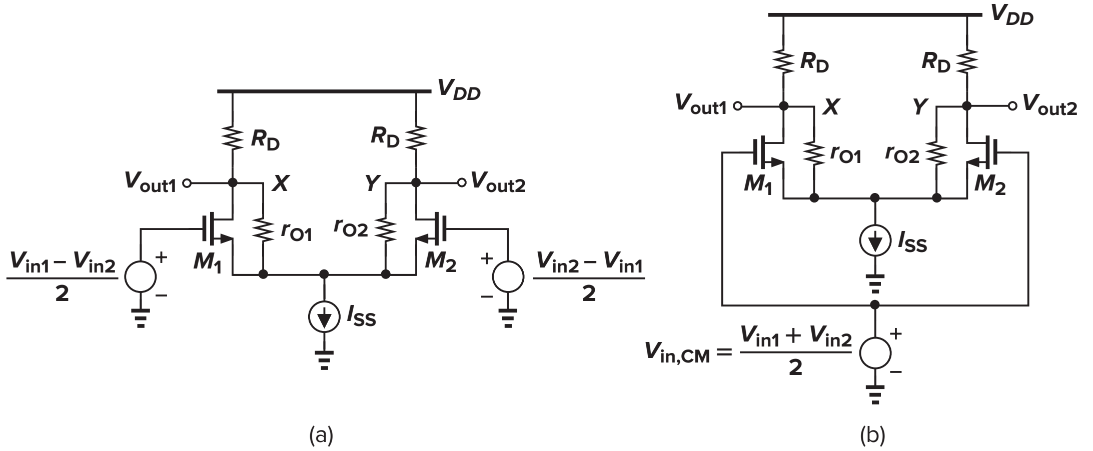


Figure 4.26

4.2.3 Degenerated Differential Pair

As with a simple common-source stage, a differential pair can incorporate resistive degeneration to improve its linearity. Shown in Fig. 4.27(a), such a topology softens the nonlinear behavior of M_1 and M_2 by R_{S1} and R_{S2}. This can be seen from the input-output characteristics of Fig. 4.27(b), where, due to degeneration, the differential voltage necessary to turn off one side increases in magnitude. We can

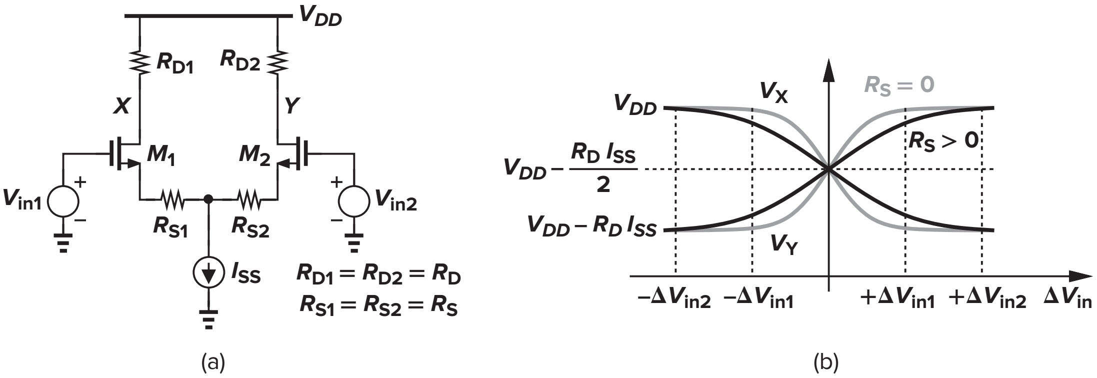


Figure 4.27 (a) Degenerated differential pair, and (b) characteristics with and without degeneration.

readily prove this point. Suppose that at $V_{in1} - V_{in2} = \Delta V_{in2}$, M_2 turns off and $I_{D1} = I_{SS}$. We then have $V_{GS2} = V_{TH}$, and hence

$$V_{in1} - V_{GS1} - R_S I_{SS} = V_{in2} - V_{TH} \tag{4.31}$$

which yields

$$V_{in1} - V_{in2} = V_{GS1} - V_{TH} + R_S I_{SS} \tag{4.32}$$

$$= \sqrt{\frac{2I_{SS}}{\mu_n C_{ox}\dfrac{W}{L}}} + R_S I_{SS} \tag{4.33}$$

We recognize the first term on the right-hand side as ΔV_{in1} (the input difference necessary for turning off M_2 if $R_S = 0$). It follows that

$$\Delta V_{in2} - \Delta V_{in1} = R_S I_{SS} \tag{4.34}$$

suggesting that the linear input range is widened by approximately $\pm R_S I_{SS}$.

The small-signal voltage gain of the degenerated differential pair can be obtained by applying the half-circuit concept. The half circuit is simply a degenerated CS stage, exhibiting a gain of

$$|A_v| = \frac{R_D}{\dfrac{1}{g_m} + R_S} \tag{4.35}$$

if $\lambda = \gamma = 0$. The circuit thus trades gain for linearity—as is also observed from the slopes of the characteristics in Fig. 4.27(b). Note that A_v is less sensitive to g_m variations in this case.

In addition to reducing the gain, the degeneration resistors in Fig. 4.27(a) also consume voltage headroom. In the equilibrium condition, each resistor sustains a voltage drop of $R_S I_{SS}/2$, as if the tail current source itself required this much more headroom. The input common-mode level must therefore be higher by this amount, and so must be the minimum voltage at X or Y. In other words, the maximum allowable *differential* output swing is reduced by $R_S I_{SS}$. This issue can be resolved as shown in Fig. 4.28, where the tail current source is split in half, with each half directly tied to a source. In equilibrium, no current flows through the degeneration resistance, and hence no headroom is sacrificed.[7] Other methods of linearizing differential pairs are described in Chapter 14.

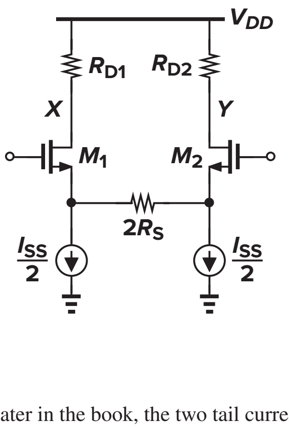


Figure 4.28 Degenerated differential pair with split tail current source.

[7] But, as explained later in the book, the two tail current sources do contribute differential noise and offset in this case.

4.3 ■ Common-Mode Response

An important attribute of differential amplifiers is their ability to suppress the effect of common-mode perturbations. Example 4.8 portrays an idealized case of common-mode response. In reality, neither is the circuit fully symmetric nor does the current source exhibit an infinite output impedance. As a result, a fraction of the input CM variation appears at the output.

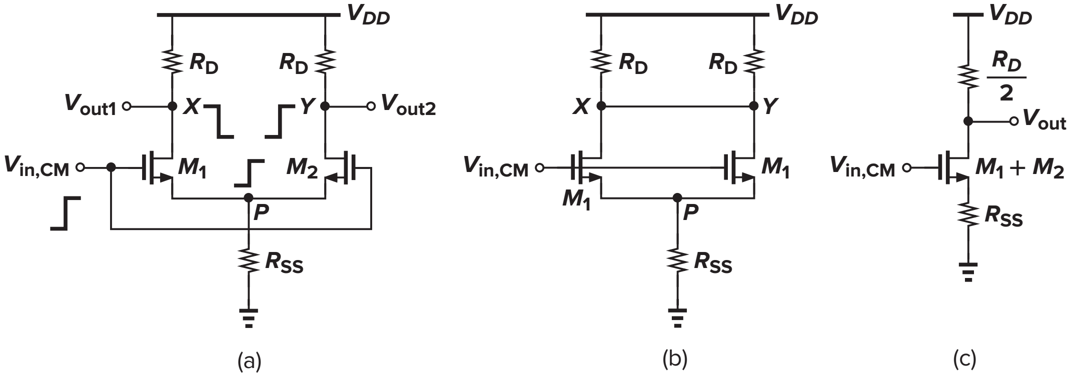


Figure 4.29 (a) Differential pair sensing CM input; (b) simplified version of (a); (c) equivalent circuit of (b).

We first assume that the circuit is symmetric, but the current source has a finite output impedance, R_{SS} [Fig. 4.29(a)]. As $V_{in,CM}$ changes, so does V_P, thereby increasing the drain currents of M_1 and M_2 and lowering both V_X and V_Y. Owing to symmetry, V_X remains equal to V_Y and, as depicted in Fig. 4.29(b), the two nodes can be shorted together. Since M_1 and M_2 are now "in parallel," i.e., they share all of their respective terminals, the circuit can be reduced to that in Fig. 4.29(c). Note that the composite device, $M_1 + M_2$, has twice the width and the bias current of each of M_1 and M_2 and, therefore, twice their transconductance. The "common-mode gain" of the circuit is thus equal to

$$A_{v,CM} = \frac{V_{out}}{V_{in,CM}} \tag{4.36}$$

$$= -\frac{R_D/2}{1/(2g_m) + R_{SS}} \tag{4.37}$$

where g_m denotes the transconductance of each of M_1 and M_2 and $\lambda = \gamma = 0$.

What is the significance of this calculation? In a symmetric circuit, input CM variations disturb the bias points, altering the small-signal gain and possibly limiting the output voltage swings. This can be illustrated by an example.

▶ **Example 4.9**

The circuit of Fig. 4.30 uses a resistor rather than a current source to define a tail current of 1 mA. Assume that $(W/L)_{1,2} = 25/0.5$, $\mu_n C_{ox} = 50\ \mu\text{A/V}^2$, $V_{TH} = 0.6$ V, $\lambda = \gamma = 0$, and $V_{DD} = 3$ V.

(a) What is the required input CM voltage for which R_{SS} sustains 0.5 V?

(b) Calculate R_D for a differential gain of 5.

(c) What happens at the output if the input CM level is 50 mV higher than the value calculated in (a)?

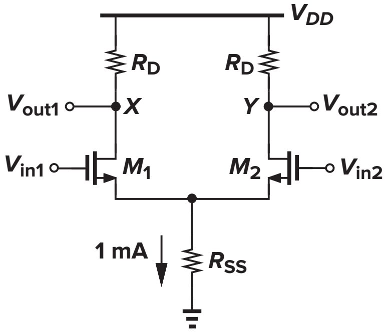


Figure 4.30

Solution

(a) Since $I_{D1} = I_{D2} = 0.5$ mA, we have

$$V_{GS1} = V_{GS2} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\dfrac{W}{L}}} + V_{TH} \tag{4.38}$$

$$= 1.23 \text{ V} \tag{4.39}$$

Thus, $V_{in,CM} = V_{GS1} + 0.5 \text{ V} = 1.73$ V. Note that $R_{SS} = 500\ \Omega$.

(b) The transconductance of each device is $g_m = \sqrt{2\mu_n C_{ox}(W/L)I_{D1}} = 1/(632\ \Omega)$, requiring $R_D = 3.16$ kΩ for a gain of 5.

Note that the output bias level is equal to $V_{DD} - I_{D1}R_D = 1.42$ V. Since $V_{in,CM} = 1.73$ V and $V_{TH} = 0.6$ V, the transistors are 290 mV away from the triode region.

(c) If $V_{in,CM}$ increases by 50 mV, the equivalent circuit of Fig. 4.29(c) suggests that V_X and V_Y drop by

$$|\Delta V_{X,Y}| = \Delta V_{in,CM}\frac{R_D/2}{R_{SS} + 1/(2g_m)} \tag{4.40}$$

$$= 50 \text{ mV} \times 1.94 \tag{4.41}$$

$$= 96.8 \text{ mV} \tag{4.42}$$

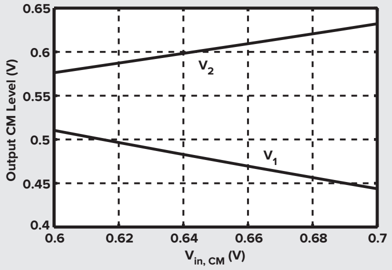
Now, M_1 and M_2 are only 143 mV away from the triode region because the input CM level has increased by 50 mV and the output CM level has decreased by 96.8 mV.

The foregoing discussion indicates that the finite output impedance of the tail current source results in some common-mode gain in a symmetric differential pair. Nonetheless, this is usually a minor concern. More troublesome is the variation of the *differential* output as a result of a change in $V_{in,CM}$, an effect that occurs because in reality the circuit is not fully symmetric, i.e., the two sides suffer from slight mismatches during manufacturing. For example, in Fig. 4.29(a), R_{D1} may not be exactly equal to R_{D2}.

We now study the effect of input common-mode variations if the circuit is asymmetric and the tail current source suffers from a finite output impedance. Suppose, as shown in Fig. 4.31, $R_{D1} = R_D$ and $R_{D2} = R_D + \Delta R_D$, where ΔR_D denotes a small mismatch and the circuit is otherwise symmetric. Assume that $\lambda = \gamma = 0$ for M_1 and M_2. What happens to V_X and V_Y as $V_{in,CM}$ increases? We recognize that M_1 and M_2 operate as one source follower,

Nanometer Design Notes

As a result of the low output impedance of tail current sources in nanometer technologies, a CM level change can "propagate." Plotted below are the output CM levels of two cascaded differential pairs as the main input CM level, $V_{in,CM}$, increases, revealing a drop in the first and a rise in the second.



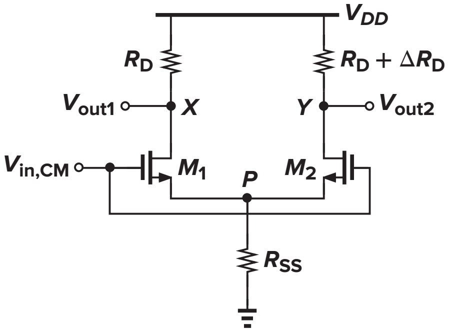


Figure 4.31 Common-mode response in the presence of resistor mismatch.

raising V_P by

$$\Delta V_P = \frac{R_{SS}}{R_{SS} + \dfrac{1}{2g_m}} \Delta V_{in,CM} \tag{4.43}$$

Since M_1 and M_2 are identical, I_{D1} and I_{D2} increase by $[g_m/(1 + 2g_m R_{SS})]\Delta V_{in,CM}$, but V_X and V_Y change by different amounts:

$$\Delta V_X = -\Delta V_{in,CM} \frac{g_m}{1 + 2g_m R_{SS}} R_D \tag{4.44}$$

$$\Delta V_Y = -\Delta V_{in,CM} \frac{g_m}{1 + 2g_m R_{SS}} (R_D + \Delta R_D) \tag{4.45}$$

Thus, a common-mode change at the input introduces a *differential* component at the output. We say that the circuit exhibits common-mode to differential conversion. This is a critical problem because if the input of a differential pair includes both a differential signal and common-mode noise, the circuit corrupts the amplified differential signal by the input CM change. The effect is illustrated in Fig. 4.32.

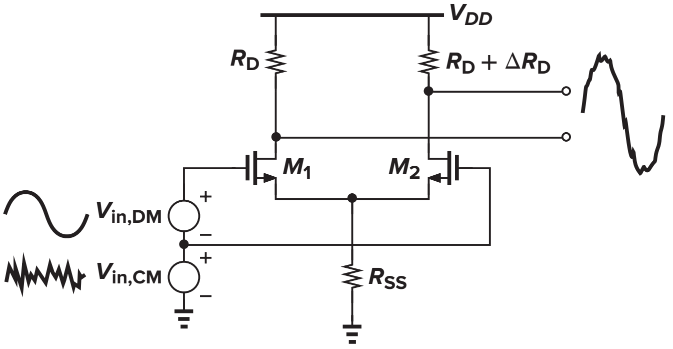


Figure 4.32 Effect of CM noise in the presence of resistor mismatch.

In summary, the common-mode response of differential pairs depends on the output impedance of the tail current source and asymmetries in the circuit, manifesting itself through two effects: variation of the output CM level (in the absence of mismatches) and conversion of input common-mode variations to differential components at the output. In analog circuits, the latter effect is much more severe than the former. For this reason, the common-mode response should usually be studied with mismatches taken into account.

How significant is common-mode to differential conversion? We make two observations. First, as the *frequency* of the CM disturbance increases, the total capacitance shunting the tail current source introduces larger tail current variations. Thus, even if the output *resistance* of the current source is high, common-mode to differential conversion becomes significant at high frequencies. Shown in Fig. 4.33, this capacitance arises from the parasitics of the current source itself as well as the source-bulk junctions of M_1 and M_2. Second, the asymmetry in the circuit stems from both the load resistors and the input transistors, the latter contributing a typically much greater mismatch.

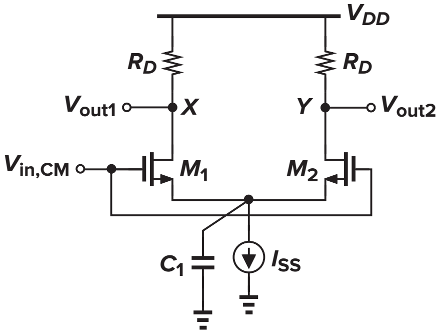


Figure 4.33 CM response with finite tail capacitance.

Let us study the asymmetry resulting from mismatches between M_1 and M_2 in Fig. 4.34(a). Owing to dimension and threshold voltage mismatches, the two transistors carry slightly different currents and exhibit unequal transconductances. We assume that $\lambda = \gamma = 0$. To calculate the small-signal gain from $V_{in,CM}$ to X and Y, we use the equivalent circuit in Fig. 4.34(b), writing $I_{D1} = g_{m1}(V_{in,CM} - V_P)$ and $I_{D2} = g_{m2}(V_{in,CM} - V_P)$. Since $(I_{D1} + I_{D2})R_{SS} = V_P$,

$$(g_{m1} + g_{m2})(V_{in,CM} - V_P)R_{SS} = V_P \tag{4.46}$$

and

$$V_P = \frac{(g_{m1} + g_{m2})R_{SS}}{(g_{m1} + g_{m2})R_{SS} + 1} V_{in,CM} \tag{4.47}$$

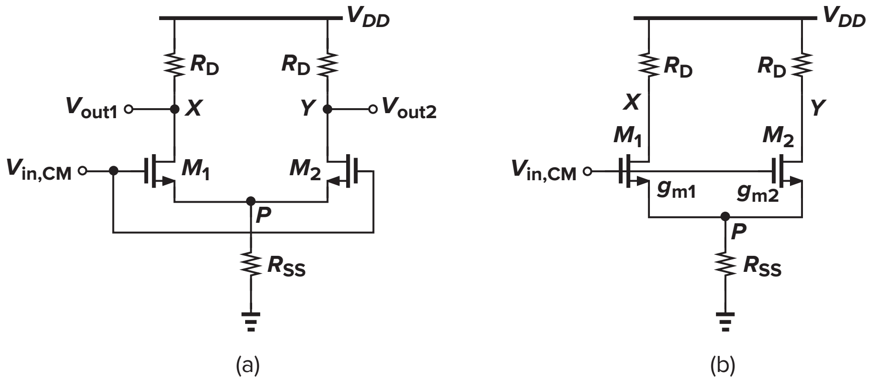


Figure 4.34 (a) Differential pair sensing CM input; (b) equivalent circuit of (a).

We now obtain the output voltages as

$$V_X = -g_{m1}(V_{in,CM} - V_P)R_D \tag{4.48}$$

$$= \frac{-g_{m1}}{(g_{m1} + g_{m2})R_{SS} + 1} R_D V_{in,CM} \tag{4.49}$$

and

$$V_Y = -g_{m2}(V_{in,CM} - V_P)R_D \tag{4.50}$$

$$= \frac{-g_{m2}}{(g_{m1} + g_{m2})R_{SS} + 1} R_D V_{in,CM} \tag{4.51}$$

The differential component at the output is therefore given by

$$V_X - V_Y = -\frac{g_{m1} - g_{m2}}{(g_{m1} + g_{m2})R_{SS} + 1} R_D V_{in,CM} \tag{4.52}$$

In other words, the circuit converts input CM variations to a differential error by a factor equal to

$$A_{CM-DM} = -\frac{\Delta g_m R_D}{(g_{m1} + g_{m2})R_{SS} + 1} \tag{4.53}$$

where A_{CM-DM} denotes common-mode to differential-mode conversion and $\Delta g_m = g_{m1} - g_{m2}$.

▶ **Example 4.10**

Two differential pairs are cascaded as shown in Fig. 4.35. Transistors M_3 and M_4 suffer from a g_m mismatch of Δg_m, the total parasitic capacitance at node P is represented by C_P, and the circuit is otherwise symmetric. What fraction of the supply noise appears as a differential component at the output? Assume that $\lambda = \gamma = 0$.

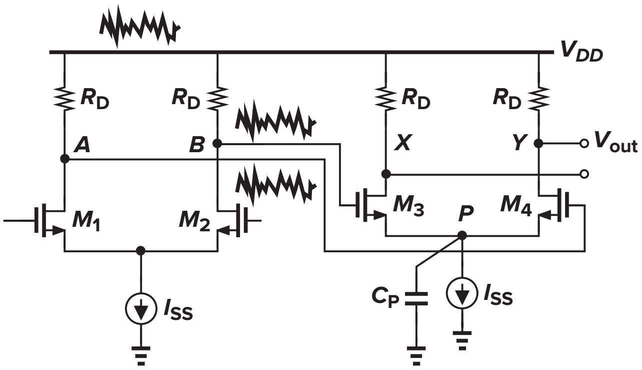


Figure 4.35

Solution

Neglecting the capacitance at nodes A and B, we note that the supply noise appears at these nodes with no attenuation. Substituting $1/(C_P s)$ for R_{SS} in (4.53) and taking the magnitude, we have

$$|A_{CM-DM}| = \frac{\Delta g_m R_D}{\sqrt{1 + (g_{m3} + g_{m4})^2 \left|\frac{1}{C_P \omega}\right|^2}} \tag{4.54}$$

The key point is that the effect becomes more noticeable as the supply noise frequency, ω, increases. ◀

For a meaningful comparison of differential circuits, the undesirable differential component produced by CM variations must be normalized to the wanted differential output resulting from amplification. We define the "common-mode rejection ratio" (CMRR) as the desired gain divided by the undesired gain:

$$\text{CMRR} = \left|\frac{A_{DM}}{A_{CM-DM}}\right| \tag{4.55}$$

If only g_m mismatch is considered, the reader can show from the analysis of Fig. 4.17 that

$$|A_{DM}| = \frac{R_D}{2}\frac{g_{m1} + g_{m2} + 4g_{m1}g_{m2}R_{SS}}{1 + (g_{m1} + g_{m2})R_{SS}} \tag{4.56}$$

and hence

$$\text{CMRR} = \frac{g_{m1} + g_{m2} + 4g_{m1}g_{m2}R_{SS}}{2\Delta g_m} \tag{4.57}$$

$$\approx \frac{g_m}{\Delta g_m}(1 + 2g_m R_{SS}) \tag{4.58}$$

where g_m denotes the mean value, that is, $g_m = (g_{m1} + g_{m2})/2$. In practice, all mismatches must be taken into account. Note that $2g_m R_{SS} \gg 1$, and hence $\text{CMRR} \approx 2g_m^2 R_{SS}/\Delta g_m$.

▶ **Example 4.11**

Our studies suggest that an ideal tail current source guarantees infinite CM rejection. Is this always true?

Solution

Interestingly, it is not. If the two transistors exhibit body-effect mismatch, then the circuit still converts an input CM change to a differential output component even if the tail impedance is infinite. As illustrated in Fig. 4.36, a change in $V_{in,CM}$ produces a change in V_P, and hence in V_{BS} of both transistors. If $g_{mb1} \neq g_{mb2}$, the change in $I_{D1}(= g_{mb1}V_{BS1})$ is not equal to that in I_{D2}, yielding a differential change at the output.

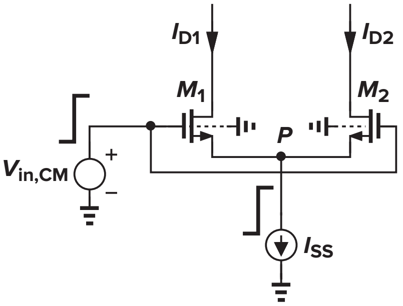


Figure 4.36

◀

4.4 ■ Differential Pair with MOS Loads

The load of a differential pair need not be implemented by linear resistors. As with the common-source stages studied in Chapter 3, differential pairs can employ diode-connected or current-source loads (Fig. 4.37). The small-signal differential gain can be derived using the half-circuit concept. For Fig. 4.37(a),

$$A_v = -g_{mN}\left(g_{mP}^{-1}\|r_{ON}\|r_{OP}\right) \tag{4.59}$$

$$\approx -\frac{g_{mN}}{g_{mP}} \tag{4.60}$$

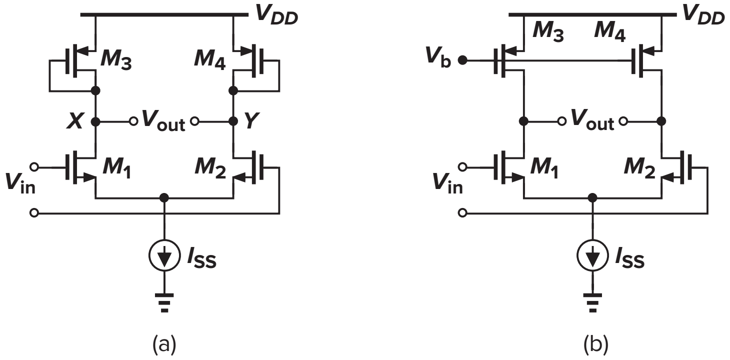


Figure 4.37 Differential pair with (a) diode-connected and (b) current-source loads.

where the subscripts N and P denote NMOS and PMOS, respectively. Expressing g_{mN} and g_{mP} in terms of device dimensions, we have

$$A_v \approx -\sqrt{\frac{\mu_n (W/L)_N}{\mu_p (W/L)_P}} \tag{4.61}$$

For Fig. 4.37(b), we have

$$A_v = -g_{mN}(r_{ON} \| r_{OP}) \tag{4.62}$$

▶ Example 4.12

It is possible to obviate the need for V_b in the circuit of Fig. 4.37(b) as shown in Fig. 4.38(a), where R_1 and $R_2 (= R_1)$ are relatively large. In the absence of signals, $V_X = V_Y = V_N = V_{DD} - |V_{GS3,4}|$. That is, M_3 and M_4 are "self-biased." Determine the differential voltage gain of this topology.

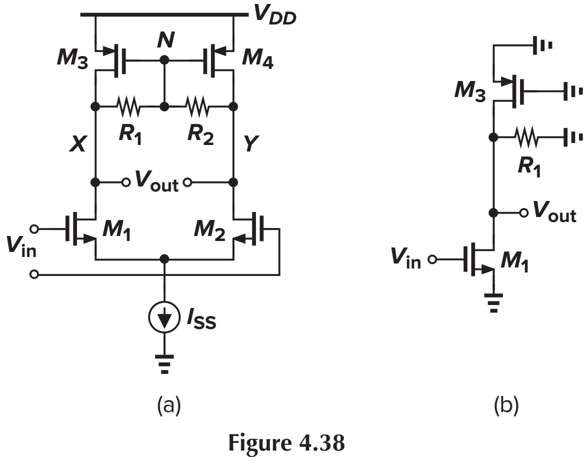


Figure 4.38

Solution

For differential outputs, V_N does not change (why?) and can be considered ac ground. Shown in Fig. 4.38(b), the half-circuit yields

$$|A_v| = g_{m1}(r_{O1} \| R_1 \| r_{O3}) \tag{4.63}$$

If the resistors are much greater than $r_{O1}||r_{O3}$, then they negligibly reduce the gain. ◀

In the circuit of Fig. 4.37(a), the diode-connected loads consume voltage headroom, thus creating a trade-off between the output voltage swings, the voltage gain, and the input CM range. Recall from Eq. (3.37) that, for given bias current and input device dimensions, the circuit's gain and the PMOS overdrive voltage scale together. To achieve a higher gain, $(W/L)_P$ must decrease, thereby increasing $|V_{GSP} - V_{THP}|$ and lowering the CM level at nodes X and Y.

In order to alleviate the above difficulty, part of the bias currents of the input transistors can be provided by PMOS current sources. Illustrated in Fig. 4.39(a), the idea is to lower the g_m of the load devices by reducing their current rather than their aspect ratio. For example, if the "auxiliary" current sources, M_5 and M_6, carry 80% of the drain current of M_1 and M_2, the current through M_3 and M_4 is reduced by a factor of five. For a given $|V_{GSP} - V_{THP}|$, this translates to a fivefold reduction in the transconductance of M_3 and M_4 because the aspect ratio of the devices can be lowered by the same factor. Thus, the differential gain is now five times that of the case with no PMOS current sources (if $\lambda = 0$).

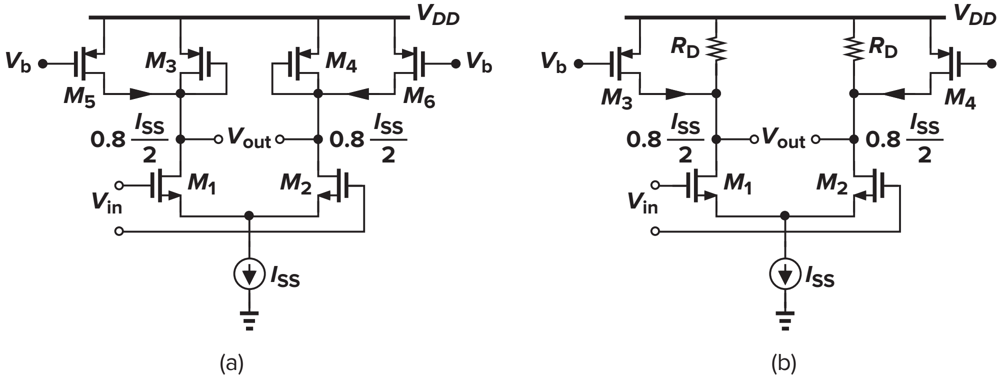


Figure 4.39 Addition of current sources to increase the voltage gain with (a) diode-connected loads and (b) resistive loads.

Since the voltage headroom consumed by diode-connected devices cannot be less than V_{TH} (if subthreshold conduction is neglected), the topology of Fig. 4.39(a) allows limited output voltage swings. We therefore prefer the alternative shown in Fig. 4.39(b), where the loads are realized by resistors—and the maximum voltage at each output node is equal to $V_{DD} - |V_{GS3,4} - V_{TH3,4}|$ rather than $V_{DD} - |V_{TH3,4}|$. For a given output CM level and 80% auxiliary currents, R_D can be five times as large, yielding a voltage gain of

$$|A_v| = g_{mN}(R_D||r_{ON}||r_{OP}) \tag{4.64}$$

If the PMOS devices are long (and, necessarily, wide), then $r_{OP} \gg r_{ON}$ and the gain is limited by $R_D||r_{ON}$. The circuit of Fig. 4.39(b) approaches that in Fig. 4.37(b) if $R_D \to \infty$, with the PMOS current sources providing *all* of the bias currents of M_1 and M_2.

The small-signal gain of the differential pair with current-source loads is relatively low—in the range of 5 to 10 in nanometer technologies. How do we increase the voltage gain? Borrowing ideas from the amplifiers in Chapter 3, we increase the output impedance of both the PMOS and the NMOS devices by cascoding, in essence creating a differential version of the cascode stage introduced in Chapter 3. The result is depicted in Fig. 4.40(a). To calculate the gain, we construct the half circuit of Fig. 4.40(b), which is similar to the cascode stage of Fig. 3.70. It follows that

$$|A_v| \approx g_{m1}[(g_{m3}r_{O3}r_{O1})\|(g_{m5}r_{O5}r_{O7})] \tag{4.65}$$

Cascoding therefore increases the differential gain substantially, but at the cost of consuming more voltage headroom. We return to this circuit in Chapter 9.

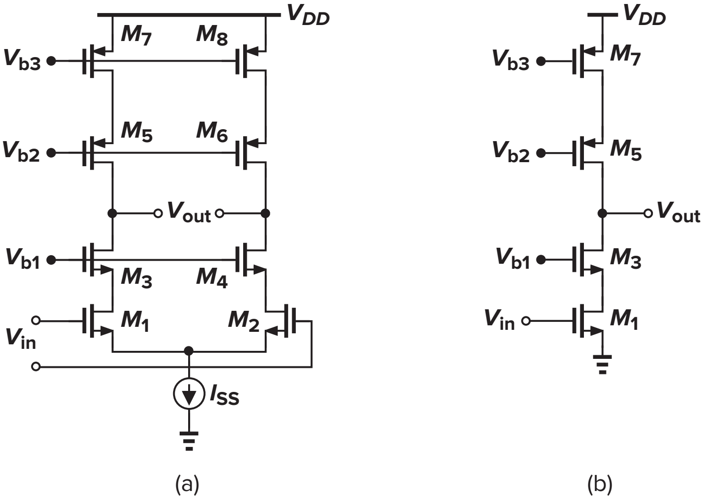


Figure 4.40 (a) Cascode differential pair; (b) half circuit of (a).

As a final note, we should mention that high-gain fully differential amplifiers require a means of defining the output common-mode level. For example, in Fig. 4.37(b), the output common-mode level is not well-defined, whereas in Fig. 4.37(a), diode-connected transistors define the output CM level as $V_{DD} - V_{GSP}$. We revisit this issue in Chapter 9.

4.5 ■ Gilbert Cell

Our study of differential pairs reveals two important aspects of their operation: (1) the small-signal gain of the circuit is a function of the tail current, and (2) the two transistors in a differential pair provide a simple means of steering the tail current to one of two destinations. By combining these two properties, we can develop a versatile building block.

Suppose we wish to construct a differential pair whose gain is varied by a control voltage. This can be accomplished as depicted in Fig. 4.41(a), where the control voltage defines the tail current and hence the gain. In this topology, $A_v = V_{out}/V_{in}$ varies from zero (if $I_{D3} = 0$) to a maximum value given by voltage headroom limitations and device dimensions. This circuit is a simple example of a "variable-gain amplifier" (VGA). VGAs find application in systems where the signal amplitude may experience large variations and hence requires inverse changes in the gain.

Now suppose we seek an amplifier whose gain can be continuously varied from a *negative* value to a *positive* value. Consider two differential pairs that amplify the input by opposite gains [Fig. 4.41(b)]. We now have $V_{out1}/V_{in} = -g_m R_D$ and $V_{out2}/V_{in} = +g_m R_D$, where g_m denotes the transconductance of each transistor in equilibrium. If I_1 and I_2 vary in opposite directions, so do $|V_{out1}/V_{in}|$ and $|V_{out2}/V_{in}|$.

But how should V_{out1} and V_{out2} be combined into a single output? As illustrated in Fig. 4.42(a), the two voltages can be summed, producing $V_{out} = V_{out1}+V_{out2} = A_1 V_{in}+A_2 V_{in}$, where A_1 and A_2 are controlled by V_{cont1} and V_{cont2}, respectively. The actual implementation is in fact quite simple: since $V_{out1} = R_D I_{D1} - R_D I_{D2}$ and $V_{out2} = R_D I_{D4} - R_D I_{D3}$, we have $V_{out1} + V_{out2} = R_D(I_{D1} + I_{D4}) - R_D(I_{D2} + I_{D3})$. Thus, rather than add V_{out1} and V_{out2}, we simply short the corresponding drain terminals to sum the currents and subsequently generate the output voltage [Fig. 4.42(b)]. Note that if $I_1 = 0$, then $V_{out} = +g_m R_D$, and if $I_2 = 0$, then $V_{out} = -g_m R_D$. For $I_1 = I_2$, the gain drops to zero.

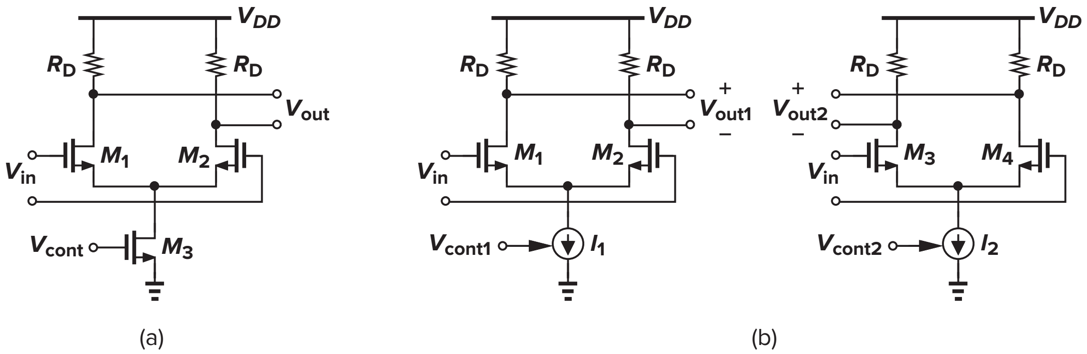


Figure 4.41 (a) Simple VGA; (b) two stages providing variable gain.

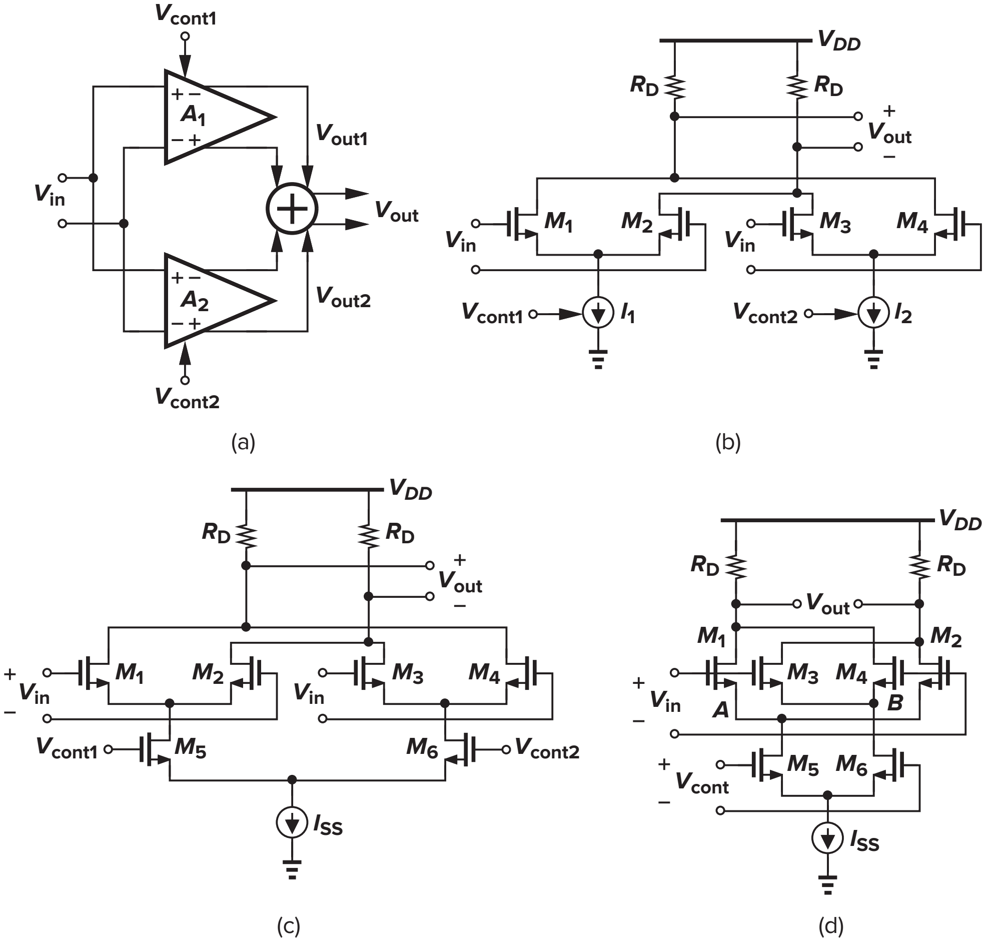


Figure 4.42 (a) Summation of the output voltages of two amplifiers; (b) summation in the current domain; (c) use of M_5-M_6 to control the gain; (d) Gilbert cell.

In the circuit of Fig. 4.42(b), V_{cont1} and V_{cont2} must change I_1 and I_2 in opposite directions such that the gain of the amplifier changes monotonically. What circuit can vary two currents in opposite directions? A differential pair provides such a characteristic, leading to the topology of Fig. 4.42(c). Note that for a large $|V_{cont1} - V_{cont2}|$, all of the tail current is steered to one of the top differential pairs and the gain from V_{in} to V_{out} is at its most positive or most negative value. If $V_{cont1} = V_{cont2}$, the gain is zero. For simplicity, we redraw the circuit as shown in Fig. 4.42(d). Called the "Gilbert cell" [2], this topology is widely used in many analog and communication systems. In a typical design, M_1–M_4 are identical, and so are M_5 and M_6.

▶ Example 4.13

Explain why the Gilbert cell can operate as an analog voltage multiplier.

Solution

Since the gain of the circuit is a function of $V_{cont} = V_{cont1} - V_{cont2}$, we have $V_{out} = V_{in} \cdot f(V_{cont})$. Expanding $f(V_{cont})$ in a Taylor series and retaining only the first-order term, αV_{cont}, we have $V_{out} = \alpha V_{in} V_{cont}$. Thus, the circuit can multiply voltages. This property accompanies any voltage-controlled variable-gain amplifier. ◀

As with a cascode structure, the Gilbert cell consumes a greater voltage headroom than a simple differential pair does. This is because the two differential pairs M_1–M_2 and M_3–M_4 are "stacked" on top of the control differential pair. To understand this point, suppose the differential input, V_{in}, in Fig. 4.42(d) has a common-mode level $V_{CM,in}$. Then, $V_A = V_B = V_{CM,in} - V_{GS1}$, where M_1–M_4 are assumed identical. For M_5 and M_6 to operate in saturation, the CM level of V_{cont}, $V_{CM,cont}$, must be such that $V_{CM,cont} \leq V_{CM,in} - V_{GS1} + V_{TH5,6}$. Since $V_{GS1} - V_{TH5,6}$ is roughly equal to one overdrive voltage, we conclude that the control CM level must be lower than the input CM level by at least this value.

In arriving at the Gilbert cell topology, we opted to vary the gain of each differential pair through its tail current, thereby applying the control voltage to the bottom pair and the input signal to the top pairs. Interestingly, the order can be exchanged while still obtaining a VGA. Illustrated in Fig. 4.43(a), the idea is to convert the input voltage to current by means of M_5 and M_6 and route the current through M_1–M_4 to the output nodes. If, as shown in Fig. 4.43(b), V_{cont} is very positive, then only M_1 and M_3 are on and $V_{out} = g_{m5,6} R_D V_{in}$. Similarly, if V_{cont} is very negative [Fig. 4.43(c)], then only M_2 and M_4 are on and $V_{out} = -g_{m5,6} R_D V_{in}$. For a zero differential control voltage, $V_{out} = 0$. The input differential pair may incorporate degeneration to provide a linear voltage-to-current conversion.

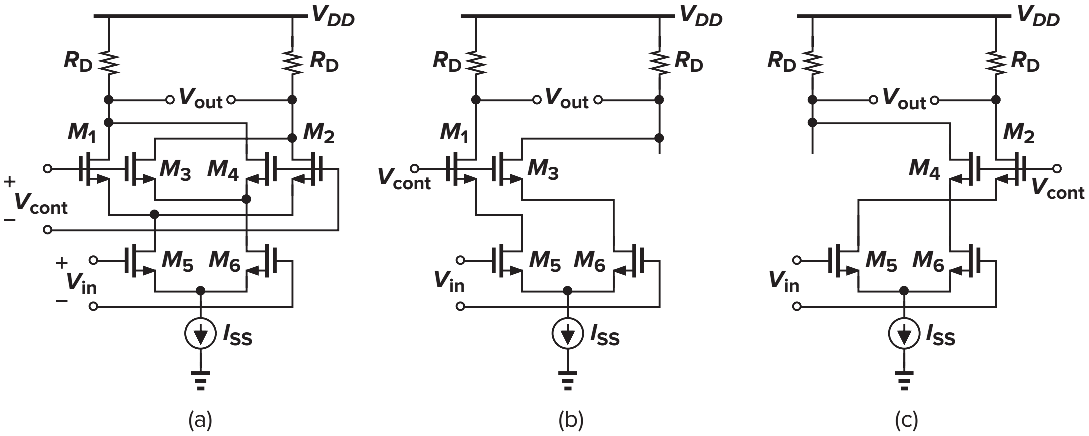


Figure 4.43 (a) Gilbert cell sensing the input voltage by the bottom differential pair; (b) signal path for very positive V_{cont}; (c) signal path for very negative V_{cont}.

References

[1] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3d ed. (New York: Wiley, 1993).
[2] B. Gilbert, “A Precise Four-Quadrant Multiplier with Subnanosecond Response,” *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 365–373, Dec. 1968.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. All device dimensions are effective values and in microns.

4.1. Suppose the total capacitance between adjacent lines in Fig. 4.2 is 10 fF and the capacitance from the drains of M_1 and M_2 to ground is 100 fF.
(a) What is the amplitude of the glitches in the analog output in Fig. 4.2(a) for a clock swing of 3 V?
(b) If in Fig. 4.2(b), the capacitance between L_1 and L_2 is 10% less than that between L_1 and L_3, what is the amplitude of the glitches in the differential analog output for a clock swing of 3 V?

4.2. Sketch the small-signal differential voltage gain of the circuit shown in Fig. 4.9(a) if V_{DD} varies from 0 to 3 V. Assume that $(W/L)_{1-3} = 50/0.5$, $V_{in,CM} = 1.3$ V, and $V_b = 1$ V.

4.3. Construct the plots of Fig. 4.9(c) for a differential pair using PMOS transistors.

4.4. In the circuit of Fig. 4.11, $(W/L)_{1,2} = 50/0.5$ and $I_{SS} = 0.5$ mA.
(a) What is the maximum allowable output voltage swing if $V_{in,CM} = 1.2$ V?
(b) What is the voltage gain under this condition?

4.5. A differential pair uses input NMOS devices with $W/L = 50/0.5$ and a tail current of 1 mA.
(a) What is the equilibrium overdrive voltage of each transistor?
(b) How is the tail current shared between the two sides if $V_{in1} - V_{in2} = 50$ mA?
(c) What is the equivalent G_m under this condition?
(d) For what value of $V_{in1} - V_{in2}$ does the G_m drop by 10%? By 90%?

4.6. Repeat Problem 4.5 with $W/L = 25/0.5$ and compare the results.

4.7. Repeat Problem 4.5 with a tail current of 2 mA and compare the results.

4.8. Sketch I_{D1} and I_{D2} in Fig. 4.19 versus $V_{in1} - V_{in2}$. For what value of $V_{in1} - V_{in2}$ are the two currents equal?

4.9. Consider the circuit of Fig. 4.32, assuming $(W/L)_{1,2} = 50/0.5$ and $R_D = 2$ kΩ. Suppose R_{SS} represents the output impedance of an NMOS current source with $(W/L)_{SS} = 50/0.5$ and a drain current of 1 mA. The input signal consists of $V_{in,DM} = 10\ \text{mV}_{pp}$ and $V_{in,CM} = 1.5\ \text{V} + V_n(t)$, where $V_n(t)$ denotes noise with a peak-to-peak amplitude of 100 mV. Assume that $\Delta R/R = 0.5\%$.
(a) Calculate the output differential signal-to-noise ratio, defined as the signal amplitude divided by the noise amplitude.
(b) Calculate the CMRR.

4.10. Repeat Problem 4.9 if $\Delta R = 0$, but M_1 and M_2 suffer from a threshold voltage mismatch of 1 mV.

4.11. Suppose the differential pair of Fig. 4.37(a) is designed with $(W/L)_{1,2} = 50/0.5$, $(W/L)_{3,4} = 10/0.5$, and $I_{SS} = 0.5$ mA. Also, I_{SS} is implemented with an NMOS device having $(W/L)_{SS} = 50/0.5$.
(a) What are the minimum and maximum allowable input CM levels if the differential swings at the input and output are small?
(b) For $V_{in,CM} = 1.2$ V, sketch the small-signal differential voltage gain as V_{DD} goes from 0 to 3 V.

4.12. In Problem 4.11, suppose M_1 and M_2 have a threshold voltage mismatch of 1 mV. What is the CMRR?

4.13. In Problem 4.11, suppose $W_3 = 10\ \mu$m, but $W_4 = 11\ \mu$m. Calculate the CMRR.

4.14. For the differential pairs of Fig. 4.37(a) and (b), calculate the differential voltage gain if $I_{SS} = 1$ mA, $(W/L)_{1,2} = 50/0.5$, and $(W/L)_{3,4} = 50/1$. What is the minimum allowable input CM level if I_{SS} requires at least 0.4 V across it? Using this value for $V_{in,CM}$, calculate the maximum output voltage swing in each case.

4.15. In the circuit of Fig. 4.39(a), assume that $I_{SS} = 1$ mA and $W/L = 50/0.5$ for all the transistors.
(a) Determine the voltage gain.
(b) Calculate V_b such that $I_{D5} = I_{D6} = 0.8(I_{SS}/2)$.
(c) If I_{SS} requires a minimum voltage of 0.4 V, what is the maximum differential output swing?

4.16. Assuming that all the circuits shown in Fig. 4.44 are symmetric, sketch V_{out} as **(a)** V_{in1} and V_{in2} vary differentially from zero to V_{DD}, and **(b)** V_{in1} and V_{in2} are equal and vary from zero to V_{DD}.

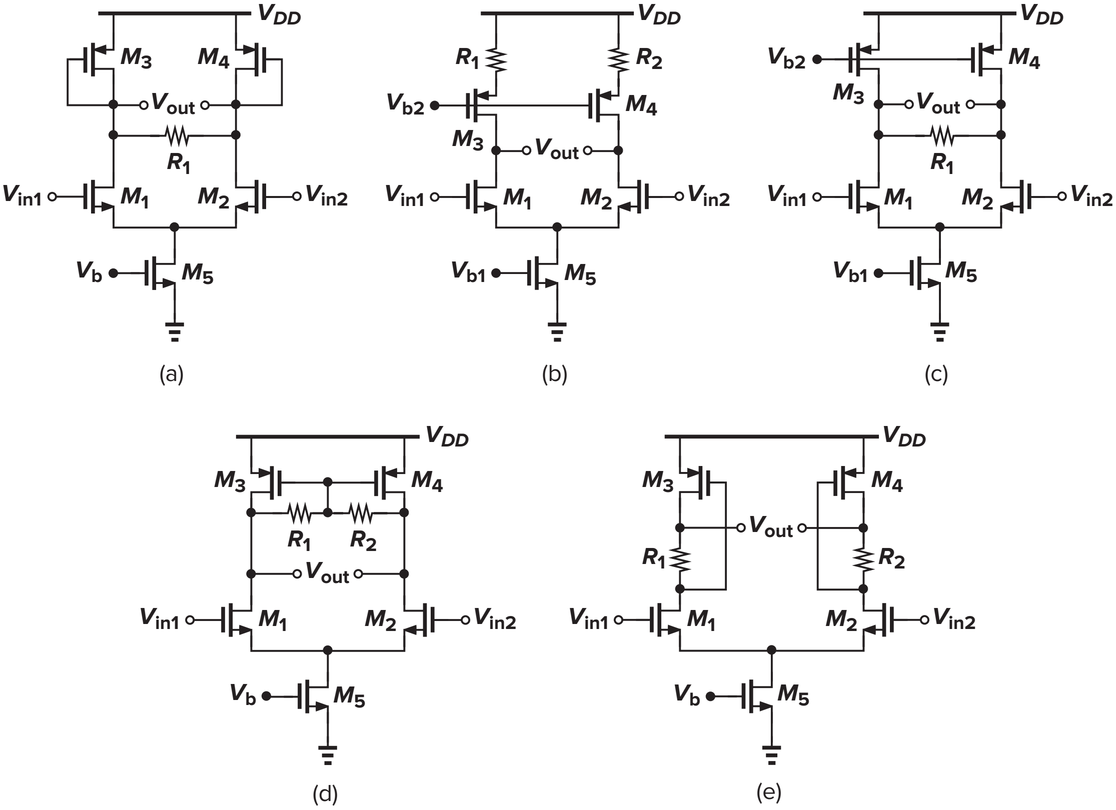


Figure 4.44

4.17. Assuming that all the circuits shown in Fig. 4.45 are symmetric, sketch V_{out} as **(a)** V_{in1} and V_{in2} vary differentially from zero to V_{DD}, and **(b)** V_{in1} and V_{in2} are equal and vary from zero to V_{DD}.

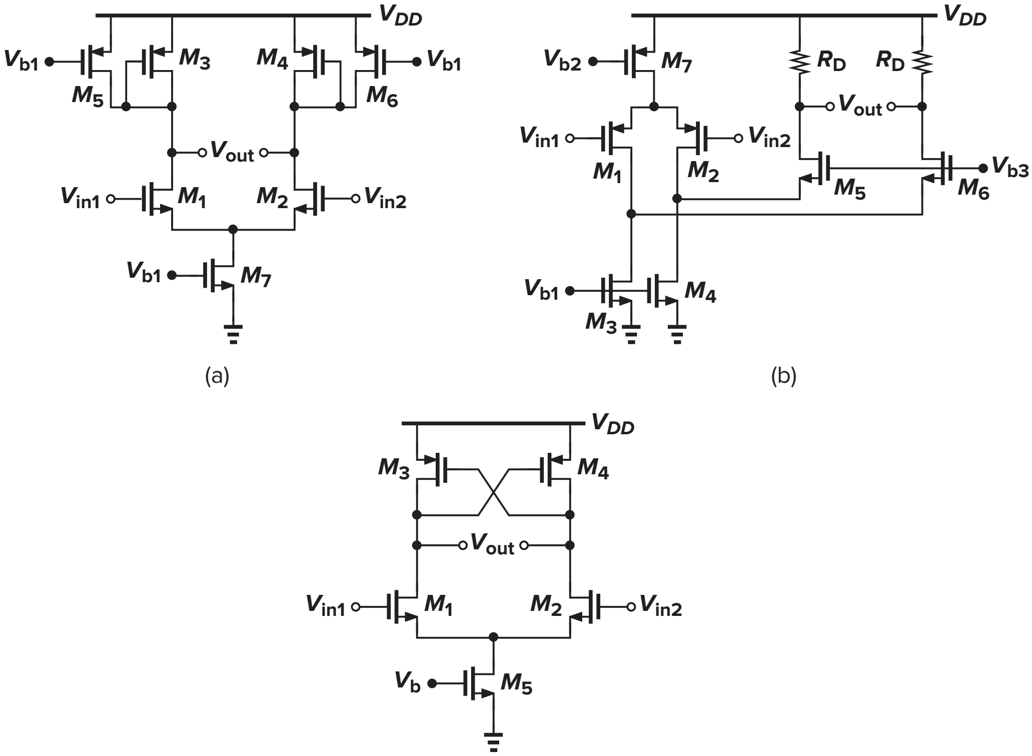
4.18. Assuming that all the transistors in the circuits of Figs. 4.44 and 4.45 are saturated and $\lambda \neq 0$, calculate the small-signal differential voltage gain of each circuit.

4.19. Consider the circuit shown in Fig. 4.46.
(a) Sketch V_{out} as V_{in1} and V_{in2} vary differentially from zero to V_{DD}.
(b) If $\lambda = 0$, obtain an expression for the voltage gain. What is the voltage gain if $W_{3,4} = 0.8W_{5,6}$?

4.20. For the circuit shown in Fig. 4.47,
(a) Sketch V_{out}, V_X, and V_Y as V_{in1} and V_{in2} vary differentially from zero to V_{DD}.
(b) Calculate the small-signal differential voltage gain.

4.21. Assuming no symmetry in the circuit of Fig. 4.48 and using no equivalent circuits, calculate the small-signal voltage gain $(V_{out})/(V_{in1} - V_{in2})$ if $\lambda = 0$ and $\gamma \neq 0$.

4.22. Due to a manufacturing defect, a large parasitic resistance has appeared between the drain and source terminals of M_1 in Fig. 4.49. Assuming $\lambda = \gamma = 0$, calculate the small-signal gain, common-mode gain, and CMRR.



(a)

(b)

(c)

Figure 4.45

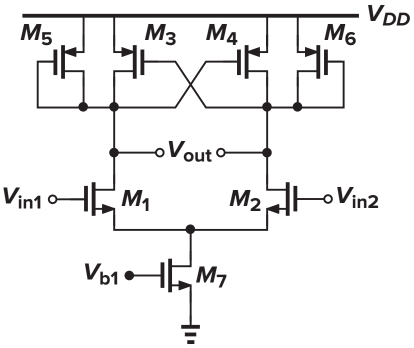


Figure 4.46

4.23. Due to a manufacturing defect, a large parasitic resistance has appeared between the drains of M_1 and M_4 in the circuit of Fig. 4.50. Assuming $\lambda = \gamma = 0$, calculate the small-signal gain, common-mode gain, and CMRR.

4.24. In the circuit of Fig. 4.51, all of the transistors have a W/L of 50/0.5, and M_3 and M_4 are to operate in the deep triode region with an on-resistance of 2 kΩ. Assuming that $I_{D5} = 20\ \mu\text{A}$ and $\lambda = \gamma = 0$, calculate the input common-mode level that yields such resistance. Sketch V_{out1} and V_{out2} as V_{in1} and V_{in2} vary differentially from 0 to V_{DD}.

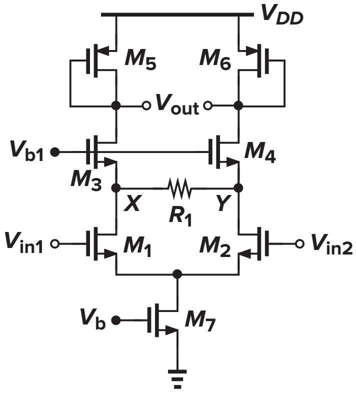


Figure 4.47

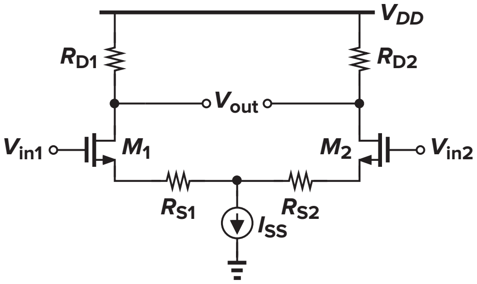


Figure 4.48

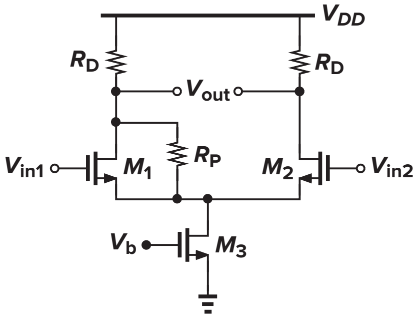


Figure 4.49

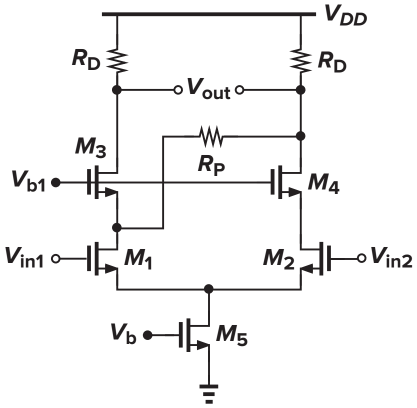


Figure 4.50

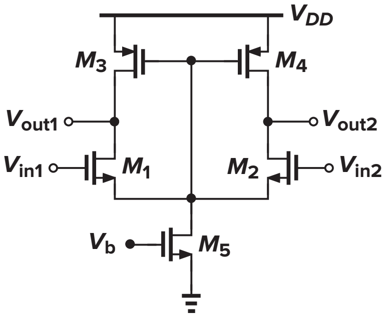


Figure 4.51

4.25. In the circuit of Fig. 4.37(b), $(W/L)_{1-4} = 50/0.5$ and $I_{SS} = 1$ mA.
(a) What is the small-signal differential gain?
(b) For $V_{in,CM} = 1.5$ V, what is the maximum allowable output voltage swing?

4.26. In the circuit of Fig. 4.39, assume that M_5 and M_6 have a small threshold voltage mismatch of ΔV and I_{SS} has an output impedance R_{SS}. Calculate the CMRR.

4.27. What happens if R_{SS} in Eq. (4.56) becomes very large? Can we obtain the same result by analyzing a differential pair having an ideal tail current source but $g_{m1} \neq g_{m2}$?

4.28. In Example 4.5, how much input dc imbalance can be tolerated if the small-signal gain must not drop by more than 5%?

4.29. In the lemma illustrated in Fig. 4.20, suppose channel-length modulation is not neglected. Assuming the two devices are connected to two equal load resistors, explain intuitively why the lemma still holds.

4.30. Does the lemma in Fig. 4.20 still hold if the devices have body effect? Explain.

4.31. Repeat Example 4.7 using Method I.

4.32. Prove the lemma illustrated in Fig. 4.20 if the tail current source is replaced by a resistor R_T.

4.33. What happens to the plots on Fig. 4.13 as W/L increases? Determine the area under the G_m plot and use the result to explain why the peak G_m must increase as W/L increases.

4.34. Assuming that I_1 and I_{SS} in Fig. 4.52 are ideal and $\lambda, \gamma > 0$, determine V_{out1}/V_{in} and V_{out2}/V_{in}.

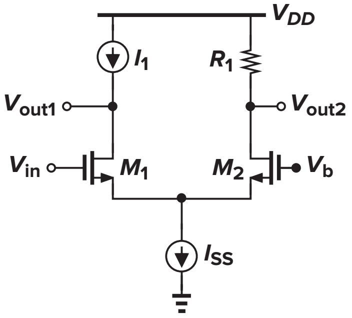


Figure 4.52

4.35. In Problem 4.11, suppose M_3 and M_4 have a threshold voltage mismatch of 1 mV. Calculate the CMRR.

CHAPTER

5

Current Mirrors and Biasing Techniques

Our study of single-stage and differential amplifiers in Chapters 3 and 4 points to the wide usage of current sources. In these circuits, current sources act as a large resistor without consuming excessive voltage headroom. We also noted that MOS devices operating in saturation can act as a current source.

Current sources find other applications in analog design as well. For example, some digital-to-analog (D/A) converters employ an array of current sources to produce an analog output proportional to the digital input. Also, current sources, in conjunction with "current mirrors," can perform useful functions on analog signals.

This chapter deals with the design of current mirrors and bias circuits. Following a review of basic current mirrors, we study the cascode mirror. Next, we analyze active current mirrors and describe the properties of differential pairs using such circuits as loads. Finally, we introduce various biasing techniques for amplifier stages.

5.1 ■ Basic Current Mirrors

Figure 5.1 illustrates two examples in which a current source proves useful. From our study in Chapter 2, recall that the output resistance and capacitance and the voltage headroom of a current source trade with the magnitude of the output current. In addition to these issues, several other aspects of current sources are important: supply, process, and temperature dependence; output noise current; and matching with other current sources. We defer the noise and matching considerations to Chapters 7 and 14, respectively.

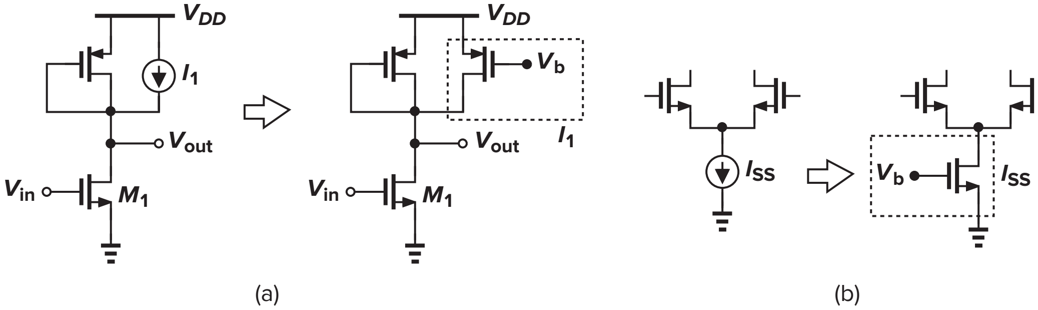


Figure 5.1 Applications of current sources.

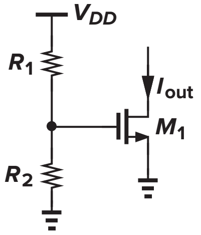


Figure 5.2 Definition of current by resistive divider.

How should a MOSFET be biased so as to operate as a stable current source? To gain a better view of the issues, let us consider the simple resistive biasing shown in Fig. 5.2. Assuming M_1 is in saturation, we can write

$$I_{out} \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left(\frac{R_2}{R_1+R_2}V_{DD}-V_{TH}\right)^2 \tag{5.1}$$

This expression reveals various PVT dependencies of I_{out}. The overdrive voltage is a function of V_{DD} and V_{TH}; the threshold voltage may vary by 50 to 100 mV from wafer to wafer. Furthermore, both μ_n and V_{TH} exhibit temperature dependence. Thus, I_{out} is poorly defined. The issue becomes more severe as the device is biased with a smaller overdrive voltage, e.g., to consume less headroom and support greater voltage swings at the drain. With a nominal overdrive of, say, 200 mV, a 50-mV error in V_{TH} results in a 44% error in the output current.

It is important to note that process and temperature dependencies exist even if the gate voltage is not a function of the supply voltage. In other words, if the gate-source *voltage* of a MOSFET is precisely defined, then its drain *current* is not! For this reason, we must seek other methods of biasing MOS current sources.

The design of current sources in analog circuits is based on "copying" currents from a reference, with the assumption that *one* precisely-defined current source is already available. While this method may appear to entail an endless loop, it is carried out as illustrated in Fig. 5.3. A relatively complex circuit—sometimes requiring external adjustments—is used to generate a stable reference current, I_{REF}, which is then "cloned" to create many current sources in the system. We study the copying operation here and the reference generator (which is based on "bandgap" techniques) in Chapter 12.

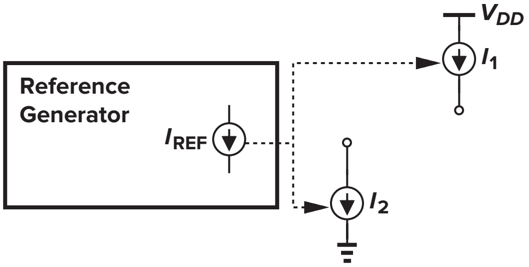


Figure 5.3 Use of a reference to generate various currents.

How do we generate copies of a reference current? For example, in Fig. 5.4, how do we guarantee that $I_{out} = I_{REF}$? For a MOSFET, if $I_D = f(V_{GS})$, where $f(\cdot)$ denotes the dependence of I_D upon V_{GS}, then $V_{GS} = f^{-1}(I_D)$. That is, if a transistor is biased at I_{REF}, then it produces $V_{GS} = f^{-1}(I_{REF})$ [Fig. 5.5(a)]. Thus, if this voltage is applied to the gate and source terminals of a second MOSFET, the resulting current is $I_{out} = f[f^{-1}(I_{REF})] = I_{REF}$ [Fig. 5.5(b)]. From another point of view, two identical MOS devices that have equal gate-source voltages and operate in saturation carry equal currents (if $\lambda = 0$).

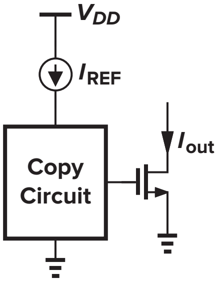


Figure 5.4 Conceptual means of copying currents.

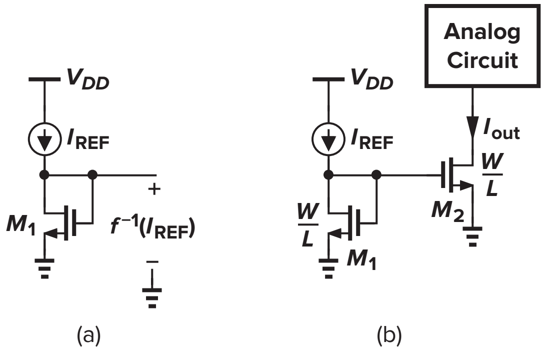


Figure 5.5 (a) Diode-connected device providing inverse function; (b) basic current mirror.

The structure consisting of M_1 and M_2 in Fig. 5.5(b) is called a "current mirror." In the general case, the transistors need not be identical. Neglecting channel-length modulation, we can write

$$I_{REF} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_1 (V_{GS} - V_{TH})^2 \tag{5.2}$$

$$I_{out} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_2 (V_{GS} - V_{TH})^2 \tag{5.3}$$

obtaining

$$I_{out} = \frac{(W/L)_2}{(W/L)_1} I_{REF} \tag{5.4}$$

The key property of this topology is that it allows precise copying of the current with no dependence on process and temperature. The translation from I_{REF} to I_{out} merely involves the *ratio* of device dimensions, a quantity that can be controlled with reasonable accuracy.

It is important to appreciate the cause-and-effect relationships stipulated by $V_{GS} = f^{-1}(I_{REF})$ and $f[f^{-1}(I_{REF})] = I_{REF}$. The former suggests that we must *generate* a V_{GS} *from* I_{REF}; i.e., I_{REF} is the cause and V_{GS} is the effect. A MOSFET can perform this function only if it is configured as a diode while carrying a current of I_{REF} [M_1 in Fig. 5.5(b)]. Similarly, the latter equation indicates that a transistor must sense $f^{-1}(I_{REF})$ $(= V_{GS})$ and generate $f[f^{-1}(I_{REF})]$. In this case, the cause is V_{GS} and the effect is the output current, $f[f^{-1}(I_{REF})]$ [as provided by M_2 in Fig. 5.5(b)].

With the aid of these observations, we can understand why a circuit such as that in Fig. 5.6 does not perform current *copying*. Here, V_b is *not* caused by I_{REF}, and hence I_{out} does not track I_{REF}.

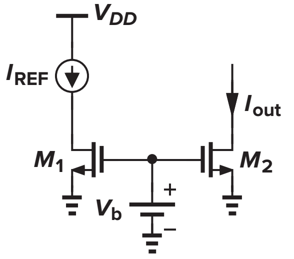


Figure 5.6 Circuit incapable of copying current.

▶ **Example 5.1**

In Fig. 5.7, find the drain current of M_4 if all of the transistors are in saturation.

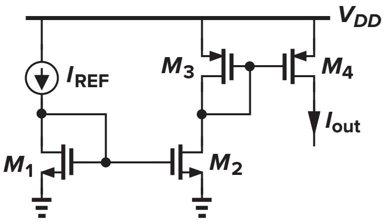


Figure 5.7

Solution

We have $I_{D2} = I_{REF}[(W/L)_2/(W/L)_1]$. Also, $|I_{D3}| = |I_{D2}|$ and $I_{D4} = I_{D3} \times [(W/L)_4/(W/L)_3]$. Thus, $|I_{D4}| = \alpha\beta I_{REF}$, where $\alpha = (W/L)_2/(W/L)_1$ and $\beta = (W/L)_4/(W/L)_3$. Proper choice of α and β can establish large or small ratios between I_{D4} and I_{REF}. For example, $\alpha = \beta = 5$ yields a magnification factor of 25. Similarly, $\alpha = \beta = 0.2$ can be utilized to generate a small, well-defined current. ◀

We should also remark that the copy of a copy may not be as "clear" as the original. Owing to random "mismatches" between M_1 and M_2 in the above example, I_{D2} slightly deviates from its nominal value. Similarly, as I_{D2} is copied onto I_{D4}, additional errors accumulate. We must therefore avoid long current mirror chains.

Current mirrors find wide application in analog circuits. Figure 5.8 illustrates a typical case, where a differential pair is biased by means of an NMOS mirror for the tail current source and a PMOS mirror

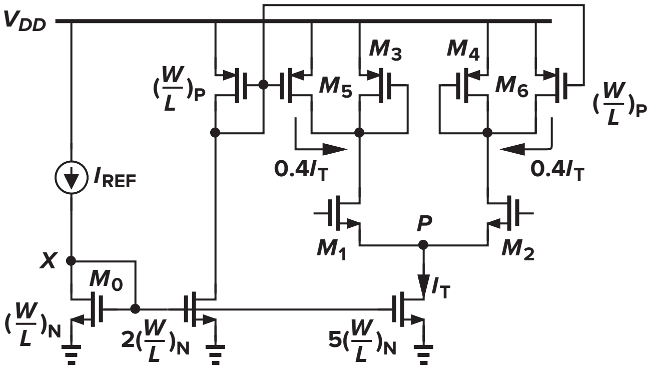


Figure 5.8 Current mirrors used to bias a differential amplifier.

for the load current sources. The device dimensions shown establish a drain current of $0.4I_T$ in M_5 and M_6, reducing the drain current of M_3 and M_4 and hence increasing the amplifier's gain.

Sizing Issues Current mirrors usually employ the same *length* for all of the transistors so as to minimize errors due to the side-diffusion of the source and drain areas (L_D). For example, in Fig. 5.8, the NMOS current sources must have the same channel length as M_0. This is because if L_{drawn} is, say, doubled, then $L_{eff} = L_{drawn} - 2L_D$ is not. Furthermore, the threshold voltage of short-channel devices exhibits some dependence on the channel length (Chapter 17). Thus, current ratioing is achieved by scaling only the width of the transistors.

Suppose we wish to copy a reference current, I_{REF}, and generate $2I_{REF}$. We begin with a width of W_{REF} for the diode-connected reference transistor and hence choose $2W_{REF}$ for the current source [Fig. 5.9(a)]. Unfortunately, direct scaling of the width also faces difficulties. As illustrated in Fig. 5.9(b), since the "corners" of the gate are poorly defined, if the drawn W is doubled, the actual width does not exactly double. We thus prefer to employ a "unit" transistor and create copies by *repeating* such a device [Fig. 5.9(c)].

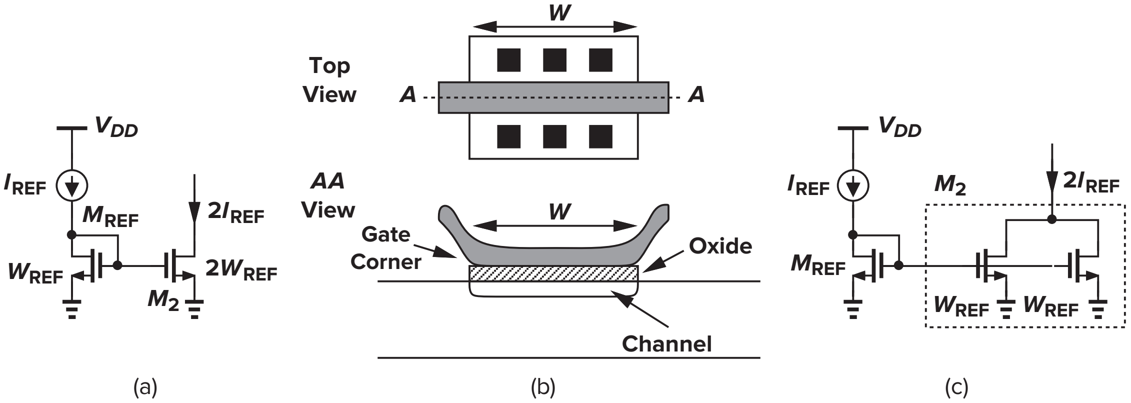


Figure 5.9 (a) Current mirror multiplying I_{REF} by 2, (b) effect of gate corner on current accuracy, and (c) more accurate current multiplication.

But how do we generate a current equal to $I_{REF}/2$ from I_{REF}? In this case, the diode-connected device itself must consist of *two* units, each carrying $I_{REF}/2$. Figure 5.10(a) depicts an example for the generation of both $2I_{REF}$ and $I_{REF}/2$; each unit has a width of W_0 (and the same length).

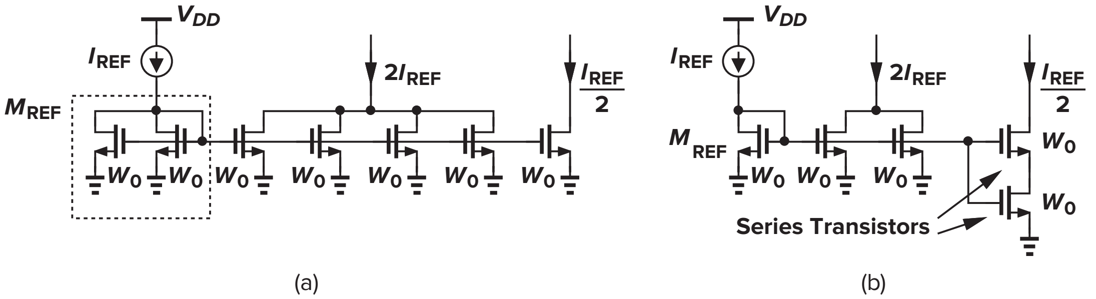


Figure 5.10 Current mirrors providing $I_{REF}/2$ from I_{REF} by (a) half-width device and (b) series transistors.

The above approach requires a large number of unit transistors if many different currents must be generated. It is possible to reduce the complexity by scaling the *lengths*, but not directly. In order to avoid the errors due to L_D, we can, for example, double the equivalent length by placing two unit transistors *in series*. Illustrated in Fig. 5.10(b), this approach preserves an effective length of $L_{drawn} - 2L_D$ for each unit, yielding an equivalent length of $2(L_{drawn} - 2L_D)$ for the composite device and hence halving the current. Note that this structure is *not* a cascode because the bottom device is in the triode region (why?).

We should also mention that current mirrors can process *signals* as well. In Fig. 5.5(b), for example, if I_{REF} increases by ΔI, then I_{out} increases by $\Delta I (W/L)_2/(W/L)_1$. That is, the circuit *amplifies* the small-signal current if $(W/L)_2/(W/L)_1 > 1$ (but at the cost of proportional multiplication of the bias current).

▶ **Example 5.2**

Calculate the small-signal voltage gain of the circuit shown in Fig. 5.11.

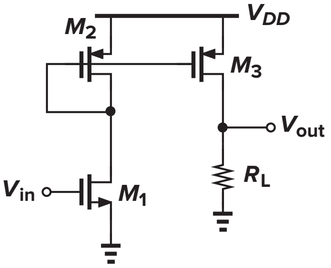


Figure 5.11

Solution

The small-signal drain current of M_1 is equal to $g_{m1}V_{in}$. Since $I_{D2} = I_{D1}$ and $I_{D3} = I_{D2}(W/L)_3/(W/L)_2$, the small-signal drain current of M_3 is equal to $g_{m1}V_{in}(W/L)_3/(W/L)_2$, yielding a voltage gain of $g_{m1}R_L(W/L)_3/(W/L)_2$.

◀

5.2 ■ Cascode Current Mirrors

In our discussion of current mirrors thus far, we have neglected channel-length modulation. In practice, this effect produces significant error in copying currents, especially if minimum-length transistors are used so as to minimize the width and hence the output capacitance of the current source. For the simple mirror of Fig. 5.5(b), we can write

$$I_{D1} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_1 (V_{GS} - V_{TH})^2(1 + \lambda V_{DS1}) \tag{5.5}$$

$$I_{D2} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)_2 (V_{GS} - V_{TH})^2(1 + \lambda V_{DS2}) \tag{5.6}$$

and hence

$$\frac{I_{D2}}{I_{D1}} = \frac{(W/L)_2}{(W/L)_1} \cdot \frac{1 + \lambda V_{DS2}}{1 + \lambda V_{DS1}} \tag{5.7}$$

While $V_{DS1} = V_{GS1} = V_{GS2}$, V_{DS2} may not equal V_{GS2} because of the circuitry fed by M_2. For example, in Fig. 5.8, the potential at node P is determined by the input common-mode level and the gate-source voltage of M_1 and M_2, and it may not equal V_X.

In order to suppress the effect of channel-length modulation in Fig. 5.5(b), we can (1) force V_{DS2} to be equal to V_{DS1}, or (2) force V_{DS1} to be equal to V_{DS2}. These two principles lead to two different topologies.

First Approach We begin with the first principle and wish to ensure that V_{DS2} in Fig. 5.5(b) is both *constant* and equal to V_{DS1}. Recall from Chapter 3 that a cascode device can shield a current source, thereby reducing the voltage variations across it. As shown in Fig. 5.12(a), even though the analog circuit may allow V_P to vary substantially, V_Y remains relatively constant. But how do we ensure that $V_{DS2} = V_{DS1}$? We must generate V_b such that $V_b - V_{GS3} = V_{DS1}$ $(= V_{GS1})$, i.e., $V_b = V_{GS3} + V_{GS1}$. In other words, V_b can be established by two diode-connected devices in series [Fig. 5.12(b)], provided that $V_{GS0} + V_{GS1} = V_{GS3} + V_{GS1}$, and hence $V_{GS0} = V_{GS3}$. We now attach the V_b generator of Fig. 5.12(b) to the cascode current source as shown in Fig. 5.12(c). The result allows accurate copying of the current even in the presence of body effect (why?).

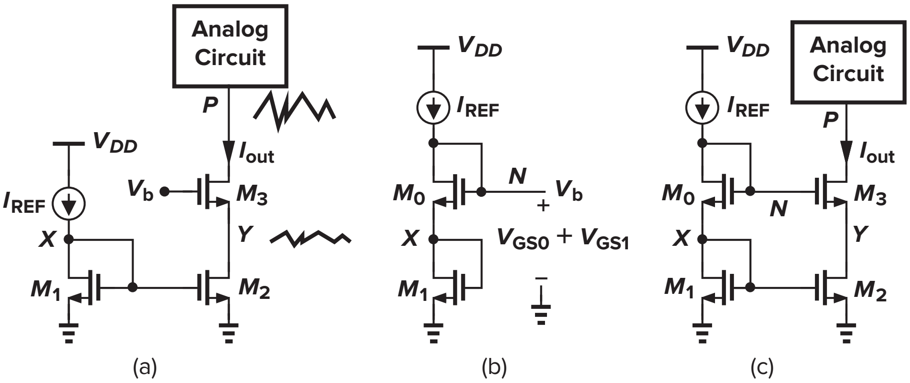


Figure 5.12 (a) Cascode current source, (b) modification of mirror circuit to generate the cascode bias voltage, and (c) cascode current mirror.

A few notes on the sizing of the transistors in Fig. 5.12(c) are warranted. As explained earlier, we typically select $L_2 = L_1$ and scale W_2 (in integer units) with respect to W_1 to obtain the desired multiple of I_{REF}. Similarly, for V_{GS3} to be equal to V_{GS0}, we choose $L_3 = L_0$ and scale W_3 with respect to W_0 by the same factor, i.e., $W_3/W_0 = W_2/W_1$. In practice, L_3 and L_0 are equal to the minimum allowable value so as to minimize their width, while L_1 and L_2 may be longer in some cases.[1]

▶ **Example 5.3**

In Fig. 5.13, sketch V_X and V_Y as a function of I_{REF}. If I_{REF} requires 0.5 V to operate as a current source, what is its maximum value?

Solution

Since M_2 and M_3 are properly ratioed with respect to M_1 and M_0, we have $V_Y = V_X \approx \sqrt{2I_{REF}/[\mu_n C_{ox}(W/L)_1]} + V_{TH1}$. The behavior is plotted in Fig. 5.13(b).

To find the maximum value of I_{REF}, we note that

$$V_N = V_{GS0} + V_{GS1} \tag{5.8}$$

$$= \sqrt{\frac{2I_{REF}}{\mu_n C_{ox}}}\left[\sqrt{\left(\frac{L}{W}\right)_0} + \sqrt{\left(\frac{L}{W}\right)_1}\right] + V_{TH0} + V_{TH1} \tag{5.9}$$

[1] To reduce channel-length modulation, mismatches, or flicker noise.

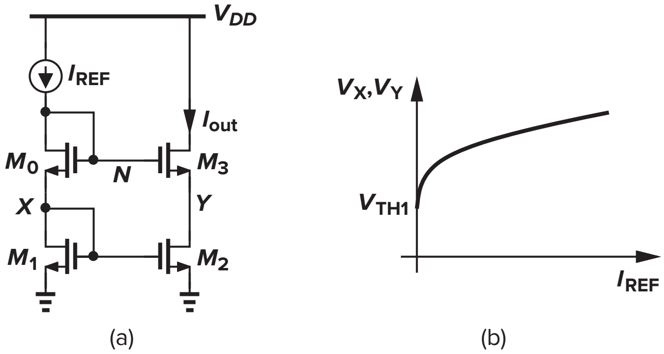


Figure 5.13

Thus,

$$V_{DD} - \sqrt{\frac{2I_{REF}}{\mu_n C_{ox}}} \left[\sqrt{\left(\frac{L}{W}\right)_0} + \sqrt{\left(\frac{L}{W}\right)_1} \right] - V_{TH0} - V_{TH1} = 0.5\text{ V} \tag{5.10}$$

and hence

$$I_{REF,max} = \frac{\mu_n C_{ox}}{2} \frac{(V_{DD} - 0.5\text{ V} - V_{TH0} - V_{TH1})^2}{(\sqrt{(L/W)_0} + \sqrt{(L/W)_1})^2} \tag{5.11}$$

While operating as a current source with a high output impedance and accurate value, the topology of Fig. 5.12(c) nonetheless consumes substantial voltage headroom. For simplicity, let us neglect the body effect and assume that all of the transistors are identical. Then, the minimum allowable voltage at node P is equal to

$$V_N - V_{TH} = V_{GS0} + V_{GS1} - V_{TH} \tag{5.12}$$

$$= (V_{GS0} - V_{TH}) + (V_{GS1} - V_{TH}) + V_{TH} \tag{5.13}$$

i.e., two overdrive voltages plus one threshold voltage. How does this value compare with that in Fig. 5.12(a) if V_b could be chosen more arbitrarily? As shown in Fig. 5.14(a), V_b could be so low

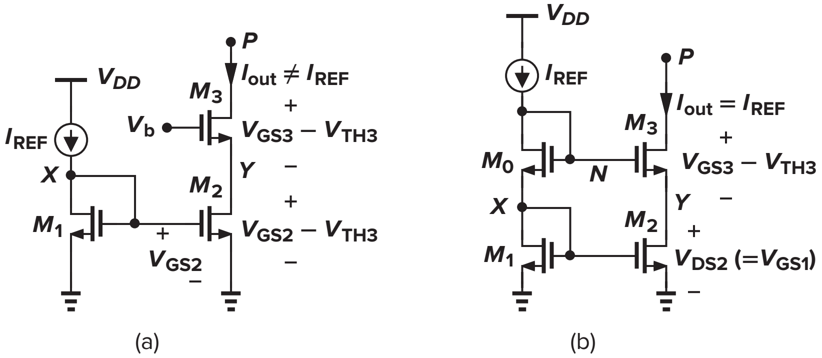


Figure 5.14 (a) Cascode current source with minimum headroom voltage; (b) headroom consumed by a cascode mirror.

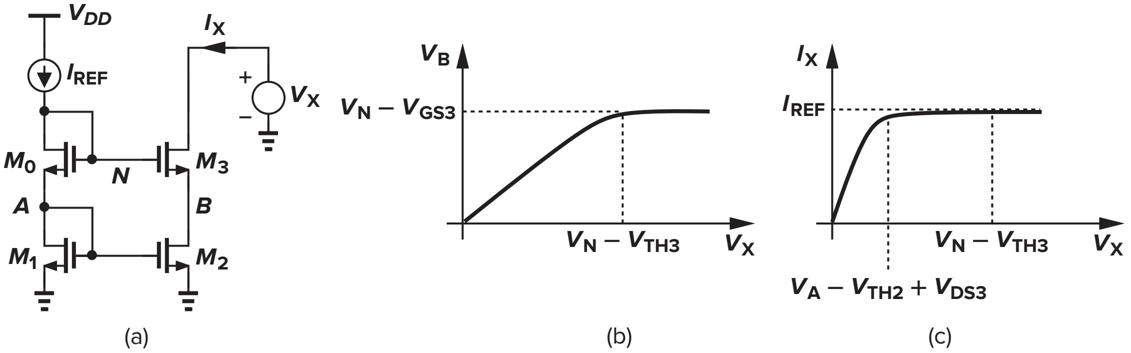
($= V_{GS3} + V_{GS2} - V_{TH2}$) that the minimum allowable voltage at P is merely two overdrive voltages. Thus, the cascode mirror of Fig. 5.12(c) "wastes" one threshold voltage in the headroom. This is because $V_{DS2} = V_{GS2}$, whereas V_{DS2} *could* be as low as $V_{GS2} - V_{TH}$ while maintaining M_2 in saturation.

Figure 5.14 summarizes our discussion. In Fig. 5.14(a), V_b is chosen to allow the lowest possible value of V_P, but the output current does not accurately track I_{REF} because M_1 and M_2 sustain unequal drain-source voltages. In Fig. 5.14(b), a higher accuracy is achieved, but the minimum level at P is higher by one threshold voltage.

Before resolving this issue, it is instructive to examine the large-signal behavior of a cascode current source.

▶ Example 5.4

In Fig. 5.15(a), assume that all of the transistors are identical and sketch I_X and V_B as V_X drops from a large positive value.



(a) (b) (c)

Figure 5.15

Solution

For $V_X \geq V_N - V_{TH}$, both M_2 and M_3 are in saturation, $I_X = I_{REF}$ and $V_B = V_A$. As V_X drops, which transistor enters the triode region first, M_3 or M_2? Suppose M_2 enters the triode region before M_3 does. For this to occur, V_{DS2} must drop and, since V_{GS2} is constant, so must I_{D2}. This means that V_{GS3} increases while I_{D3} decreases, which is not possible if M_3 is still in saturation. Thus, M_3 enters the triode region first.

As V_X falls below $V_N - V_{TH}$, M_3 enters the triode region, requiring a greater gate-source overdrive to carry the same current. Thus, as shown in Fig. 5.15(b), V_B begins to drop, causing I_{D2} and hence I_X to decrease slightly. As V_X and V_B decrease further, eventually we have $V_B < V_A - V_{TH}$, and M_2 enters the triode region. At this point, I_{D2} begins to drop sharply. For $V_X = 0$, $I_X = 0$, and M_2 and M_3 operate in the deep triode region. Note that as V_X drops below $V_N - V_{TH3}$, the output impedance of the cascode falls rapidly because g_{m3} degrades in the triode region. ◀

Second Approach In order to avoid the V_{TH} penalty in the voltage headroom of the above cascode current source, we force V_{DS1} to be equal to V_{DS2} instead. To understand this principle, we return to Fig. 5.14(a) and recognize that the V_{TH} headroom consumption is eliminated *only* if $V_b = V_{GS3} + (V_{GS2} - V_{TH2})$, i.e., only if V_{DS2} is around one overdrive voltage. How can we then ensure that $V_{DS1} = V_{DS2}$ ($= V_{GS2} - V_{TH2}$)? Since M_1 is a diode-connected device, it appears impossible to expect a V_{DS1} less than one threshold.

A simple escape from the foregoing quandary is to create a deliberate voltage difference between the gate and drain of M_1 by a means of a resistor. Illustrated in Fig. 5.16(a), the idea is to choose $R_1 I_{REF} \approx V_{TH1}$ and $V_b = V_{GS3} + (V_{GS1} - V_{TH1})$. Now, $V_{DS1} = V_{GS1} - R_1 I_{REF} \approx V_{GS1} - V_{TH1}$, which is equal to $V_b - V_{GS3}$ and hence to V_{DS2}.

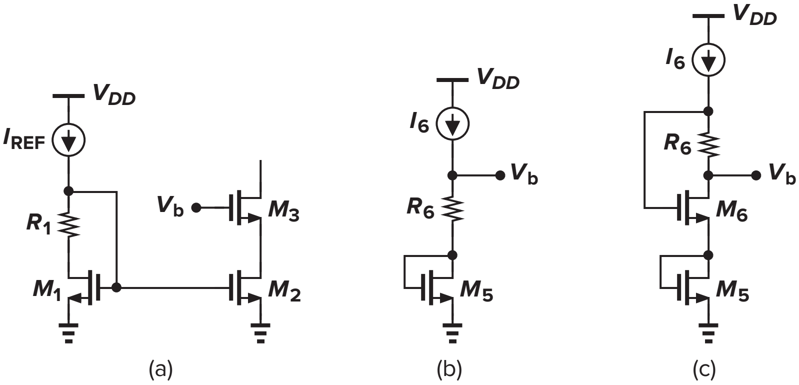


Figure 5.16 (a) Use of IR drop to improve accuracy of current mirror, (b) generation of V_b, and (c) alternative generation of V_b.

▶ Example 5.5

Is the M_1-R_1 combination in Fig. 5.16(a) a diode-connected device? Assume $\lambda > 0$.

Solution

From the small-signal equivalent shown in Fig. 5.17, we express the voltage drop across R_1 as $I_X R_1$ and write a KCL at the drain node:

$$\frac{V_X - I_X R_1}{r_O} + g_m V_X = I_X \tag{5.14}$$

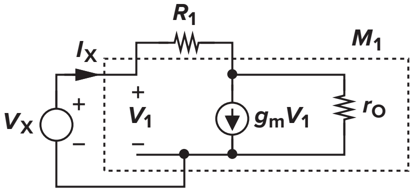


Figure 5.17

It follows that

$$\frac{V_X}{I_X} = \frac{R_1 + r_O}{1 + g_m r_O} \tag{5.15}$$

which reduces to $1/g_m$ in the absence of channel-length modulation. (Is it a coincidence that this impedance is the same as that seen at the source of a common-gate stage with $\gamma = 0$?!) Thus, from a small-signal point of view, the combination is close to a diode-connected device. From a large-signal point of view, $V_{GS1} \approx \sqrt{2I_D/[\mu_n C_{ox}(W/L)]} + V_{TH}$ if λ is small, suggesting diode-connected operation as well. ◀

The circuit of Fig. 5.16(a) entails two issues. First, in the presence of PVT variations, it may be difficult to guarantee that $R_1 I_{REF} \approx V_{TH1}$ as R_1 and V_{TH} may vary differently. Second, the generation of $V_b = V_{GS3} + (V_{GS1} - V_{TH1})$ is not straightforward. Let us address the latter issue first. We seek an arrangement that adds one gate-source voltage to an overdrive, surmising that we must begin with a diode-connected device. We consider the branch shown in Fig. 5.16(b) as a candidate and write $V_b = V_{GS5} + R_6 I_6$. We can readily choose I_6 and the dimensions of M_5 to ensure that $V_{GS5} = V_{GS3}$. However, the condition $R_6 I_6 = V_{GS1} - V_{TH1} = V_{GS1} - R_1 I_{REF}$ translates to $R_6 I_6 + R_1 I_{REF} = V_{GS1}$, which is difficult to meet because the IR products do not "track" the MOS gate-source voltage. For example, the value of the resistors may fall with temperature while V_{GS} may rise.

Depicted in Fig. 5.16(c) is another example, where M_5 establishes the V_{GS}, and M_6 and R_6 the overdrive. We select I_6 and the device parameters such that

$$V_{GS5} = V_{GS3} \tag{5.16}$$

$$V_{GS6} - R_6 I_6 = V_{GS1} - V_{TH1} \tag{5.17}$$

$$= V_{GS1} - R_1 I_{REF} \tag{5.18}$$

observing that it is now possible to ensure that V_{GS6} and V_{GS1} track each other, and so do $R_1 I_{REF}$ and $R_6 I_6$. For example, we may simply choose $I_6 = I_{REF}$, $R_6 = R_1$, and $(W/L)_6 = (W/L)_1$.[2]

To avoid the first issue mentioned above, we develop another circuit topology that forces the V_{DS} of the diode-connected device to be equal to the V_{DS} of the current source transistor. The level shift between the gate and drain voltages need not be created by a resistor. In particular, suppose we tie the output node of a cascode topology to its input [Fig. 5.18(a)]. In this case, $V_{DS1} = V_b - V_{GS0}$, and V_b can be chosen to place M_1 at the edge of saturation. We now connect this branch to the main cascode current source as shown in Fig. 5.18(b), recognizing that V_{DS1} is forced to be equal to V_{DS2} if $V_{GS0} = V_{GS3}$. Called a "low-voltage cascode," this configuration finds wider usage than the regular cascode shown in Fig. 5.14(b).

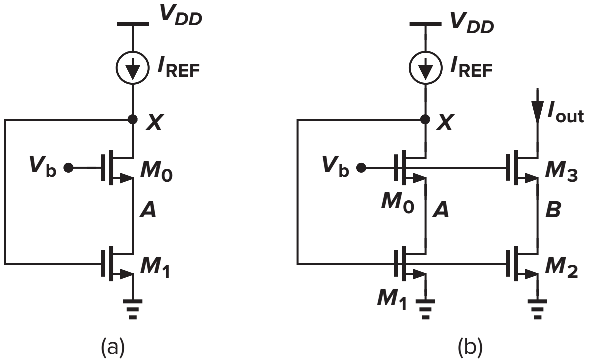


Figure 5.18 Modification of cascode mirror for low-voltage operation.

We must now answer two questions. First, how do we choose V_b in Fig. 5.18(a) so that both M_1 and M_0 are in saturation? We must have $V_b - V_{TH0} \leq V_X$ $(= V_{GS1})$ for M_0 to be saturated and $V_{GS1} - V_{TH1} \leq V_A$ $(= V_b - V_{GS0})$ for M_1 to be saturated. Thus,

$$V_{GS0} + (V_{GS1} - V_{TH1}) \leq V_b \leq V_{GS1} + V_{TH0} \tag{5.19}$$

A solution exists if $V_{GS0} + (V_{GS1} - V_{TH1}) < V_{GS1} + V_{TH0}$, i.e., if $V_{GS0} - V_{TH0} < V_{TH1}$. We must therefore size M_0 to ensure that its overdrive is well below V_{TH1}.

The second question is how to generate V_b. For minimal voltage headroom consumption, $V_A = V_{GS1} - V_{TH1}$, and hence V_b must be equal to (or slightly greater than) $V_{GS0} + (V_{GS1} - V_{TH1})$. Figure 5.19(a) depicts an example where M_5 generates $V_{GS5} \approx V_{GS0}$ and M_6 together with R_b

[2] The circuit incurs a small tracking error because M_6 experiences body effect but M_1 does not (and also because M_3 does but M_5 does not).

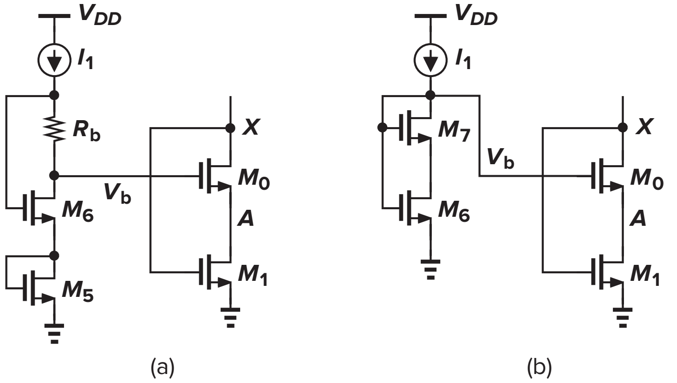


Figure 5.19 Generation of gate voltage V_b for cascode mirrors.

produces $V_{DS6} = V_{GS6} - R_bI_1 \approx V_{GS1} - V_{TH1}$. Some inaccuracy nevertheless arises because M_5 does not suffer from body effect whereas M_0 does. Also, the magnitude of R_bI_1 is not well-controlled.

A simpler alternative is shown in Fig. 5.19(b), where the diode-connected transistor M_7 provides the necessary V_{GS} and M_6 creates a V_{DS} equal to the required overdrive.

▶ Example 5.6

Shown in Fig. 5.20(a) is a differential pair along with its bias network. In this particular design, the voltage headroom is too small to allow the use of a cascode current source. Devise a method to reduce the current mirror error due to channel-length modulation.

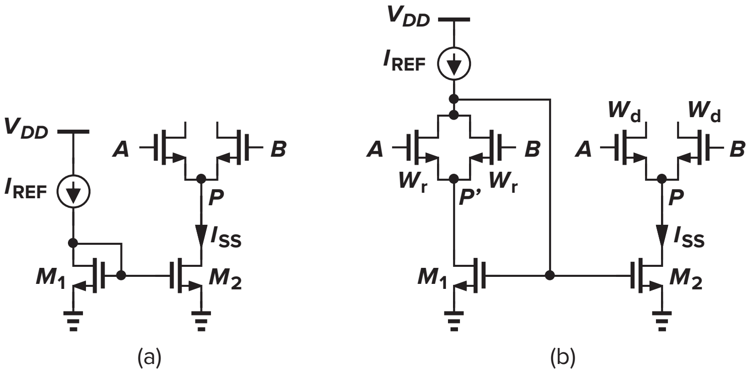


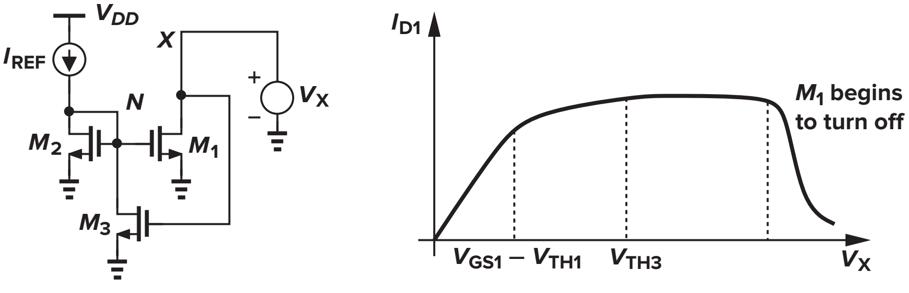
Figure 5.20

Solution

Since the limited headroom does not allow us to make V_{DS2} equal to V_{DS1}, we seek to make V_{DS1} equal to V_{DS2}. As exemplified by Fig. 5.16(a), we can simply insert a resistor in series with the drain of M_1 and select the voltage drop across it such that $V_{DS1} = V_{DS2}$. However, if variations in the circuit preceding the differential pair change the common-mode level at A and B, then $V_{DS1} \neq V_{DS2}$. We must therefore force the voltage at node P upon the drain of M_1. Let us replicate the differential pair and insert the replica as shown in Fig. 5.20(b). Now, the voltages at P' and P track even if the CM level at A and B varies. To ensure that $V_{P'} = V_P$, the two differential pairs must incorporate the same lengths and scale their widths according to $W_r/W_d = I_{REF}/I_{SS}$. Of course, if the CM level at A and B rises excessively, the replica transistors enter the triode region, introducing some error. ◀

▶ Example 5.7

Figure 5.21(a) shows an alternative current mirror exhibiting a high output impedance. Study the small-signal and large-signal properties of the circuit.



(a) (b)

Figure 5.21

Solution

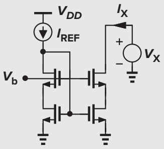
In this circuit, M_3 raises the output impedance by sensing the voltage change at node X and adjusting the voltage at node N. For example, suppose V_X rises by ΔV and tends to increase I_{D1} by $\Delta V/r_{O1}$. Transistor M_3 then draws a current change of $g_{m3}\Delta V$ from node N, causing V_N to fall by approximately $g_{m3}\Delta V/g_{m2}$ and I_{D1} to decrease by $(g_{m3}\Delta V/g_{m2})g_{m1}$. In other words, if we choose $g_{m3}g_{m1}/g_{m2} \approx r_{O1}^{-1}$, the net change in I_{D1} is small.

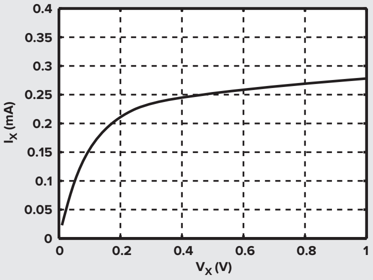
The circuit displays interesting large-signal properties. Let us sweep V_X from 0 to a high value and examine I_{D1}. At $V_X = 0$, M_1 operates in the deep triode region, carrying a zero current, and M_3 is off. As V_X rises, so does I_{D1} proportionally, up to $V_X = V_{GS1} - V_{TH1}$. Beyond this point, I_{D1} varies more gradually [Fig. 5.21(b)]. If V_X exceeds V_{TH3}, M_3 turns on and begins to "regulate" I_{D1}, creating a higher output impedance. However, for a sufficiently large V_X, M_3 absorbs all of I_{REF} and turns M_1 off.

While providing a high output impedance without a cascode device, the above circuit does pose its own voltage headroom limitation, i.e., V_X must exceed $V_{TH3}(> V_{DS,sat})$. ◀

Nanometer Design Notes

Owing to severe channel-length modulation in nanometer devices, even the cascode current mirror may exhibit substantial mismatches. We choose $W/L = 5\ \mu$m/40 nm for the devices in the circuit shown below and $I_{REF} = 0.25$ mA. As V_X is swept from low to high values, we observe that I_X still varies noticeably even though all transistors are in saturation for 0.4 V $< V_X$.





5.3 ■ Active Current Mirrors

As mentioned earlier and exemplified by the circuit of Fig. 5.11, current mirrors can also process signals, i.e., operate as "active" elements. Particularly useful is a type of mirror topology used in conjunction with differential pairs. In this section, we study this circuit and its properties. Shown in Fig. 5.22 and sometimes called a five-transistor "operational transconductance amplifier" (OTA), this topology finds application in many analog and digital systems and merits a detailed study here. Note that the output is single-ended; hence the circuit is sometimes used to convert differential signals to a single-ended output. We analyze a simpler topology with passive load before studying the OTA.

Differential Pair with Passive Load To generate a single-ended output, we may simply discard one output of a differential pair as shown in Fig. 5.23(a). Here, a current source in a "passive " mirror arrangement serves as the load. What is the small-signal gain, $A_v = V_{out}/V_{in}$, of this circuit? We

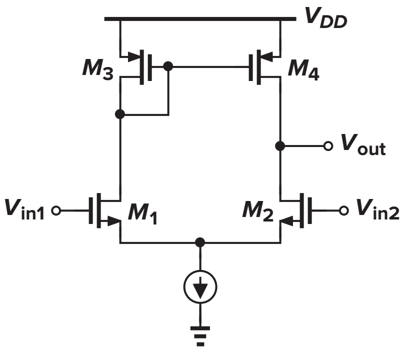


Figure 5.22 Five-transistor OTA.

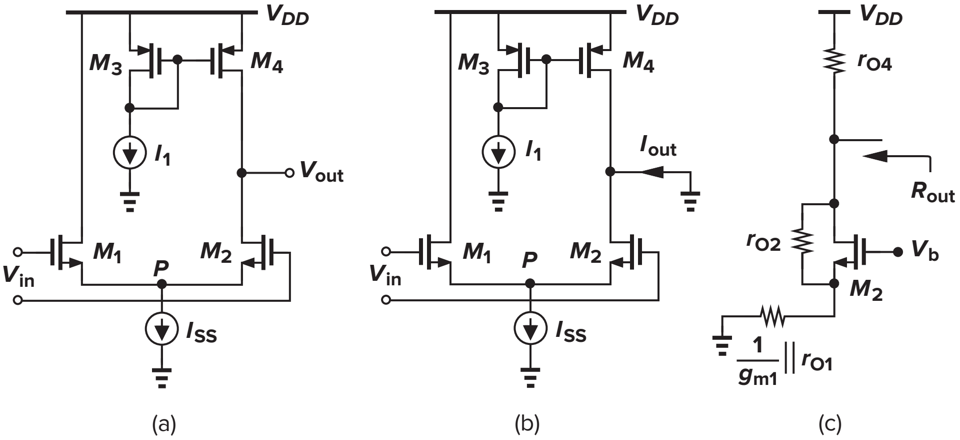


Figure 5.23 (a) Differential pair with current-source load; (b) circuit for calculation of G_m; (c) circuit for calculation of R_{out}.

calculate A_v using two different approaches, assuming $\gamma = 0$ for simplicity. Owing to the asymmetry, the half-circuit concept cannot be applied directly here.

Writing $|A_v| = G_m R_{out}$, we must compute the short-circuit transconductance, G_m, and the output resistance, R_{out}. We recognize from Fig. 5.23(b) that M_1 and M_2 become symmetric when the output is shorted to ac ground. Thus, $G_m = I_{out}/V_{in} = (g_{m1}V_{in}/2)/V_{in} = g_{m1}/2$. As illustrated in Fig. 5.23(c), for the R_{out} calculation, M_2 is degenerated by the source output impedance of M_1, $R_{deg} = (1/g_{m1})||r_{O1}$, thereby exhibiting an output impedance equal to $(1 + g_{m2}r_{O2})R_{deg} + r_{O2}a \approx 2r_{O2}$. It follows that $R_{out} = (2r_{O2})\|r_{O4}$, and

$$|A_v| = \frac{g_{m1}}{2}[(2r_{O2})\|r_{O4}] \tag{5.20}$$

Interestingly, if $r_{O4} \to \infty$, then $A_v \to -g_{m1}r_{O2}$.

In our second approach, we calculate V_P/V_{in} and V_{out}/V_P in Fig. 5.23(a) and multiply the results to obtain V_{out}/V_{in}. With the aid of Fig. 5.24 and viewing M_1 as a source follower, we write

$$\frac{V_P}{V_{in}} = \frac{R_{eq}||r_{O1}}{R_{eq}||r_{O1} + \dfrac{1}{g_{m1}}} \tag{5.21}$$

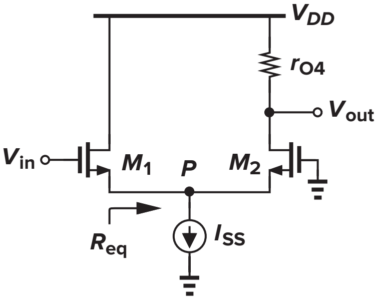


Figure 5.24 Circuit for calculation of V_P/V_{in}.

where R_{eq} denotes the resistance seen looking into the source of M_2. Since the drain of M_2 is terminated by a relatively large resistance, r_{O4}, the value of R_{eq} must be obtained from Eq. (3.117):

$$R_{eq} = \frac{r_{O2} + r_{O4}}{1 + g_{m2}r_{O2}} \tag{5.22}$$

It follows that

$$\frac{V_P}{V_{in}} = \frac{g_{m1}r_{O1}(r_{O2} + r_{O4})}{(1 + g_{m1}r_{O1})(r_{O2} + r_{O4}) + (1 + g_{m2}r_{O2})r_{O1}} \tag{5.23}$$

We now calculate V_{out}/V_P. From Fig. 5.25,

$$\frac{V_{out}}{V_P} = \frac{(1 + g_{m2}r_{O2})r_{O4}}{r_{O2} + r_{O4}} \tag{5.24}$$

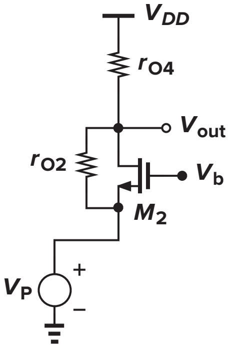


Figure 5.25 Circuit for calculation of V_{out}/V_P.

From (5.23) and (5.24), we have

$$\frac{V_{out}}{V_{in}} = \frac{g_{m2}r_{O2}r_{O4}}{2r_{O2} + r_{O4}} \tag{5.25}$$

$$= \frac{g_{m2}}{2}[(2r_{O2})\|r_{O4}] \tag{5.26}$$

Differential Pair with Active Load In the circuit of Fig. 5.23(a), the small-signal drain current of M_1 is "wasted." As conceptually shown in Fig. 5.26(a), it is desirable to utilize this current with proper polarity at the output. This can be accomplished by the five-transistor OTA shown in Fig. 5.26(b), where M_3 and M_4 are identical and operate as an active current mirror.

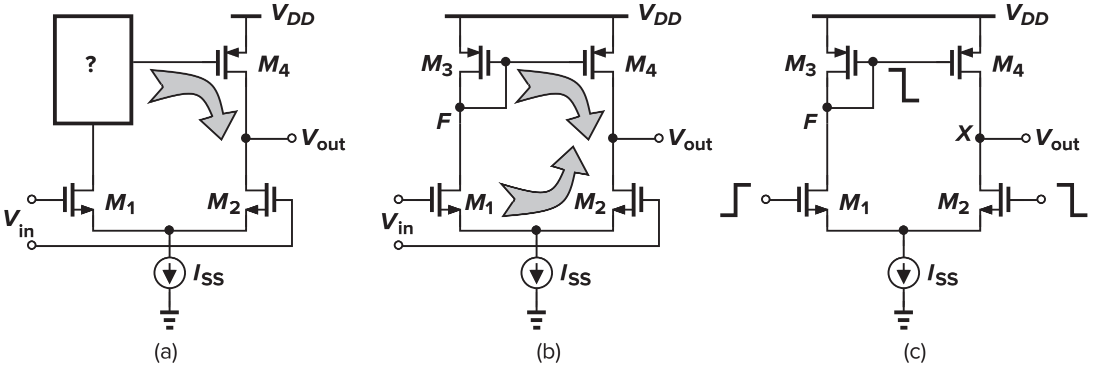


Figure 5.26 (a) Concept of combining the drain currents of M_1 and M_2, (b) realization of (a), and (c) response of the circuit to differential inputs.

To see how M_3 enhances the gain, suppose the gate voltages of M_1 and M_2 change by equal and opposite amounts [Fig. 5.26(c)]. Consequently, I_{D1} increases, V_F falls, and I_{D2} decreases. Thus, the output voltage rises by means of *two* mechanisms: M_2 draws less current from X to ground *and* M_4 pushes a greater current from V_{DD} to X. By contrast, in the circuit of Fig. 5.23(a), M_4 plays no active role in changing V_{out} because its gate voltage is constant. The five-transistor OTA is also called a differential pair with active load.

5.3.1 Large-Signal Analysis

Let us study the large-signal behavior of the five-transistor OTA. To this end, we replace the ideal tail current source by a MOSFET as shown in Fig. 5.27(a). If V_{in1} is much more negative than V_{in2}, M_1 is off, and so are M_3 and M_4. Since no current can flow from V_{DD}, both M_2 and M_5 operate in the deep triode region, carrying zero current. Thus, $V_{out} = 0$.[3] As V_{in1} approaches V_{in2}, M_1 turns on, drawing a fraction of I_{D5} from M_3 and turning M_4 on. The output voltage then depends on the difference between I_{D4} and I_{D2}. For a small difference between V_{in1} and V_{in2}, both M_2 and M_4 are saturated, providing a high gain [Fig. 5.27(b)]. As V_{in1} becomes more positive than V_{in2}, I_{D1}, $|I_{D3}|$, and $|I_{D4}|$ increase and I_{D2} decreases, allowing V_{out} to rise and eventually driving M_4 into the triode region. If $V_{in1} - V_{in2}$ is sufficiently large, M_2 turns off, M_4 operates in the deep triode region with zero current, and $V_{out} = V_{DD}$.

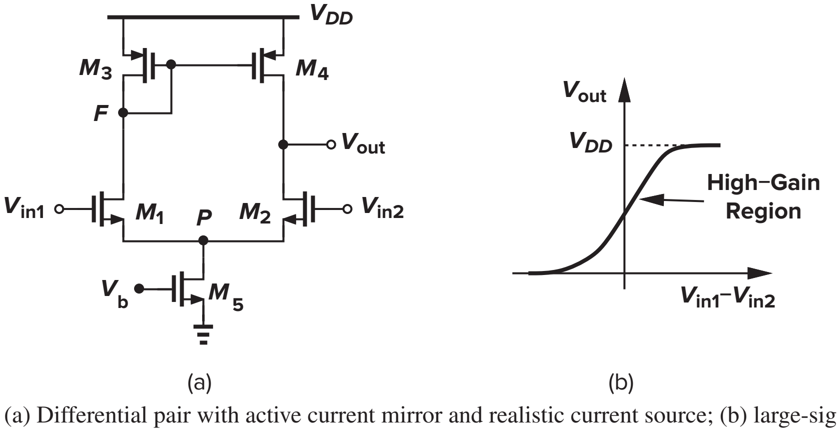


Figure 5.27 (a) Differential pair with active current mirror and realistic current source; (b) large-signal input-output characteristic.

[3] If V_{in1} is greater than one threshold voltage with respect to ground, M_5 may draw a small current from M_1, raising V_{out} slightly.

Note that if $V_{in1} > V_F + V_{TH}$, then M_1 enters the triode region. Also, V_{out} is in-phase with respect to V_{in1} but 180° out of phase with respect to V_{in2}.

The choice of the input common-mode voltage of the circuit is also important. For M_2 to be saturated, the output voltage cannot be less than $V_{in,CM} - V_{TH}$. Thus, to allow maximum output swings, the input CM level must be as low as possible, with the minimum given by $V_{GS1,2} + V_{DS5,min}$. The constraint imposed by the input CM level upon the output swing in this circuit is a critical drawback.

What is the output voltage of the circuit when $V_{in1} = V_{in2}$? With perfect symmetry, $V_{out} = V_F = V_{DD} - |V_{GS3}|$. This can be proved by contradiction as well. Suppose, for example, that $V_{out} < V_F$. Then, due to channel-length modulation, M_1 must carry a greater current than M_2 (and M_4 a greater current than M_3). In other words, the total current through M_1 is greater than half of I_{SS}. But this means that the total current through M_3 also exceeds $I_{SS}/2$, violating the assumption that M_4 carries more current than M_3. In reality, however, asymmetries in the circuit may result in a large deviation in V_{out}, possibly driving M_2 or M_4 into the triode region. For example, if the threshold voltage of M_2 is slightly smaller than that of M_1, the former carries a greater current than the latter even with $V_{in1} = V_{in2}$, causing V_{out} to drop significantly. For this reason, the circuit is rarely used in an open-loop configuration to amplify small signals. Nonetheless, the open-loop OTA proves useful as a differential to a single-ended converter for large swings, as illustrated by the following example.

▶ **Example 5.8**

Some digital circuits operate with differential (complementary) signals having voltage swings less than V_{DD}. For example, the single-ended swing can be 300 mV$_{pp}$. Explain how a five-transistor OTA can convert the moderate-swing differential signals to a single-ended rail-to-rail signal.

Solution

Consider the OTA shown in Fig. 5.28, where M_1 and M_2 sense swings equal to $V_2 - V_1 = 300$ mV. With proper choice of $(W/L)_{1,2}$ and I_{SS}, we can guarantee that such a swing turns off one side. For example, if M_1 carries all of I_{SS}, then M_2 remains off, allowing M_4 to pull V_{out} to V_{DD}. Conversely, when M_2 hogs I_{SS}, M_1, M_2, and M_4 turn off, M_2 and M_5 remain on with zero current, and $V_{out} = 0$. The "push-pull" action between M_2 and M_4 thus produces rail-to-rail swings at the output.

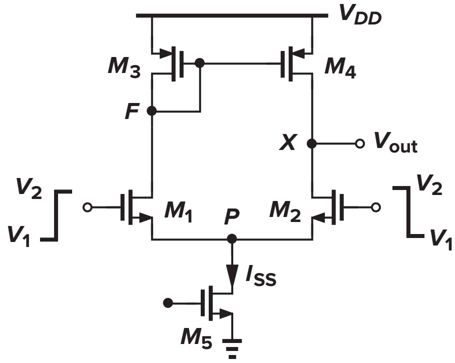


Figure 5.28

In practice, V_{out} does not reach exactly V_{DD} or zero if $V_1 > V_{TH1,2}$. The proof is left as an exercise for the reader. (Hint: if M_2 and M_5 are in the deep triode region, then V_P approaches zero, possibly turning on M_1.) For this reason, the OTA is typically followed by a CMOS inverter to obtain rail-to-rail swings.

◀

▶ **Example 5.9**

Assuming perfect symmetry, sketch the output voltage of the circuit in Fig. 5.29(a) as V_{DD} varies from 3 V to zero. Assume that for $V_{DD} = 3$ V, all of the devices are saturated.

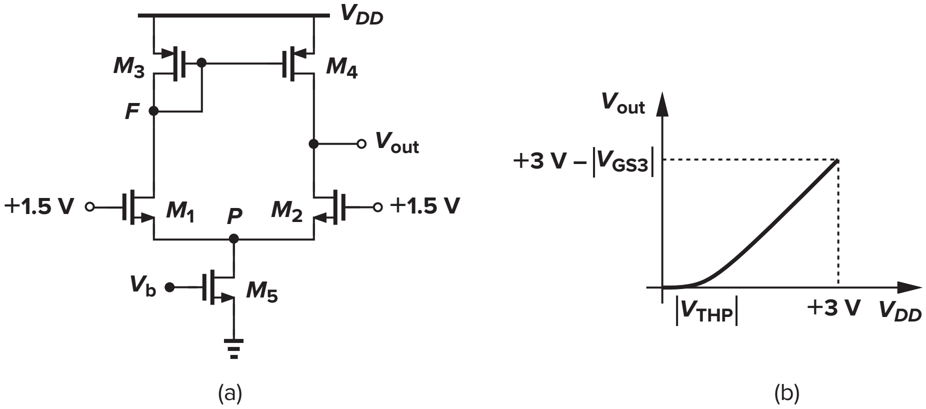


Figure 5.29

Solution

For $V_{DD} = 3$ V, symmetry requires that $V_{out} = V_F$. As V_{DD} drops, so do V_F and V_{out} with a slope close to unity [Fig. 5.29(b)]. As V_F and V_{out} fall below $+1.5\text{ V} - V_{THN}$, M_1 and M_2 enter the triode region, but their drain currents are constant if M_5 is saturated. Further decrease in V_{DD} and hence V_F and V_{out} causes V_{GS1} and V_{GS2} to increase, eventually driving M_5 into the triode region. Thereafter, the bias current of all of the transistors drops, lowering the rate at which V_{out} decreases. For $V_{DD} < |V_{THP}|$, we have $V_{out} = 0$.

▶ **Example 5.10**

Sketch the large-signal input-output characteristic of the unity-gain buffer shown in Fig. 5.30(a) if the op amp is realized as a five-transistor OTA.

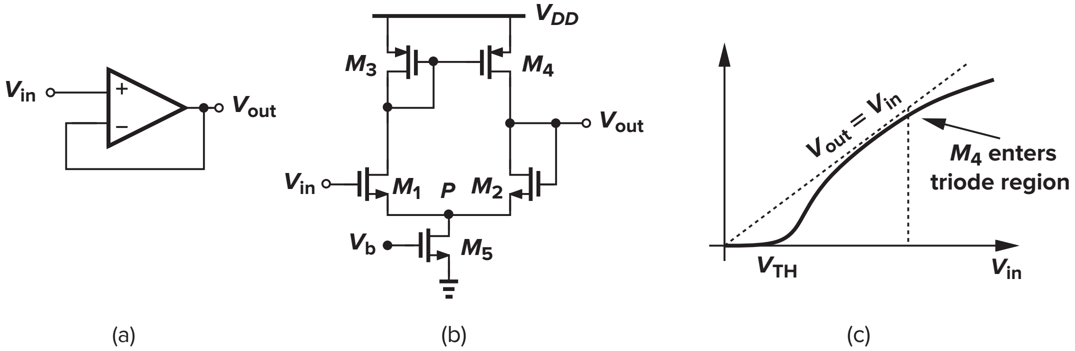


Figure 5.30

Solution

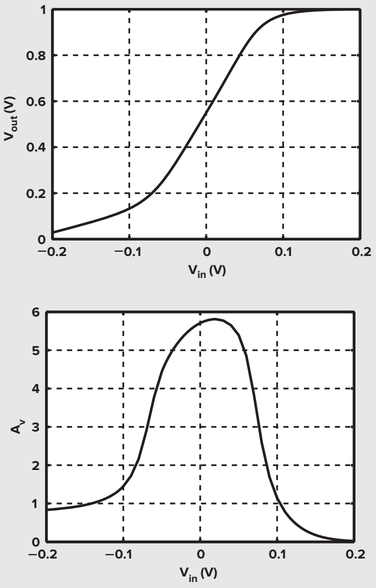
Drawing the circuit as shown in Fig. 5.30(b), we begin with $V_{in} = 0$ and note that M_1, M_3, and M_4 are off. Thus, M_5 enters the triode region with zero drain current and the diode-connected device M_2 sustains a zero V_{GS}.[4] We therefore have $V_{out} = V_P = 0$ [Fig. 5.30(c)]. As V_{in} rises and exceeds one threshold, M_1 begins to draw current from M_3,

[4] In constructing input-output characteristics, we assume that the input is changing slowly, and hence the subthreshold currents have enough time to reduce V_{GS} to zero.

turning M_4 and hence M_2 on. Note that, since $I_{D3} \approx I_{D4}$, we have $I_{D1} \approx I_{D2}$ and $V_{GS1} \approx V_{GS2}$. That is, $V_{out} \approx V_{in}$. This unity-gain action continues as V_{in} increases. For a sufficiently high V_{in}, two phenomena occur: (a) M_1 enters the triode region if $V_{in} > V_{DD} - |V_{GS3}| + V_{TH1}$, and (b) M_4 enters the triode region if $V_{out} > V_{DD} - |V_{GS4} - V_{TH4}|$, and hence $V_{in} > V_{DD} - |V_{GS4} - V_{TH4}|$. These two values are roughly equal if V_{TH1} and $|V_{TH4}|$ are comparable. Beyond this point, $|I_{D4}| < |I_{D3}|$ (why?), and hence $V_{GS1} > V_{GS2}$, yielding $V_{out} < V_{in}$. If $V_{in} = V_{DD}$, then M_4 carries little current and V_{out} incurs substantial error.

Nanometer Design Notes

The five-transistor OTA provides a limited gain in nanometer technologies and across a limited output range. With $V_{DD} = 1$ V, $W/L = 5\ \mu/40$ nm, a tail current of 0.25 mA, and input CM level of 0.5 V, we obtain the characteristic shown below. The slope plot also shows the steep drop in the gain as the NMOS or PMOS device tied to the output enters the triode region.



5.3.2 Small-Signal Analysis

We now analyze the small-signal properties of the circuit shown in Fig. 5.27(a), assuming $\gamma = 0$ for simplicity. Can we apply the half-circuit concept to calculate the differential gain here? As illustrated in Fig. 5.31, with small differential inputs, the voltage swings at nodes F and X are vastly different. This is because the diode-connected device M_3 yields a much lower voltage gain from the input to node F than that from the input to node X. As a result, the effects of V_F and V_X at node P (through r_{O1} and r_{O2}, respectively) do not cancel each other, and this node cannot be considered a virtual ground. Using the lemma $|A_v| = G_m R_{out}$, we first perform an approximate analysis so as to develop insight and then carry out an exact calculation of the gain.

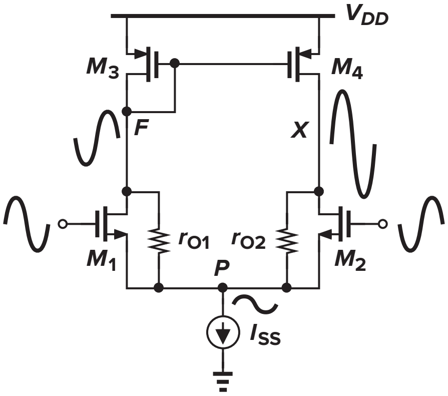


Figure 5.31 Asymmetric swings in a differential pair with active current mirror.

Approximate Analysis For the calculation of G_m, consider Fig. 5.32(a). The circuit is not quite symmetric, but because the impedance seen at node F is relatively low and the swing at this node small, the current returning from F to P through r_{O1} is negligible, and node P can be approximated by a virtual ground [Fig. 5.32(b)]. Thus, $I_{D1} = |I_{D3}| = |I_{D4}| = g_{m1,2}V_{in}/2$ and $I_{D2} = -g_{m1,2}V_{in}/2$, yielding $I_{out} = -g_{m1,2}V_{in}$, and hence $|G_m| = g_{m1,2}$. Note that, by virtue of active current mirror operation, this value is twice the transconductance of the circuit of Fig. 5.23(b).

Calculation of R_{out} is less straightforward. We may surmise that the output resistance of this circuit is equal to that of the circuit in Fig. 5.23(c), namely, $(2r_{O2})\|r_{O4}$. In reality, however, the active mirror operation yields a different value because when a voltage is applied to the output to measure R_{out}, the gate voltage of M_4 does not remain constant. Rather than draw the entire equivalent circuit, we observe that for small signals, I_{SS} is open [Fig. 5.33(a)], any current flowing into M_1 must flow out of M_2, and the role of the two transistors can be represented by a resistor $R_{XY} = 2r_{O1,2}$ [Fig. 5.33(b)]. As a result, the current drawn from V_X by R_{XY} is mirrored by M_3 onto M_4 with unity gain. This current is equal to

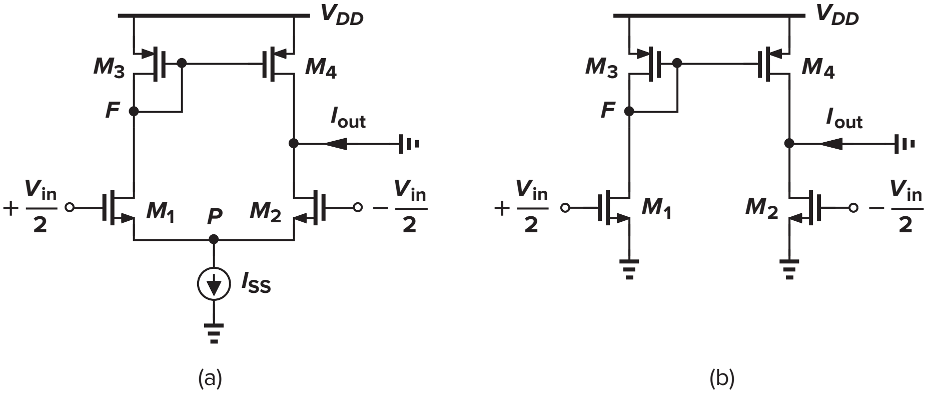


Figure 5.32 (a) Circuit for calculation of G_m; (b) circuit of (a) with node P grounded.

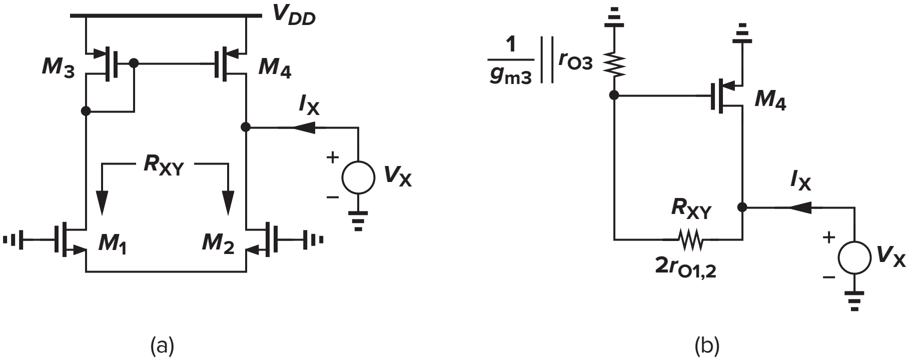


Figure 5.33 (a) Circuit for calculating R_{out}; (b) substitution of a resistor for M_1 and M_2.

$V_X/[2r_{O1,2} + (1/g_{m3})||r_{O3}]$. We multiply this current by $(1/g_{m3})||r_{O3}$ to obtain the gate-source voltage of M_4 and then multiply the result by g_{m4}. It follows that

$$I_X = \frac{V_X}{2r_{O1,2} + \dfrac{1}{g_{m3}}||r_{O3}} \left[1 + \left(\frac{1}{g_{m3}}||r_{O3}\right) g_{m4}\right] + \frac{V_X}{r_{O4}} \tag{5.27}$$

For $2r_{O1,2} \gg (1/g_{m3})\|r_{O3}$, we have

$$R_{out} \approx r_{O2}\|r_{O4} \tag{5.28}$$

The overall voltage gain is approximately equal to $|A_v| = G_m R_{out} = g_{m1,2}(r_{O2}\|r_{O4})$, somewhat higher than that of the circuit in Fig. 5.23(a).

Exact Analysis We must compute both the G_m and R_{out} of the OTA. Let us determine the G_m, without grounding node P, by solving the equivalent circuit shown in Fig. 5.34. For the sake of brevity, we use the subscript 1 to denote both M_1 and M_2. Since the current flowing downward through $(1/g_{m3})||r_{O3}$ (denoted by r_d hereafter) is $-V_4/r_d$, r_{O1} sustains a voltage equal to $(-V_4/r_d - g_{m1}V_1)r_{O1}$. Adding this voltage to $V_P = V_{in1} - V_1$, we have

$$\left(-\frac{V_4}{r_d} - g_{m1}V_1\right) r_{O1} + V_{in1} - V_1 = V_4 \tag{5.29}$$

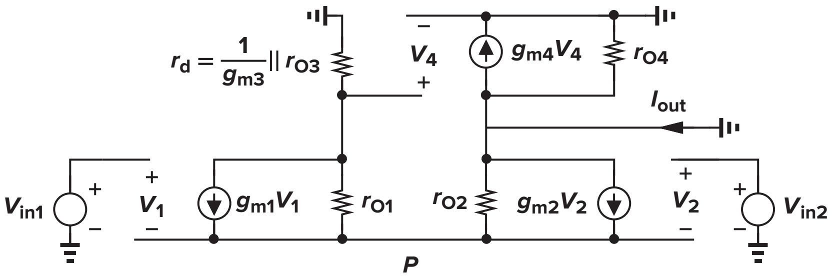


Figure 5.34 Equivalent circuit of five-transistor OTA

We also recognize that the sum of $g_{m2}V_2$ and the current flowing through r_{O2} is equal to V_4/r_d (why?). That is

$$g_{m2}V_2 - \frac{V_{in2} - V_2}{r_{O2}} = \frac{V_4}{r_D} \tag{5.30}$$

Obtaining V_1 and V_2 from these equations in terms of V_4 and noting that $V_1 - V_2 = V_{in1} - V_{in2}$ and $I_{out} = g_{m4}V_4 + V_4/r_d$, we arrive at

$$I_{out} = -g_{m1}r_{O1}\frac{g_{m4}r_d + 1}{r_d + 2r_{O1}}(V_{in1} - V_{in2}) \tag{5.31}$$

It follows that

$$G_m = -g_{m1}r_{O1}\frac{g_{m4}r_d + 1}{r_d + 2r_{O1}} \tag{5.32}$$

In the next step, we calculate R_{out}. Let us express the output admittance from Eq. (5.27) as

$$\frac{I_X}{V_X} = \frac{1 + g_{m4}r_d}{2r_{O1} + r_d} + \frac{1}{r_{O4}} \tag{5.33}$$

$$= \frac{(1 + g_{m4}r_d)r_{O4} + 2r_{O1} + r_d}{(2r_{O1} + r_d)r_{O4}} \tag{5.34}$$

and hence

$$G_m R_{out} = -g_{m1}r_{O1}\frac{(g_{m4}r_d + 1)r_{O4}}{(g_{m4}r_d + 1)r_{O4} + 2r_{O1} + r_d} \tag{5.35}$$

Since $r_d = r_{O3}/(1 + g_{m3}r_{O3})$, this expression reduces to

$$G_m R_{out} = -g_{m1}r_{O1}r_{O4}\frac{2g_{m3}r_{O3} + 1}{(2g_{m3}r_{O3} + 1)r_{O4} + 2r_{O1}(1 + g_{m3}r_{O3}) + r_{O3}} \tag{5.36}$$

$$= -\frac{g_{m1}r_{O1}r_{O4}}{r_{O1} + r_{O3}} \cdot \frac{2g_{m3}r_{O3} + 1}{2(g_{m3}r_{O3} + 1)} \tag{5.37}$$

We thus obtain a simple but exact expression for the gain:

$$|A_v| = g_{m1}(r_{O1}||r_{O4})\frac{2g_{m4}r_{O4} + 1}{2(g_{m4}r_{O4} + 1)} \tag{5.38}$$

We can view this result as our approximate solution, $g_{m1}(r_{O1}||r_{O4})$, multiplied by a "correction" factor that is *less* than unity. For example, if $g_{m4}r_{O4} = 5$, then $|A_v| = 0.92g_{m1}(r_{O1}||r_{O4})$.

▶ **Example 5.11**

With the aid of the above results, determine the output response to an input CM change if mismatches are neglected.

Solution

To represent an input CM change, we choose $V_{in1} = V_{in2}$ in Fig. 5.34, obtaining from Eq. (5.31) $I_{out} = 0$. The single-ended output voltage is therefore free from the input CM change.

◀

▶ **Example 5.12**

Calculate the small-signal voltage gain of the circuit shown in Fig. 5.35. How does the performance of this circuit compare with that of a differential pair with active mirror?

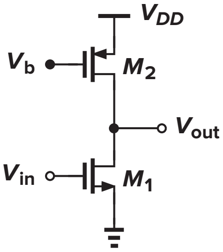


Figure 5.35

Solution

We have $A_v = g_{m1}(r_{O1}\|r_{O2})$, similar to the value derived above. For given device dimensions, this circuit requires half of the bias current to achieve the same gain as a differential pair. However, advantages of differential operation (less sensitivity to CM noise and less distortion) often outweigh the power penalty.

◀

The above calculations of the gain have assumed an ideal tail current source. In reality, the output impedance of this source affects the gain, but the error is relatively small.

Headroom Issues The five-transistor OTA does not easily lend itself to low-voltage operation as the diode-connected PMOS device tends to consume a substantial voltage headroom. To arrive at a modification, we observe that the gate voltage of this device need not be equal to its drain voltage. As shown in Fig. 5.36, we insert a resistor in series with the gate and draw a constant current from it, thereby

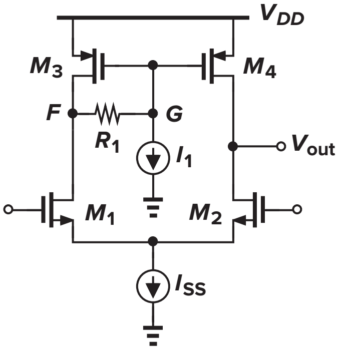


Figure 5.36 OTA headroom improvement by level shift.

allowing V_G to be below V_F by $R_1 I_1 \leq V_{TH3}$. With this level shift, the input CM level can be higher, easing the design of the preceding stage and the tail current source. The value of I_1 must be much less than $I_{SS}/2$ so as to introduce negligible asymmetry between the two sides of the circuit. The reader is encouraged to compute the input-referred offset voltage arising from I_1.

5.3.3 Common-Mode Properties

Let us now study the common-mode properties of the differential pair with active current mirror. We assume $\gamma = 0$ for simplicity and leave a more general analysis including body effect for the reader. Our objective is to predict the consequences of a finite output impedance in the tail current source. As depicted in Fig. 5.37, a change in the input CM level leads to a change in the bias current of all of the transistors. How do we define the common-mode gain here? Recall from Chapter 4 that the CM gain represents the *corruption* of the output signal of interest due to variations in the input CM level. In the circuits of Chapter 3, the output signal was sensed differentially, and hence the CM gain was defined in terms of the output differential component generated by the input CM change. In the circuit of Fig. 5.37, on the other hand, the output signal of interest is sensed with respect to ground. Thus, we define the CM gain in terms of the single-ended output component produced by the input CM change:

$$A_{CM} = \frac{\Delta V_{out}}{\Delta V_{in,CM}} \tag{5.39}$$

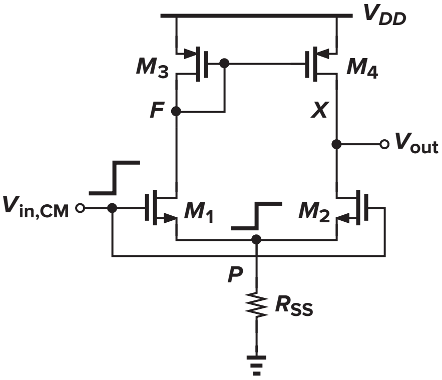


Figure 5.37 Differential pair with active current mirror sensing a common-mode change.

To determine A_{CM}, we observe that if the transistors are symmetric, $V_{out} = V_F$ for any input CM level (Section 5.3.1). For example, as $V_{in,CM}$ increases, V_F drops and so does V_{out}. In other words, nodes F and X can be shorted [Fig. 5.38(a)], resulting in the equivalent circuit shown in Fig. 5.38(b). Here, M_1 and M_2 appear in parallel and so do M_3 and M_4. It follows that

$$A_{CM} \approx \frac{-\dfrac{1}{2g_{m3,4}} \Big\| \dfrac{r_{O3,4}}{2}}{\dfrac{1}{2g_{m1,2}} + R_{SS}} \tag{5.40}$$

$$= \frac{-1}{1 + 2g_{m1,2}R_{SS}} \frac{g_{m1,2}}{g_{m3,4}} \tag{5.41}$$

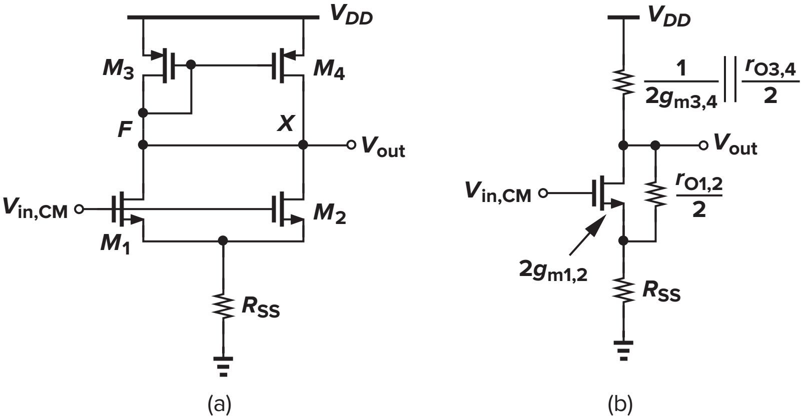


Figure 5.38 (a) Simplified circuit of Fig. 5.37; (b) equivalent circuit of (a).

where we have assumed that $1/(2g_{m3,4}) \ll r_{O3,4}$ and neglected the effect of $r_{O1,2}/2$. The CMRR is then given by

$$\text{CMRR} = \left|\frac{A_{DM}}{A_{CM}}\right| \tag{5.42}$$

$$= g_{m1,2}(r_{O1,2}\|r_{O3,4})\frac{g_{m3,4}(1+2g_{m1,2}R_{SS})}{g_{m1,2}} \tag{5.43}$$

$$= (1+2g_{m1,2}R_{SS})g_{m3,4}(r_{O1,2}\|r_{O3,4}) \tag{5.44}$$

For example, if $R_{SS} = r_O$ and $2g_{m1,2}r_O \gg 1$, then CMRR is on the order of $(g_m r_O)^2$.

Equation (5.41) indicates that, even with perfect symmetry, the output signal is corrupted by input CM variations. High-frequency common-mode noise therefore degrades the performance considerably as the capacitance shunting the tail current source exhibits a lower impedance.

▶ **Example 5.13**

The CM gain of the circuit of Fig. 5.37 can be shown to be *zero* by a (flawed) argument. As shown in Fig. 5.39(a), if $V_{in,CM}$ introduces a change of ΔI in the drain current of each input transistor, then I_{D3} also experiences the same change, and so does I_{D4}. Thus, M_4 seemingly provides the additional current required by M_2, and the output voltage need not change, i.e., $A_{CM} = 0$. Explain the flaw in this proof.

Solution

The assumption that ΔI_{D4} completely cancels the effect of ΔI_{D2} is incorrect. Consider the equivalent circuit shown in Fig. 5.39(b). Since

$$\Delta V_F = \Delta I_1\left(\frac{1}{g_{m3}}\middle\| r_{O3}\right) \tag{5.45}$$

we have

$$|\Delta I_{D4}| = g_{m4}\Delta V_F \tag{5.46}$$

$$= g_{m4}\Delta I_1\frac{r_{O3}}{1+g_{m3}r_{O3}} \tag{5.47}$$

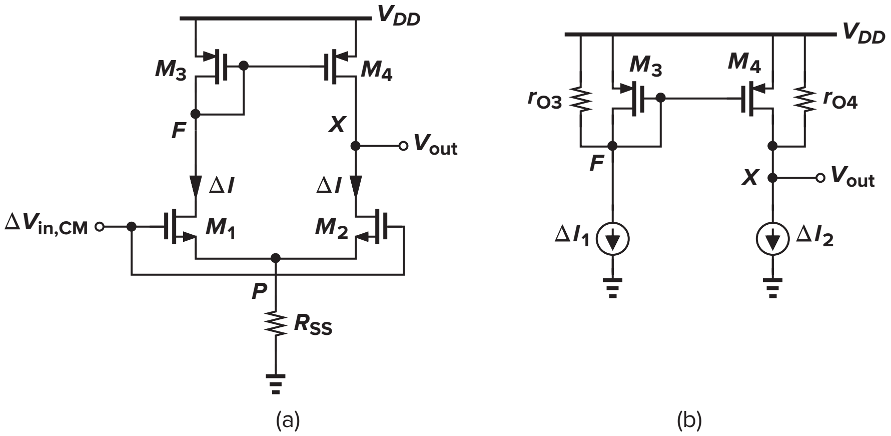


Figure 5.39

This current and ΔI_2 $(= \Delta I_1 = \Delta I)$ give a net voltage change equal to

$$\Delta V_{out} = (\Delta I_1 g_{m4} \frac{r_{O3}}{1 + g_{m3} r_{O3}} - \Delta I_2) r_{O4} \tag{5.48}$$

$$= -\Delta I \frac{1}{g_{m3} r_{O3} + 1} r_{O4} \tag{5.49}$$

which is equal to the voltage change at node F.

Effect of Mismatches It is also instructive to calculate the common-mode gain in the presence of mismatches. As an example, we consider the case where the input transistors exhibit slightly different transconductances [Fig. 5.40(a)]. How does V_{out} depend on $V_{in,CM}$? Since the change at nodes F and X is relatively small, we can compute the change in I_{D1} and I_{D2} while neglecting the effect of r_{O1} and r_{O2}. As shown in Fig. 5.40(b), the voltage change at P can be obtained by considering M_1 and M_2 as a single transistor (in a source follower configuration) with a transconductance equal to $g_{m1} + g_{m2}$, i.e.,

$$\Delta V_P = \Delta V_{in,CM} \frac{R_{SS}}{R_{SS} + \dfrac{1}{g_{m1} + g_{m2}}} \tag{5.50}$$

where body effect is neglected. The changes in the drain currents of M_1 and M_2 are therefore given by

$$\Delta I_{D1} = g_{m1}(\Delta V_{in,CM} - \Delta V_P) \tag{5.51}$$

$$= \frac{\Delta V_{in,CM}}{R_{SS} + \dfrac{1}{g_{m1} + g_{m2}}} \frac{g_{m1}}{g_{m1} + g_{m2}} \tag{5.52}$$

$$\Delta I_{D2} = g_{m2}(\Delta V_{in,CM} - \Delta V_P) \tag{5.53}$$

$$= \frac{\Delta V_{in,CM}}{R_{SS} + \dfrac{1}{g_{m1} + g_{m2}}} \frac{g_{m2}}{g_{m1} + g_{m2}} \tag{5.54}$$

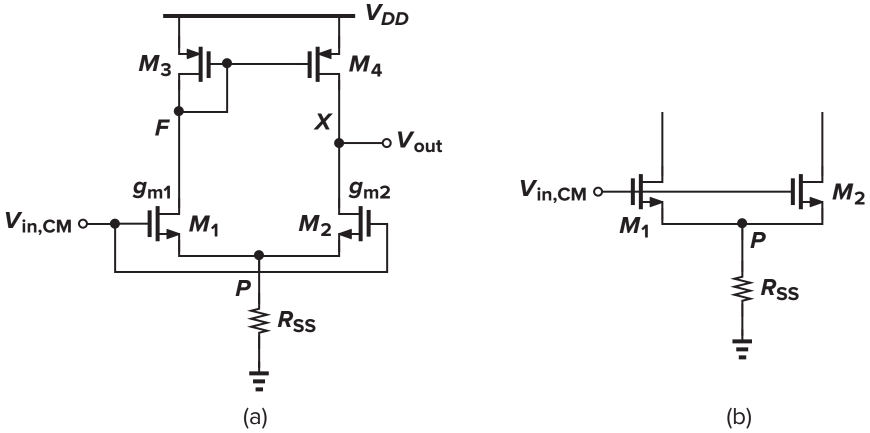


Figure 5.40 Differential pair with g_m mismatch.

The change ΔI_{D1} multiplied by $(1/g_{m3})\|r_{O3}$ yields $|\Delta I_{D4}| = g_{m4}[(1/g_{m3})\|r_{O3}]\Delta I_{D1}$. The difference between this current and ΔI_{D2} flows through the output impedance of the circuit, which is equal to r_{O4} because we have neglected the effect of r_{O1} and r_{O2}:

$$\Delta V_{out} = \left[\frac{g_{m1}\Delta V_{in,CM}}{1+(g_{m1}+g_{m2})R_{SS}}\frac{r_{O3}}{r_{O3}+\dfrac{1}{g_{m3}}} - \frac{g_{m2}\Delta V_{in,CM}}{1+(g_{m1}+g_{m2})R_{SS}}\right] r_{O4} \tag{5.55}$$

$$= \frac{\Delta V_{in,CM}}{1+(g_{m1}+g_{m2})R_{SS}}\frac{(g_{m1}-g_{m2})r_{O3}-g_{m2}/g_{m3}}{r_{O3}+\dfrac{1}{g_{m3}}} r_{O4} \tag{5.56}$$

If $r_{O3} \gg 1/g_{m3}$, we have

$$\frac{\Delta V_{out}}{\Delta V_{in,CM}} \approx \frac{(g_{m1}-g_{m2})r_{O3}-g_{m2}/g_{m3}}{1+(g_{m1}+g_{m2})R_{SS}} \tag{5.57}$$

Compared to Eq. (5.41), this result contains the additional term $(g_{m1}-g_{m2})r_{O3}$ in the numerator, revealing the effect of transconductance mismatch on the common-mode gain.

5.3.4 Other Properties of Five-Transistor OTA

The five-transistor OTA suffers from two drawbacks with respect to the fully-differential topologies studied in Chapter 4. First, the circuit exhibits a finite CMRR even with perfectly-matched transistors. As depicted in Fig. 5.41(a), an input CM change directly corrupts V_{out} in this OTA, but not the *differential* output in the fully-differential version [Fig. 5.41(b)].

Second, the supply rejection of this OTA is inferior. To understand this point, let us tie the inputs to a constant voltage and change V_{DD} by a small amount, ΔV_{DD} [Fig. 5.42(a)]. How much does V_F change? Viewing M_1 as a constant current source with a high output impedance, we recognize that V_{GS3} must remain relatively constant. That is, $\Delta V_F \approx \Delta V_{DD}$. With symmetric transistors, V_{out} must also change by ΔV_{DD}. In other words, the gain from V_{DD} to V_{out} is about unity.

Now consider the fully-differential topology in Fig. 5.42(b), where the PMOS current sources are biased by a current mirror arrangement. How do V_X and V_Y change here in response to a supply change

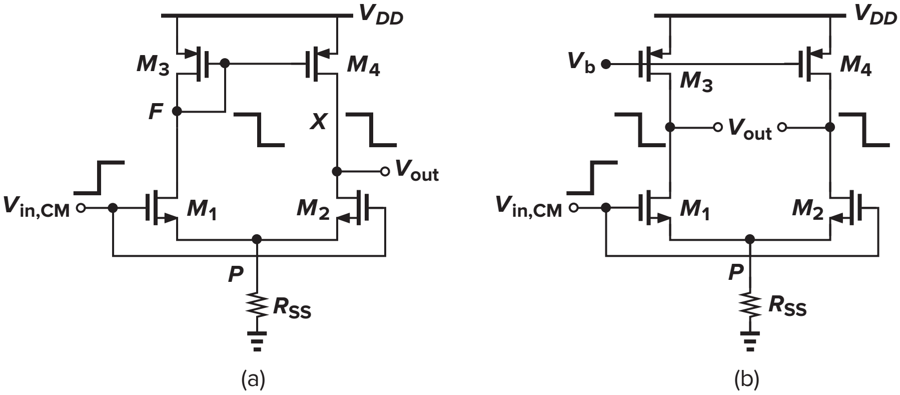


Figure 5.41 Input CM response of (a) five-transistor OTA and (b) fully-differential amplifier with current-source loads.

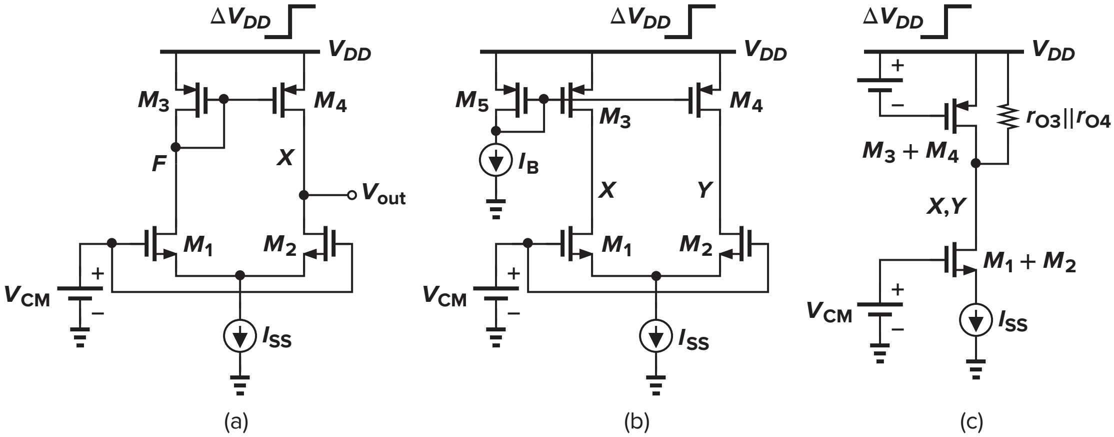


Figure 5.42 (a) OTA with supply step, (b) fully-differential circuit with supply step, and (c) equivalent circuit of (b).

of ΔV_{DD}? We note that V_{GS5} and hence V_{GS3} and V_{GS4} are constant, and, by virtue of symmetry, V_X and V_Y must change by equal amounts. We thus short X and Y and merge M_3 with M_4 and M_1 with M_2 [Fig. 5.42(c)]. If the output impedance of the cascode circuit consisting of $M_1 + M_2$ and I_{SS} is very high, then $\Delta V_X = \Delta V_Y \approx \Delta V_{DD}$ (why?). In this case, too, the output voltages change by ΔV_{DD}, but their *difference* remains intact. We should caution the reader that this circuit requires common-mode feedback (Chapter 9).

5.4 ■ Biasing Techniques

The amplifier stages studied thus far must be properly biased so that, in the absence of the input signal, each transistor carries the required current and sustains the necessary terminal voltages. We recognize that the current establishes the transconductance and output resistance of the transistor while the terminal voltages determine the headroom and hence the allowable voltage swings. In this section, we consider a number of biasing techniques for CMOS circuits.

5.4.1 CS Biasing

Simple CS Stage We wish to create a certain drain current and desired V_{GS} and V_{DS} for a transistor in a CS configuration. Using the transistor's I/V characteristics, we have determined its dimensions and must now tie the gate to a proper bias voltage [Fig. 5.43(a)]. But how do we ensure that V_B does not "fight" V_{in}? One solution is to couple V_{in} capacitively and establish a *high* impedance for V_B so that X has the same dc voltage as V_B and the same signal voltage as V_{in} [Fig. 5.43(b)]. Noting that C_B and R_B form a high-pass filter, we select $1/(2\pi R_B C_B)$ lower than the *lowest* input frequency so that the ac gain from V_{in} to V_X is near unity in the frequency range of interest.

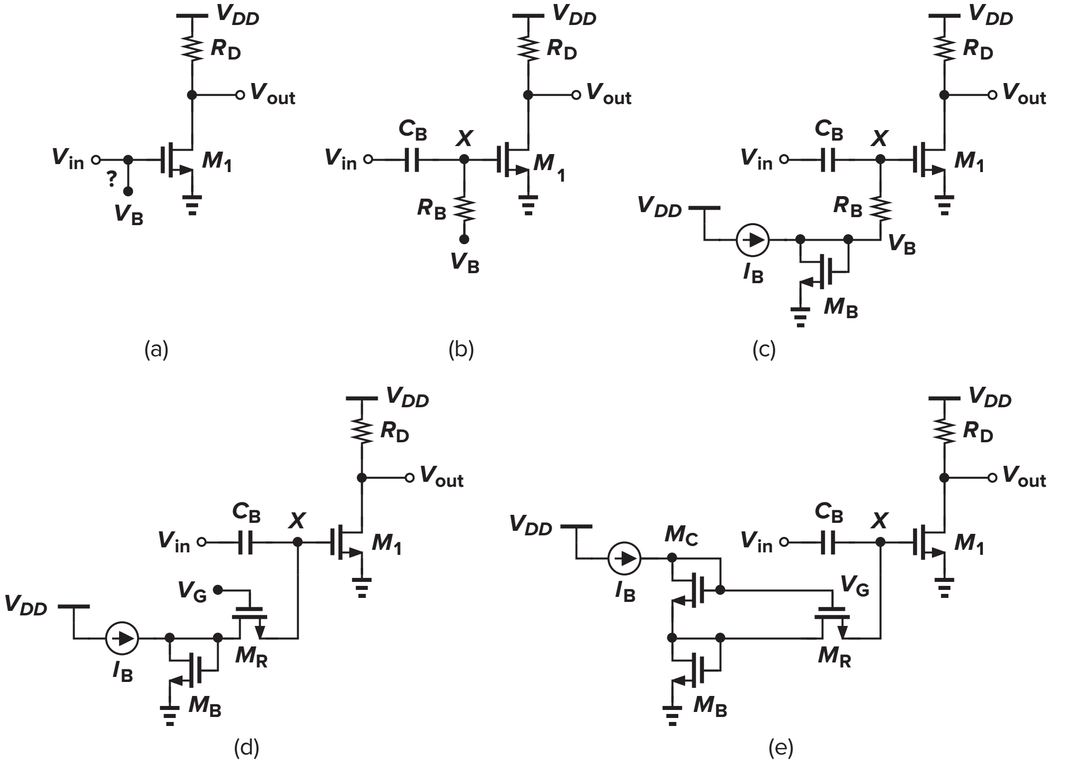


Figure 5.43 CS stage biasing with (a) V_B fighting V_{in}, (b) ac coupling to set the dc value of V_X to V_B, (c) use of a current mirror, (d) a large resistor realized by M_R, and (e) accurate V_{GS} generation for M_R.

We now make several remarks. (1) Node X in Fig. 5.43(b) *must* have a dc path to a voltage; if R_B is removed, X floats, sustaining a poorly-defined voltage.[5] (2) As explained in Sec. 5.1, the bias voltage, V_B, must *not* be constant; rather, it must be generated by a diode-connected device [Fig. 5.43(c)]. (3) We typically select I_B about one-tenth to one-fifth of I_{D1} so as to minimize the power consumed by the bias network. (4) The capacitor and the resistor may occupy a large chip area if V_{in} contains very low frequencies, e.g., in the audio range. (5) The capacitor introduces its own parasitics in the signal path (Chapter 19), degrading the high-frequency performance; even if chip area is not critical, the value of the capacitor is limited by these parasitic effects.

In applications requiring a large $R_B C_B$ product, one can replace R_B with a long, narrow MOSFET operating in the deep triode region and bias this transistor with a small overdrive voltage, thus maximizing

[5] In reality, the gate leakage current of M_1 would discharge X to zero.

its on-resistance [Fig. 5.43(d)]. But how do we guarantee that M_R does not turn *off* with PVT variations? While small, the overdrive of M_R must still be well-controlled, i.e., $V_G - V_B$ must still be around V_{TH}. This difference can be created by means of a diode-connected device [Fig. 5.43(e)]. If $(W/L)_C$ is large, $V_{GS,C} \approx V_{TH}$, producing a high resistance in M_R. Using a long-channel model, the reader can prove that, in strong inversion,

$$R_{on,R} = \frac{(W/L)_C}{(W/L)_R}\frac{1}{g_{m,C}} \tag{5.58}$$

We conclude that $(W/L)_C$ must be maximized and $(W/L)_R$ minimized. In Problem 5.24, we reexamine the circuit in the subthreshold region.

Is it possible to remove the input coupling capacitor and provide the bias voltage from the preceding stage? Figure 5.44 depicts an example, where $V_{DD} - R_{D2}I_{D2}$ serves as the bias gate voltage of M_1. The principal difficulty here is that the bias conditions of M_1 are influenced by those of M_2. For example, if I_{D2} varies with PVT variations, so do V_X and hence I_{D1}. In such a cascade, the PVT variations are *amplified* because they are indistinguishable from the signal. Nonetheless, one can employ direct coupling between two stages if each has a low gain, e.g., around 2 or 3. For a larger number of stages or higher gains, negative feedback may become necessary, especilly if the load resistors are replaced with current sources (Chapter 8).

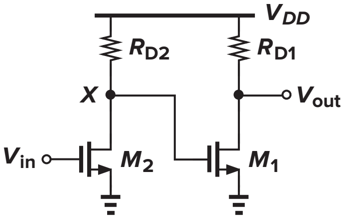


Figure 5.44 Direct coupling between two stages.

CS Stage with Current-Source Load We now turn our attention to the common-source stage with current-source load [Fig. 5.45(a)]. The foregoing techniques can be readily applied to both M_1 and M_2, yielding the circuit shown in Fig. 5.45(b). We note that I_{D1} and I_{D2} are copied from I_{B1} and I_{B2}, respectively.

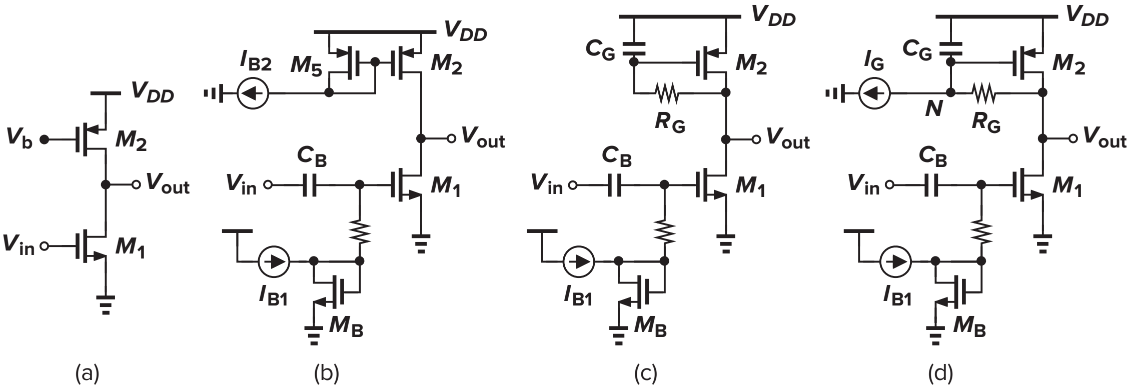


Figure 5.45 (a) CS stage with current-source load; (b) simple biasing; (c) self-biasing of current source; (d) use of I_G to shift the output.

The CS stage with current-source load exemplifies a situation sometimes encountered in analog design: two high-impedance current sources, M_1 and M_2, fight each other. That is, if the copied currents in Fig. 5.45(b) are not exactly equal, each transistor wants to impose its own current. (Imagine what happens if two unequal ideal current sources are placed in series.) For example, if I_{D1} tends to be greater than $|I_{D2}|$, then V_{out} falls—possibly driving M_1 into the triode region—until I_{D1} becomes equal to $|I_{D2}|$. To resolve this issue, we modify the circuit as shown in Fig. 5.45(c), where M_2 acts as a diode-connected device at dc, happily carrying the current imposed by M_1. At high frequencies, C_G shorts the gate of M_2 to ground, yielding a small-signal gain equal to

$$A_v = -g_{m1}(r_{O1}||r_{O2}||R_G) \tag{5.59}$$

We therefore select $R_G \gg r_{O1}||r_{O2}$ and $1/(2\pi R_G C_G)$ less than the lowest signal frequency of interest.

In the above CS stage, M_2 forces the bias value of V_{out} to be as low as $V_{DD} - |V_{GS2}|$. We can draw a constant current of I_G from R_G [Fig. 5.45(d)] so that V_N is still low enough to provide the necessary V_{GS} for M_2, but $V_{out} = V_N + I_G R_G$ is higher. The value of I_G is chosen much less than the bias current.

▶ **Example 5.14**

Compare the maximum allowable voltage swings in Figs. 5.45(c) and (d).

Solution

In Fig. 5.45(c), V_{out} begins from $V_{DD} - |V_{GS2}|$ and can rise to $V_{DD} - |V_{GS2} - V_{TH2}|$ and fall to $V_{GS1} - V_{TH1}$. However, as illustrated in Fig. 5.46(a), since the down swing is limited to $V_{DD} - |V_{GS2}| - (V_{GS1} - V_{TH1})$, the up swing cannot reach its maximum. Thus, the allowable peak-to-peak swing is about $2[V_{DD} - |V_{GS2}| - (V_{GS1} - V_{TH1})]$.

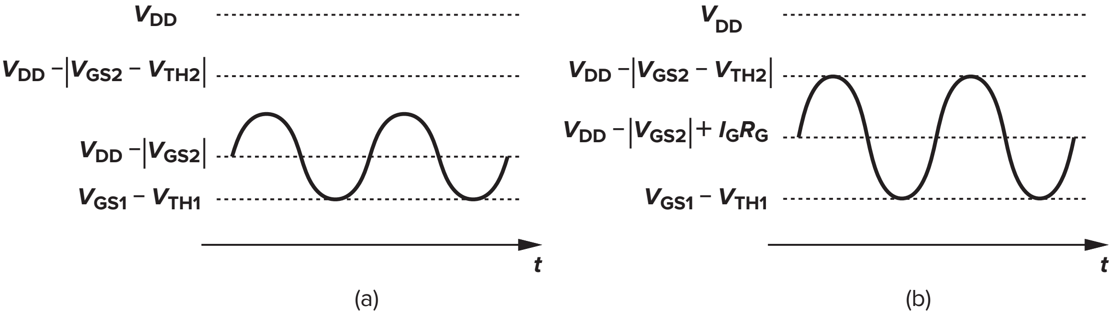


Figure 5.46

In Fig. 5.45(d), on the other hand, $I_G R_G$ can shift the operating point such that the down swing and the up swing are approximately equal. From Fig. 5.46(b), we have

$$V_{DD} - |V_{GS2}| + I_G R_G - (V_{GS1} - V_{TH1}) \approx V_{DD} - |V_{GS2} - V_{TH2}| - [V_{DD} - |V_{GS2}| + I_G R_G] \tag{5.60}$$

If the NMOS and PMOS overdrives are roughly equal, we must choose

$$I_G R_G \approx |V_{GS2}| - \frac{V_{DD}}{2} \tag{5.61}$$

in which case the output peak-to-peak swing can reach $2[V_{DD}/2 - (V_{GS1} - V_{TH1})]$. Alternatively, we can choose $|V_{GS2}| = V_{DD}/2$ and apply no I_G. As explained in Chapter 7, M_2 contributes less noise as its overdrive increases (while its bias current remains constant), making the former topology more attractive. ◀

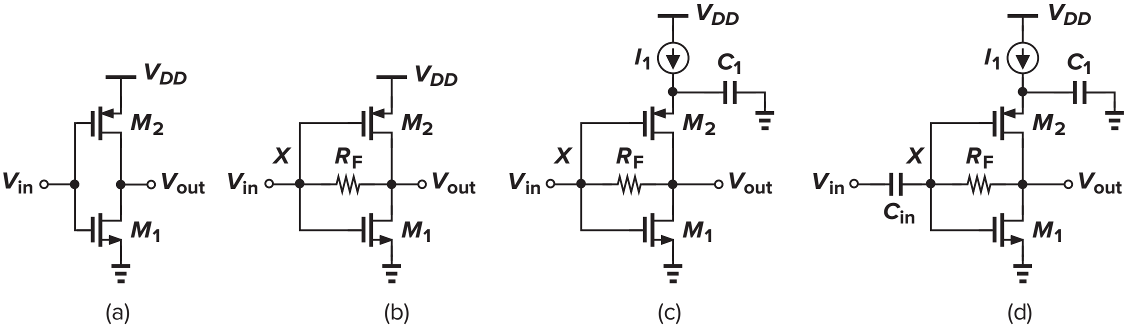


Figure 5.47 (a) Complementary CS stage, (b) self-biased topology, (c) accurate definition of bias current, and (d) use of ac coupling at input.

Complementary CS Stage Let us now consider the problem of biasing for the CS stage with active current source [Fig. 5.47(a)]. As explained in Chapter 3, this topology exhibits considerable PVT dependence because $V_{GS1} + |V_{GS2}| = V_{DD}$. Also, in a manner similar to the CS stage of Fig. 5.45(b), M_1 and M_2 fight each other.

As a first step, consider the arrangement shown in Fig. 5.47(b), where a large resistor is tied between the drains and gates of the transistors. In the absence of signals, no current flows through R_F and $V_{out} = V_X$; in essence, each transistor is configured as a diode-connected device and guaranteed to operate in saturation. The two devices therefore do not fight anymore: if for example, M_1 tends to carry a larger drain current, then V_{out} and hence V_X fall so that $I_{D1} = |I_{D2}|$.

To define the bias current accurately, we modify the circuit as shown in Fig. 5.47(c). Here, I_1 establishes the drain currents of M_1 and M_2, and C_1 creates a short circuit at the lowest signal frequency of interest, ω_{min}. The value of C_1 is chosen such that M_2 experiences negligible degeneration:

$$\frac{1}{C_1\omega_{min}} \ll \frac{1}{g_{m2}} \tag{5.62}$$

Note that I_1 consumes additional voltage headroom in this case.

Since the bias voltage at node X must track V_{out}, the input must be capacitively coupled [Fig. 5.47(d)]. In Problem 5.25, we compute the corner frequency of the high-pass filter formed by C_{in} and the remainder of the circuit. This frequency must be chosen lower than ω_{min}. With sufficiently large values for C_{in}, R_F, and C_1, the voltage gain of the amplifier at signal frequencies of interest is still given by $(g_{m1} + g_{m2})(r_{O1}||r_{O2})$.

5.4.2 CG Biasing

In a common-gate stage, the transistor must carry a bias current while sensing the input at its source terminal. Thus, the source cannot be directly tied to the ground, requiring an intervening element that passes dc, e.g., a resistor, a current source, or an inductor. Figure 5.48(a) depicts an example where M_1 and M_B form a current mirror so that I_{D1} is a multiple of I_B. For proper copying of I_B, we must ensure that $V_{GS1} = V_{GS,B}$. We therefore choose $(W/L)_1/(W/L)_B$ equal to the desired ratio for I_{D1} and I_B (e.g., in the range of 5 to 10) and apply the same ratio to R_B/R_S, i.e., $R_B/R_S = I_{D1}/I_B$.

The circuit of Fig. 5.48(a) faces difficulties in low-voltage design. In the presence of a finite driving impedance, R_1 (i.e., the output impedance of the preceding stage), the signal experiences additional attenuation due to R_S. Neglecting channel-length modulation, we write the voltage gain from V_{in}

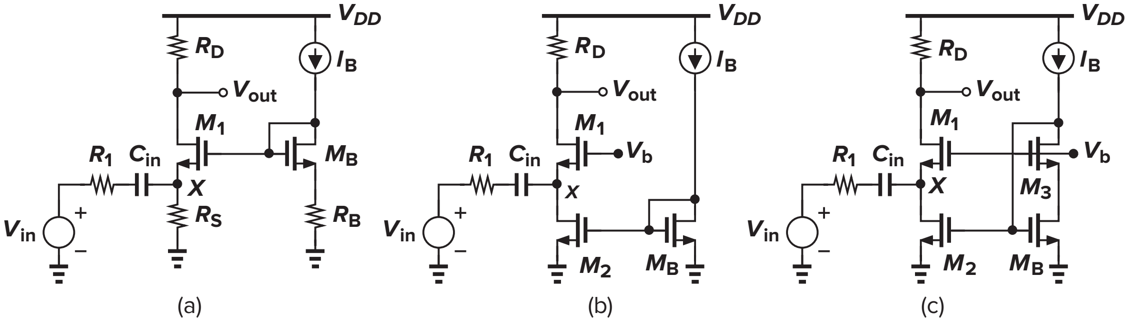


Figure 5.48 CG stage with (a) resistive path from source to ground, (b) current-source biasing, and (c) low-voltage current mirror.

to V_X as

$$\frac{V_X}{V_{in}} = \frac{\dfrac{1}{g_{m1}+g_{mb1}}||R_S}{\dfrac{1}{g_{m1}+g_{mb1}}||R_S + R_1} \tag{5.63}$$

concluding that R_S must be much greater than $1/(g_{m1} + g_{mb1})$ to minimize this attenuation. However, since the gain from V_X to V_{out} is equal to $(g_{m1} + g_{mb1})R_D$, this means that R_S may reach or even exceed R_D. Thus, R_S may sustain a *large* dc voltage drop, limiting the dc drop across R_D and hence the voltage gain.

To remedy the situation, we replace R_S with a current source [Fig. 5.48(b)]. Here, M_2 exhibits a high impedance but does not necessarily require a high V_{DS}. Copied from I_B, the drain current of M_2 does incur some error due to channel-length modulation because $V_{DS2} < V_{DS,B}$. This issue is reminiscent of the cascode current mirror studied in Section 5.2 and can be resolved by means of a low-voltage cascode topology [Fig. 5.48(c)]. The bias voltage, V_b, is also generated as explained in Section 5.2.

5.4.3 Source Follower Biasing

Source followers are typically biased by means of a current source as shown in Fig. 5.49(a). If the mismatch between I_{D2} and I_B due to channel-length modulation proves undesirable, a resistor can be placed in series with the drain of M_B (Sec. 5.2). Defined by M_2, the bias current of M_1 is less sensitive to its gate voltage than in a CS amplifier, allowing direct connection to the preceding stage. In applications where the input dc voltage may vary considerably, capacitive coupling can be used [Fig. 5.49(b)]. Note that the gate voltage of M_1 begins at V_{DD} and can swing up by one threshold before the transistor enters the triode region.

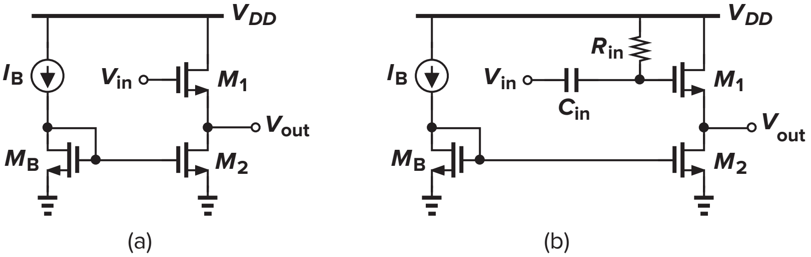


Figure 5.49 Source follower biasing with (a) current source and (b) ac coupling at input.

▶ **Example 5.15**

A source follower serves as an output buffer for a CS stage. Study the performance with and without capacitive coupling between the two stages.

Solution

In Fig. 5.50(a), the minimum drain voltage of M_3 is given by $V_{GS1} + V_{DS2,min}$, leaving little for the allowable voltage drop across R_D. The CS voltage gain is therefore severely limited. In the circuit of Fig. 5.50(b), on the other hand, the first stage's gain can be independently maximized.

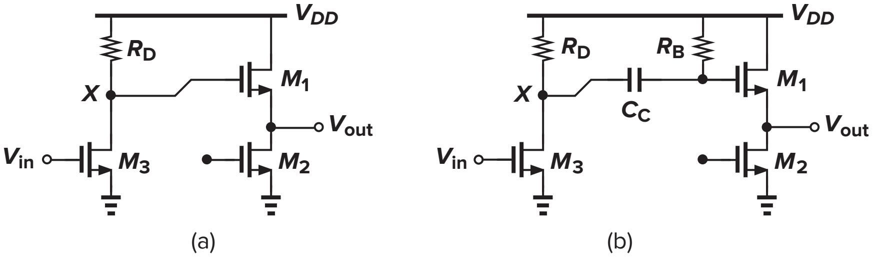


Figure 5.50

◀

5.4.4 Differential Pair Biasing

In addition to the tail current source, the gate voltage of a differential pair must also be defined. To maximize the voltage gain and/or output swings, we select the *lowest* input CM level, as shown in Fig. 5.51(a), equal to $V_{GS1,2} + V_{DS3,min}$. This choice allows the drain voltages of M_1 and M_2 to be as low as $(V_{GS1,2} - V_{TH1,2}) + V_{DS3,min}$ (two overdrive voltages above ground) and hence maximum R_D.

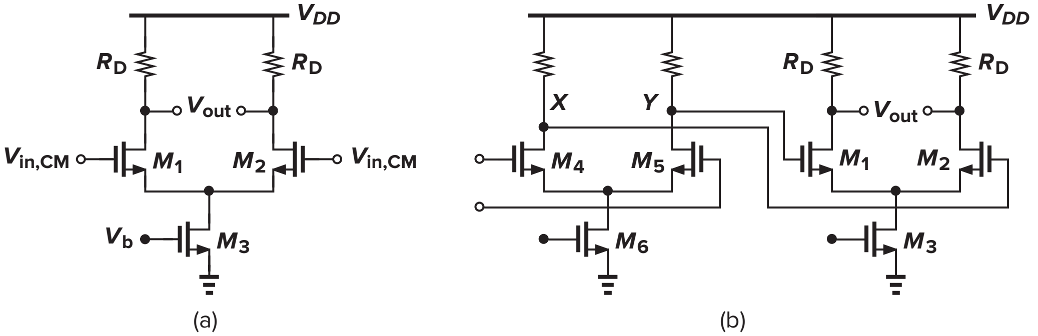


Figure 5.51 (a) Choice of input CM level for a differential pair, and (b) cascaded pairs.

Since the bias currents of M_1 and M_2 in Fig. 5.51(a) are relatively insensitive to their gate voltages, we can directly connect their gates to the preceding stage [Fig. 5.51(b)]. This approach, however, constrains the overall voltage gain: if the bias value of V_X and V_Y is chosen equal to two overdrives above ground so as to maximize the gain of the first stage, then it is an excessively low common-mode level for the *second* stage (why?). For this reason, we may resort to capacitive coupling in some cases.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. All device dimensions are effective values and in microns.

5.1. In Fig. 5.2, assume that $(W/L)_1 = 50/0.5$, $\lambda = 0$, $I_{out} = 0.5$ mA, and M_1 is saturated.
(a) Determine R_2/R_1.
(b) Calculate the sensitivity of I_{out} to V_{DD}, defined as $\partial I_{out}/\partial V_{DD}$ and normalized to I_{out}.
(c) How much does I_{out} change if V_{TH} changes by 50 mV?
(d) If the temperature dependence of μ_n is expressed as $\mu_n \propto T^{-3/2}$ but V_{TH} is independent of temperature, how much does I_{out} vary if T changes from 300 °K to 370 °K?
(e) What is the worst-case change in I_{out} if V_{DD} changes by 10%, V_{TH} changes by 50 mV, and T changes from 300 °K to 370 °K?

5.2. Consider the circuit of Fig. 5.7. Assuming I_{REF} is *ideal*, sketch I_{out} versus V_{DD} as V_{DD} varies from 0 to 3 V.

5.3. In the circuit of Fig. 5.8, $(W/L)_N = 10/0.5$, $(W/L)_P = 10/0.5$, and $I_{REF} = 100\ \mu$A. The input CM level applied to the gates of M_1 and M_2 is equal to 1.3 V.
(a) Assuming $\lambda = 0$, calculate V_P and the drain voltage of the PMOS diode-connected transistors.
(b) Now take channel-length modulation into account to determine I_T and the drain current of the PMOS diode-connected transistors more accurately.

5.4. In the circuit of Fig. 5.11, sketch V_{out} versus V_{DD} as V_{DD} varies from 0 to 3 V.

5.5. Consider the circuit of Fig. 5.12(a), assuming $(W/L)_{1-3} = 40/0.5$, $I_{REF} = 0.3$ mA, and $\gamma = 0$.
(a) Determine V_b such that $V_X = V_Y$.
(b) If V_b deviates from the value calculated in part (a) by 100 mV, what is the mismatch between I_{out} and I_{REF}?
(c) If the circuit fed by the cascode current source changes V_P by 1 V, how much does V_Y change?

5.6. The circuit of Fig. 5.18(b) is designed with $(W/L)_{1,2} = 20/0.5$, $(W/L)_{3,0} = 60/0.5$, and $I_{REF} = 100\ \mu$A.
(a) Determine V_X and the acceptable range of V_b.
(b) Estimate the deviation of I_{out} from 300 μA if the drain voltage of M_3 is higher than V_X by 1 V.

5.7. The circuit of Fig. 5.23(a) is designed with $(W/L)_{1-4} = 50/0.5$ and $I_{SS} = 2I_1 = 0.5$ mA.
(a) Calculate the small-signal voltage gain.
(b) Determine the maximum output voltage swing if the input CM level is 1.3 V.

5.8. Consider the circuit of Fig. 5.29(a) with $(W/L)_{1-5} = 50/0.5$ and $I_{D5} = 0.5$ mA.
(a) Calculate the deviation of V_{out} from V_F if $|V_{TH3}|$ is 1 mV less than $|V_{TH4}|$.
(b) Determine the CMRR of the amplifier.

5.9. Sketch V_X and V_Y as a function of V_{DD} for each circuit in Fig. 5.52. Assume the transistors in each circuit are identical.

5.10. Sketch V_X and V_Y as a function of V_{DD} for each circuit in Fig. 5.53. Assume the transistors in each circuit are identical.

5.11. For each circuit in Fig. 5.54, sketch V_X and V_Y as a function of V_1 for $0 < V_1 < V_{DD}$. Assume the transistors in each circuit are identical.

5.12. For each circuit in Fig. 5.55, sketch V_X and V_Y as a function of V_1 for $0 < V_1 < V_{DD}$. Assume the transistors in each circuit are identical.

5.13. For each circuit in Fig. 5.56, sketch V_X and V_Y as a function of I_{REF}.

5.14. For the circuit of Fig. 5.57, sketch I_{out}, V_X, V_A, and V_B as a function of **(a)** I_{REF}, **(b)** V_b.

5.15. In the circuit shown in Fig. 5.58, a source follower using a wide transistor and a small bias current is inserted in series with the gate of M_3 so as to bias M_2 at the edge of saturation. Assuming M_0–M_3 are identical and $\lambda \neq 0$, estimate the mismatch between I_{out} and I_{REF} if **(a)** $\gamma = 0$, **(b)** $\gamma \neq 0$.

5.16. Sketch V_X and V_Y as a function of time for each circuit in Fig. 5.59. Assume the transistors in each circuit are identical.

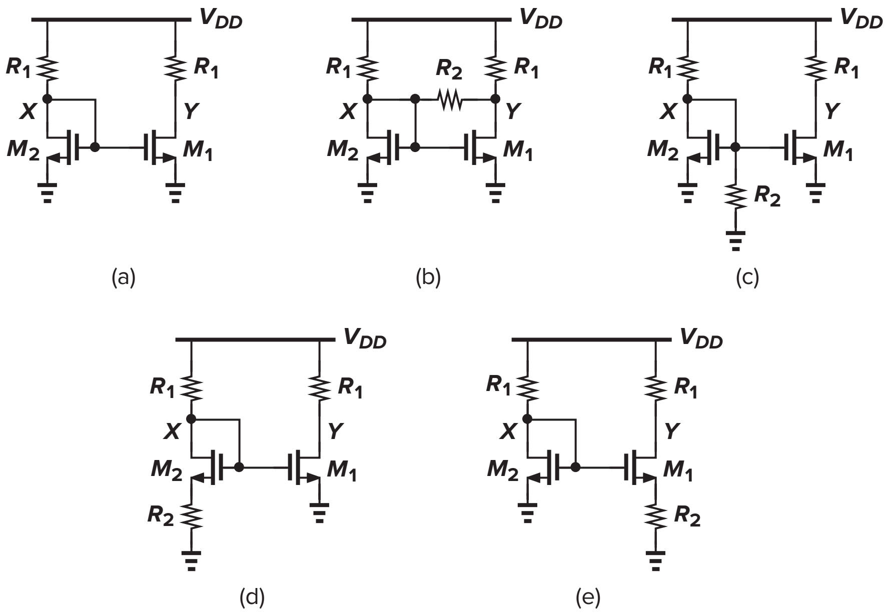


Figure 5.52

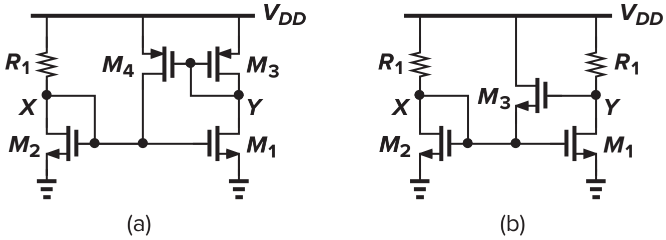


Figure 5.53

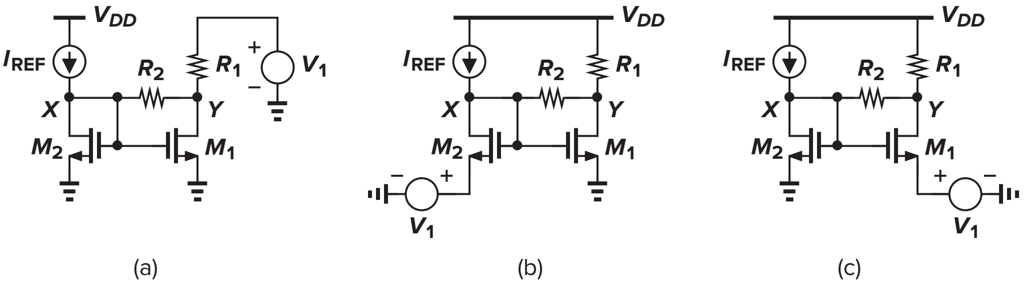


Figure 5.54

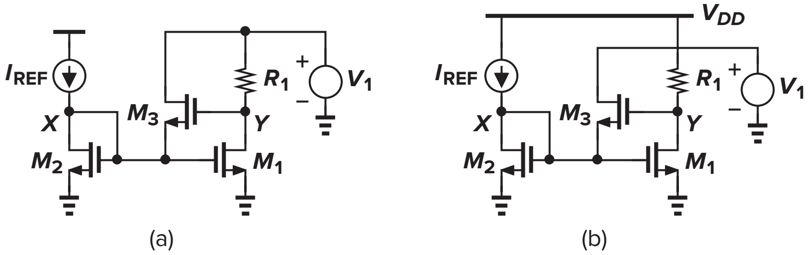


Figure 5.55

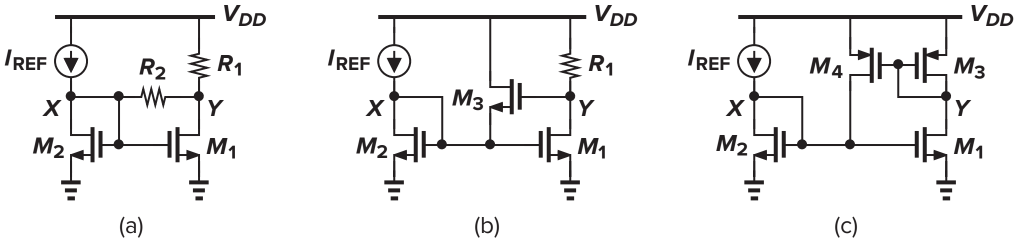


Figure 5.56

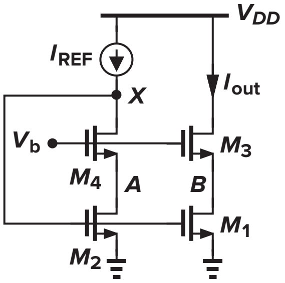


Figure 5.57

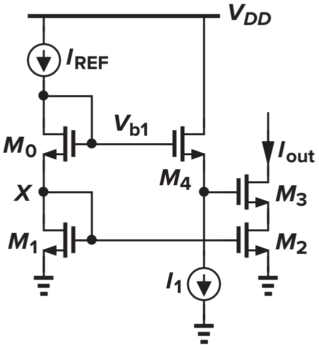


Figure 5.58

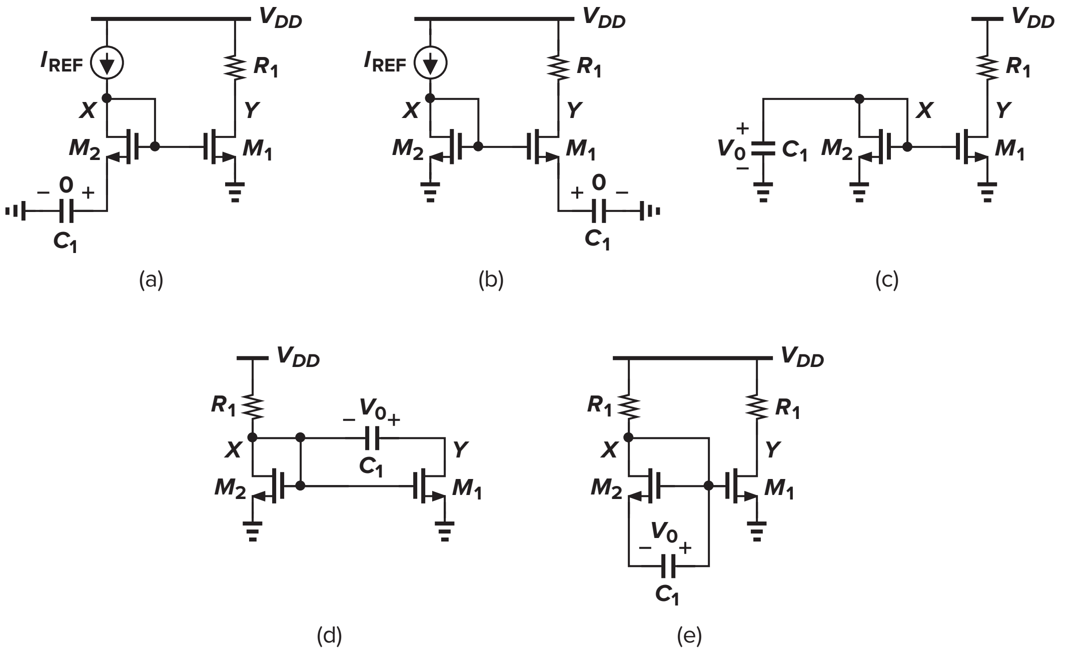


Figure 5.59

5.17. Sketch V_X and V_Y as a function of time for each circuit in Fig. 5.60. Assume the transistors in each circuit are identical.

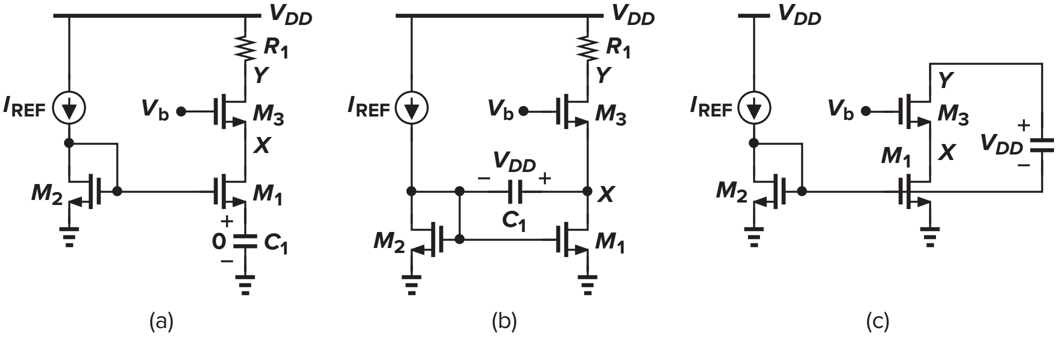


Figure 5.60

5.18. Sketch V_X and V_Y as a function of time for each circuit in Fig. 5.61. Assume the transistors in each circuit are identical.

5.19. The circuit shown in Fig. 5.62 exhibits a *negative* input inductance. Calculate the input impedance of the circuit and identify the inductive component.

5.20. Due to a manufacturing defect, a large parasitic resistance, R_1, has appeared in the circuits of Fig. 5.63. Calculate the gain of each circuit if $\lambda > 0$.

5.21. In digital circuits such as memories, a differential pair with an active current mirror is used to convert a small differential signal to a large single-ended swing (Fig. 5.64). In such applications, it is desirable that the output levels be as close to the supply rails as possible. Assuming moderate differential input swings (e.g., $\Delta V = 0.1$ V) around a common-mode level $V_{in,CM}$ and a high gain in the circuit, explain why V_{min} depends on $V_{in,CM}$.

5.22. Sketch V_X and V_Y for each circuit in Fig. 5.65 as a function of time. The initial voltage across C_1 is shown.

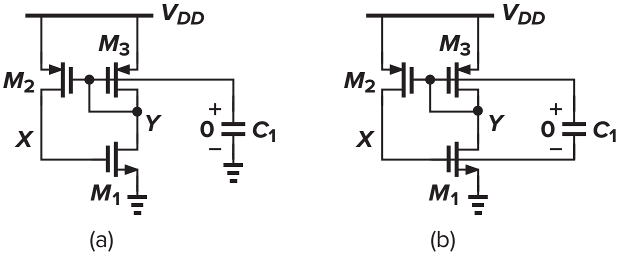


Figure 5.61

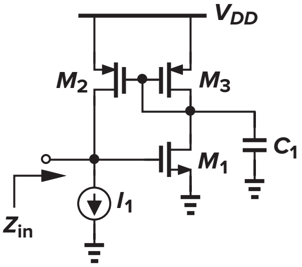


Figure 5.62

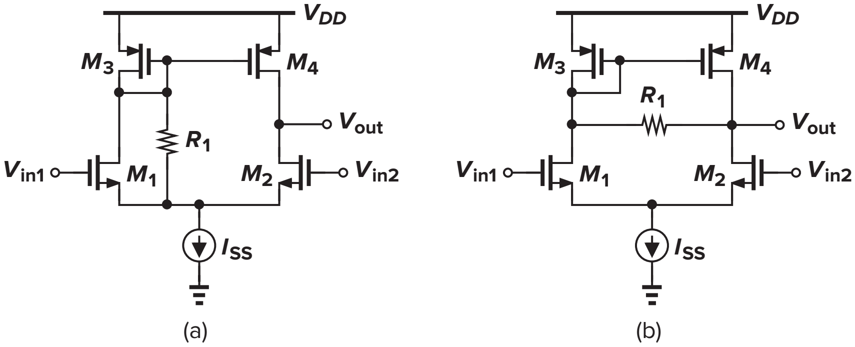


Figure 5.63

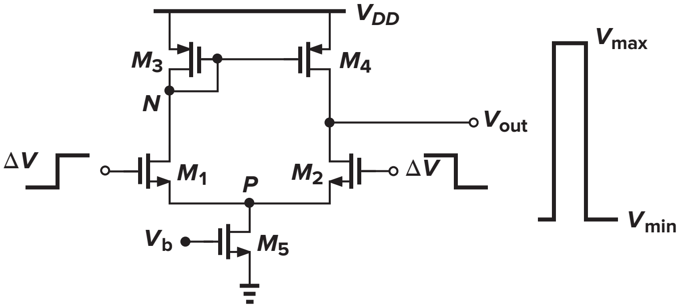


Figure 5.64

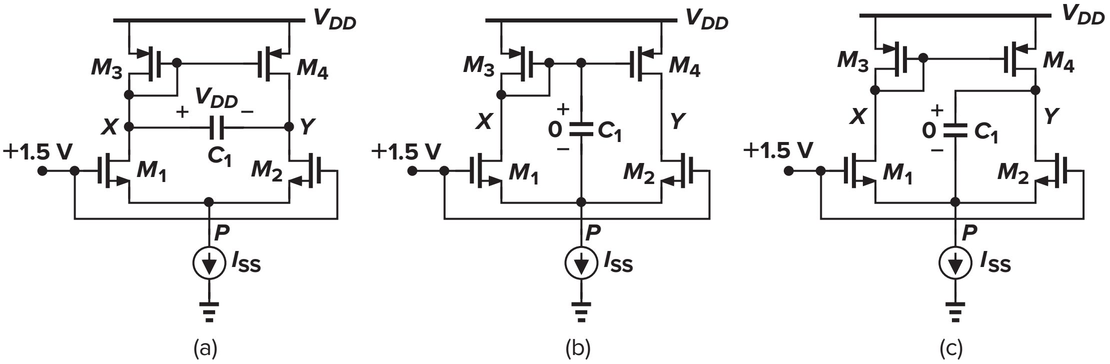


Figure 5.65

5.23. If in Fig. 5.66, ΔV is small enough that all of the transistors remain in saturation, determine the time constant and the initial and final values of V_{out}.

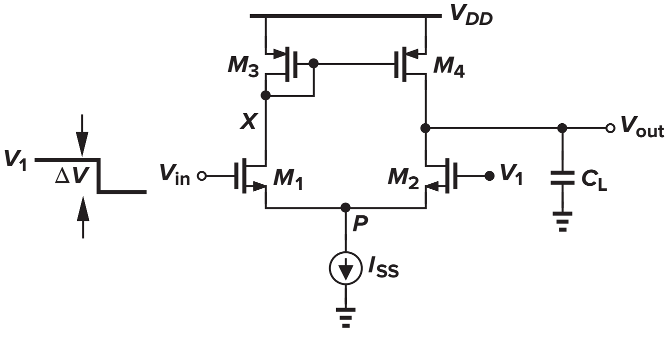


Figure 5.66

5.24. For a device operating in the subthreshold region, we have

$$I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS} - V_{TH}}{V_T} \right) \left(1 - \exp \frac{-V_{DS}}{V_T} \right) \tag{5.64}$$

(a) If the device is in the deep triode region, $V_{DS} \ll V_T$. Using $\exp(-\epsilon) \approx 1 - \epsilon$, determine the on-resistance.

(b) If the device is in saturation, $V_{DS} \gg V_T$. Compute the transconductance.

(c) Find the relation between $g_{m,B}$ and $R_{on,R}$ in Fig. 5.43(d) using the above results.

5.25. Determine the corner frequency resulting from C_{in} in Fig. 5.47(d). For simplicity, assume C_1 is a short circuit.

5.26. Determine the supply rejection of the circuit shown in Fig. 5.67.

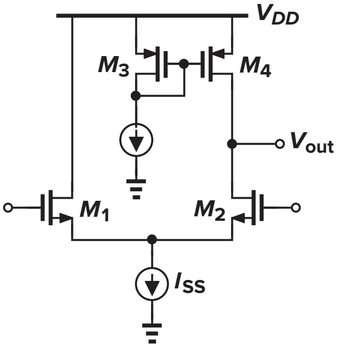


Figure 5.67

CHAPTER

6

Frequency Response of Amplifiers

Our analysis of simple amplifiers has thus far focused on low-frequency characteristics, neglecting the effect of device and load capacitances. In most analog circuits, however, the speed trades with many other parameters such as gain, power dissipation, and noise. It is therefore necessary to understand the frequency-response limitations of each circuit.

In this chapter, we study the behavior of single-stage and differential amplifiers in the frequency domain. Following a review of basic concepts, we analyze the high-frequency response of common-source and common-gate stages and source followers. Next, we deal with cascode and differential amplifiers. Finally, we consider the effect of active current mirrors on the frequency response of differential pairs.

6.1 ■ General Considerations

Recall that a MOS device exhibits four capacitances: C_{GS}, C_{GD}, C_{DB}, and C_{SB}. For this reason, the transfer function of CMOS circuits can rapidly become complicated, calling for approximations that simplify the circuit. In this section, we introduce two such approximations, namely, Miller's theorem and association of poles with nodes. We remind the reader that a two-terminal impedance, Z, is defined as $Z = V/I$, where V and I denote the voltage across and the current flowing through the device. For example, $Z = 1/(Cs)$ for a capacitor. Also, the transfer function of a circuit yields the frequency response if we replace s with $j\omega$, i.e., if we assume a sinusoidal input such as $A\cos\omega t$. For example, $H(j\omega) = (RCj\omega + 1)^{-1}$ provides the magnitude and phase of a simple low-pass filter.

In this chapter, we are primarily interested in the *magnitude* of the transfer function (with $s = j\omega$). Figure 6.1 shows examples of magnitude response. We should also remark that, even if computed exactly, some transfer functions do not offer much insight. We therefore study numerous special cases by considering extreme conditions, e.g., if the load capacitance is very small or very large.

A few basic concepts are used extensively throughout this chapter and merit a brief review. (1) The magnitude of a complex number $a + jb$ is given by $\sqrt{a^2 + b^2}$. (2) Zeros and poles are respectively defined as the roots of the numerator and denominator of the transfer function. (3) According to Bode's approximations, the slope of the magnitude of a transfer function increases by 20 dB/decade as ω passes a pole frequency and decreases by 20 dB/decade as ω passes a zero frequency.

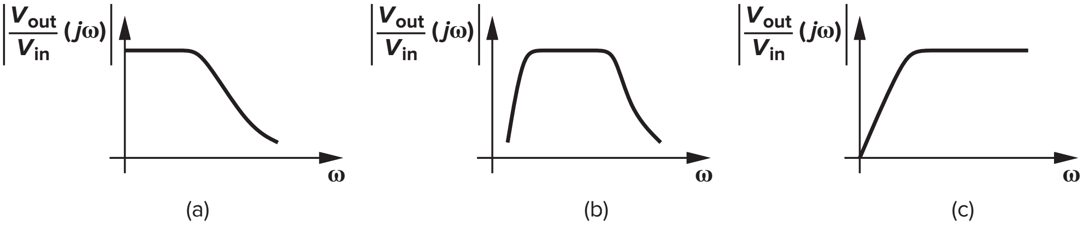


Figure 6.1 (a) Low-pass, (b) band-pass, and (c) high-pass frequency-response examples.

6.1.1 Miller Effect

An important phenomenon that occurs in many analog (and digital) circuits is related to the "Miller effect," as described by Miller in a theorem.

Miller's Theorem If the circuit of Fig. 6.2(a) can be converted to that of Fig. 6.2(b), then $Z_1 = Z/(1 - A_v)$ and $Z_2 = Z/(1 - A_v^{-1})$, where $A_v = V_Y/V_X$.

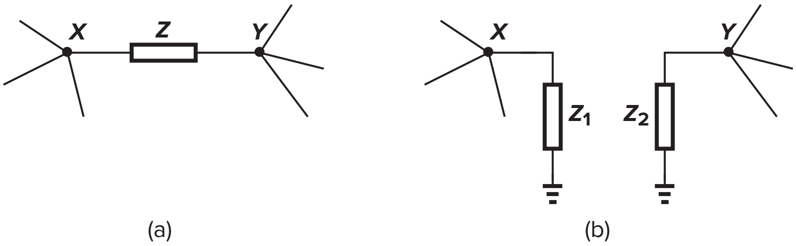


Figure 6.2 Application of Miller effect to a floating impedance.

Proof The current flowing through Z from X to Y is equal to $(V_X - V_Y)/Z$. For the two circuits to be equivalent, the same current must flow through Z_1. Thus,

$$\frac{V_X - V_Y}{Z} = \frac{V_X}{Z_1} \tag{6.1}$$

that is

$$Z_1 = \frac{Z}{1 - \dfrac{V_Y}{V_X}} \tag{6.2}$$

Similarly,

$$Z_2 = \frac{Z}{1 - \dfrac{V_X}{V_Y}} \tag{6.3}$$

This decomposition of a "floating" impedance, Z, into two "grounded" impedances proves useful in analysis and design.

▶ **Example 6.1**

Consider the circuit shown in Fig. 6.3(a), where the voltage amplifier has a negative gain equal to $-A$ and is otherwise ideal. Calculate the input capacitance of the circuit.

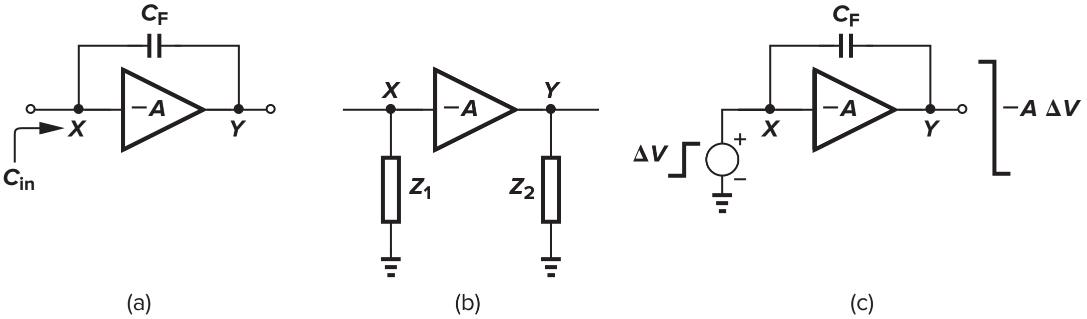


Figure 6.3

Solution

Using Miller's theorem to convert the circuit to that shown in Fig. 6.3(b), we have $Z = 1/(C_F s)$ and $Z_1 = [1/(C_F s)]/(1 + A)$. That is, the input capacitance is equal to $C_F(1 + A)$. We call this effect "Miller multiplication" of the capacitor.

Why is C_F multiplied by $1+A$? Suppose, as depicted in Fig. 6.3(c), we measure the input capacitance by applying a voltage step at the input and calculating the charge supplied by the voltage source. A step equal to ΔV at X results in a change of $-A\Delta V$ at Y, yielding a total change of $(1 + A)\Delta V$ in the voltage across C_F. Thus, the charge drawn by C_F from V_{in} is equal to $(1 + A)C_F\Delta V$ and the equivalent input capacitance equal to $(1 + A)C_F$.

▶ **Example 6.2**

A student needs a large capacitor for a filter and decides to utilize the Miller multiplication of [Fig. 6.4(a)]. Explain the issues in this approach.

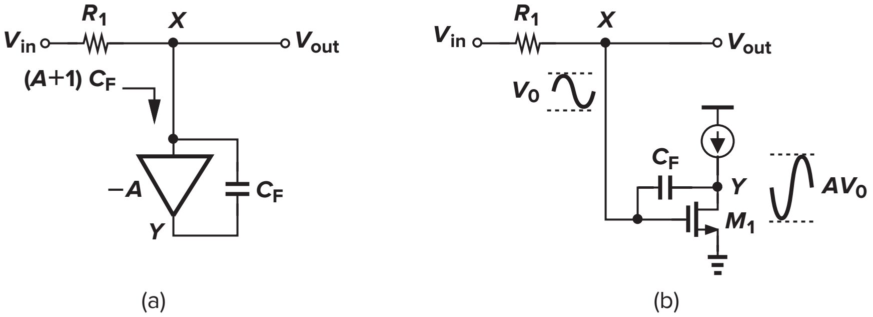


Figure 6.4

Solution

The issues relate to the amplifier, particularly to its output swing. As exemplified by the implementation in Fig. 6.4(b), if the voltage at X swings by V_0, then Y must accommodate a swing of AV_0 without saturating the amplifier. In addition, the dc level in V_{in} must be compatible with the input of the amplifier.

It is important to understand that (6.2) and (6.3) hold *if* we know a priori that the circuit of Fig. 6.2(a) can be converted to that of Fig. 6.2(b). That is, Miller's theorem does not stipulate the conditions under which this conversion is valid. If the impedance Z forms the only signal path between X and Y, then the conversion is often invalid. Illustrated in Fig. 6.5 for a simple resistive divider, the theorem gives a correct

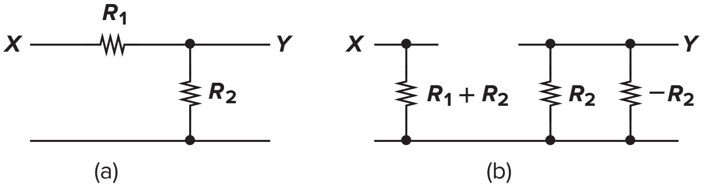


Figure 6.5 Improper application of Miller's theorem.

input impedance but an incorrect gain. Nevertheless, Miller's theorem proves useful in cases where the impedance Z appears in parallel with the main signal (Fig. 6.6).

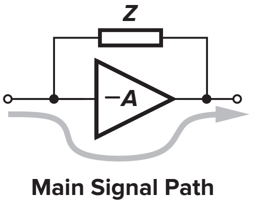


Figure 6.6 Typical case for valid application of Miller's theorem.

▶ **Example 6.3**

Calculate the input resistance of the circuit shown in Fig. 6.7(a).

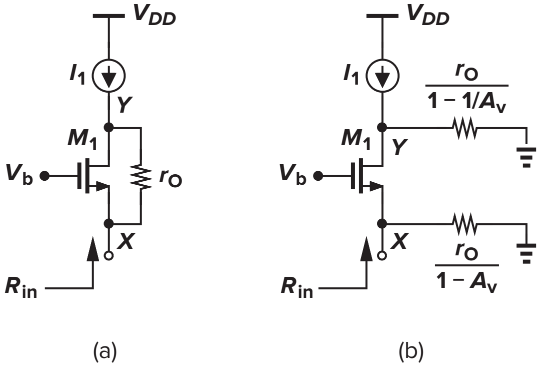


Figure 6.7

Solution

The reader can prove that the voltage gain from X to Y is equal to $1 + (g_m + g_{mb})r_O$. As shown in Fig. 6.7(b), the input resistance is given by the parallel combination of $r_O/(1 - A_v)$ and $1/(g_m + g_{mb})$. Since A_v is usually greater than unity, $r_O/(1 - A_v)$ is a *negative* resistance. We therefore have

$$R_{in} = \frac{r_O}{1 - [1 + (g_m + g_{mb})r_O]} \bigg\| \frac{1}{g_m + g_{mb}} \tag{6.4}$$

$$= \frac{-1}{g_m + g_{mb}} \bigg\| \frac{1}{g_m + g_{mb}} \tag{6.5}$$

$$= \infty \tag{6.6}$$

This is the same result as obtained in Chapter 3 (Fig. 3.54) by direct calculation. ◀

We should also mention that, strictly speaking, the value of $A_v = V_Y/V_X$ in (6.2) and (6.3) must be calculated at the frequency of interest, complicating the algebra significantly. To understand this point, let us return to Example 6.1 and assume an amplifier with a finite output resistance. Depicted in Fig. 6.8, the equivalent circuit reveals that $V_Y \neq -AV_X$ at high frequencies, and hence C_F cannot be simply multiplied by $1+A$ to yield the input capacitance. However, in many cases we use the low-frequency value of V_Y/V_X to gain insight into the behavior of the circuit. We call this approach "Miller's approximation."

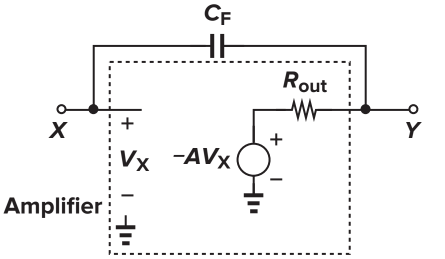


Figure 6.8 Equivalent circuit showing gain change at high frequencies.

▶ **Example 6.4**

Determine the transfer function of the circuit shown in Fig. 6.9(a) using (a) direct analysis and (b) Miller's approximation.

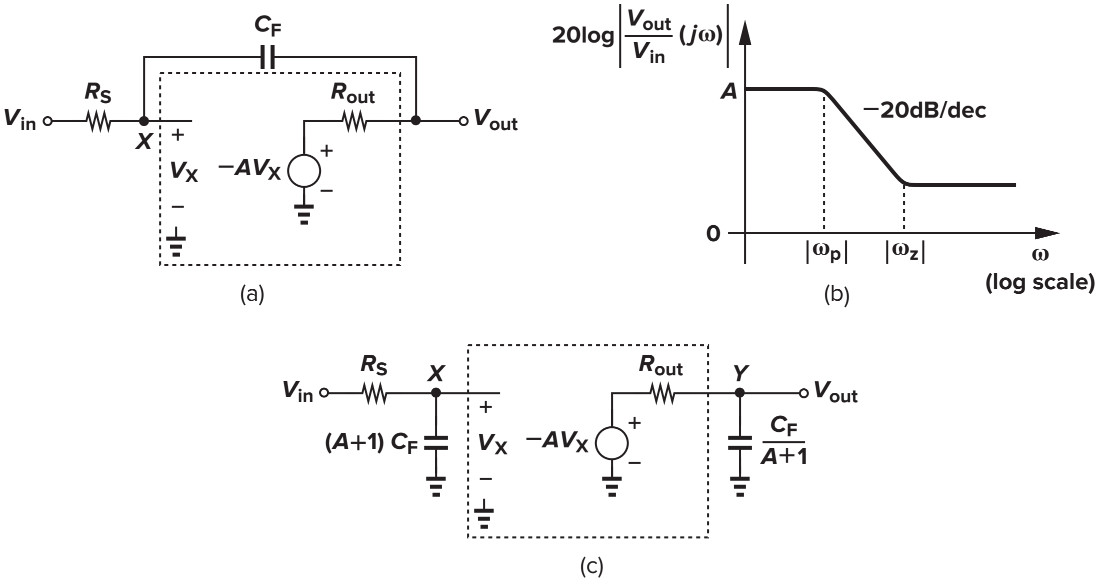


Figure 6.9

Solution

(a) We note that the current flowing through R_S is given by $(V_{in} - V_X)/R_S$, yielding a voltage drop across R_{out} equal to $(V_{in} - V_X)R_{out}/R_S$. It follows that

$$\frac{V_{in} - V_X}{R_S} R_{out} - AV_X = V_{out} \tag{6.7}$$

We also equate the currents flowing through R_S and C_F:

$$\frac{V_{in} - V_X}{R_S} = (V_X - V_{out})C_F s \tag{6.8}$$

The reader can find V_X from the first equation and substitute the result in the second, thereby obtaining

$$\frac{V_{out}}{V_{in}}(s) = \frac{R_{out}C_F s - A}{[(A+1)R_S + R_{out}]C_F s + 1} \tag{6.9}$$

The circuit thus exhibits a zero at $\omega_z = A/(R_{out}C_F)$ and a pole at $\omega_p = -1/[(A+1)R_S C_F + R_{out}C_F]$. Figure 6.9(b) plots the response for the case of $|\omega_p| < |\omega_z|$.

(b) Applying Miller's approximation, we decompose C_F into $(1+A)C_F$ at the input and $C_F/(1+A^{-1})$ at the output [Fig. 6.9(c)]. Since $V_{out}/V_{in} = (V_X/V_{in})(V_{out}/V_X)$, we first write V_X/V_{in} by considering R_S and $(1+A)C_F$ as a voltage divider:

$$\frac{V_X}{V_{in}} = \frac{\dfrac{1}{(1+A)C_F s}}{\dfrac{1}{(1+A)C_F s} + R_S} \tag{6.10}$$

$$= \frac{1}{(1+A)R_S C_F s + 1} \tag{6.11}$$

As for V_{out}/V_X, we first amplify V_X by $-A$ and subject the result to the output voltage divider,

$$\frac{V_{out}}{V_X} = \frac{-A}{\dfrac{1}{1+A^{-1}}C_F R_{out} s + 1} \tag{6.12}$$

That is

$$\frac{V_{out}}{V_{in}}(s) = \frac{-A}{[(1+A)R_S C_F s + 1]\left(\dfrac{1}{1+A^{-1}}C_F R_{out} s + 1\right)} \tag{6.13}$$

Sadly, Miller's approximation has eliminated the zero and predicted *two* poles for the circuit! Despite these shortcomings, Miller's approximation can provide intuition in many cases.[1]

If applied to obtain the input-output transfer function, Miller's theorem cannot be used simultaneously to calculate the output impedance. To derive the transfer function, we apply a voltage source to the *input* of the circuit, obtaining a value for V_Y/V_X in Fig. 6.2(a). On the other hand, to determine the output impedance, we must apply a voltage source to the *output* of the circuit, obtaining a value for V_X/V_Y that may not be equal to the inverse of the V_Y/V_X measured in the first test. For example, the circuit of Fig. 6.7(b) may suggest that the output impedance is equal to

$$R_{out} = \frac{r_O}{1 - 1/A_v} \tag{6.14}$$

$$= \frac{r_O}{1 - [1 + (g_m + g_{mb})r_O]^{-1}} \tag{6.15}$$

$$= \frac{1}{g_m + g_{mb}} + r_O \tag{6.16}$$

[1] Both of these artifacts can be avoided if we multiply C_F by $1 + A(s)$, where $A(s)$ is the actual transfer function from V_X to V_{out}, but the algebra is as lengthy as that in part (a).

whereas the actual value is equal to r_O (if X is grounded). Other subtleties of Miller's theorem are described in the Appendix C.

In summary, Miller's approximation divides a floating impedance by the low-frequency gain and faces the following limitations: (1) it may eliminate zeros, (2) it may predict additional poles, and (3) it does not correctly compute the "output" impedance.

6.1.2 Association of Poles with Nodes

Consider the simple cascade of amplifiers depicted in Fig. 6.10. Here, A_1 and A_2 are ideal voltage amplifiers, R_1 and R_2 model the output resistance of each stage, C_{in} and C_N represent the input capacitance of each stage, and C_P denotes the load capacitance. The overall transfer function can be written as

$$\frac{V_{out}}{V_{in}}(s) = \frac{A_1}{1 + R_S C_{in} s} \cdot \frac{A_2}{1 + R_1 C_N s} \cdot \frac{1}{1 + R_2 C_P s} \tag{6.17}$$

The circuit exhibits three poles, each of which is determined by the total capacitance seen from each node to ground multiplied by the total resistance seen at the node to ground. We can therefore associate each pole with one node of the circuit, i.e., $\omega_j = \tau_j^{-1}$, where τ_j is the product of the capacitance and resistance seen at node j to ground. From this perspective, we may say that "each node in the circuit contributes one pole to the transfer function."

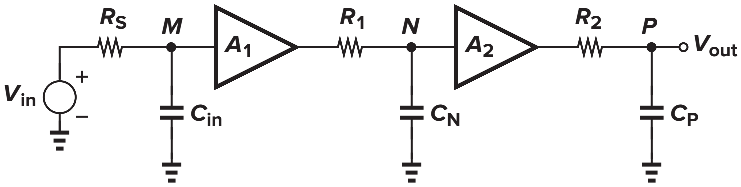


Figure 6.10 Cascade of amplifiers.

The above statement is not valid in general. For example, in the circuit of Fig. 6.11, the location of the poles is difficult to calculate because R_3 and C_3 create interaction between X and Y. Nevertheless, in many circuits, association of one pole with each node provides an intuitive approach to estimating the transfer function: we simply multiply the total equivalent capacitance by the total incremental (small-signal) resistance (both from the node of interest to ground), thus obtaining an equivalent time constant and hence a pole frequency.

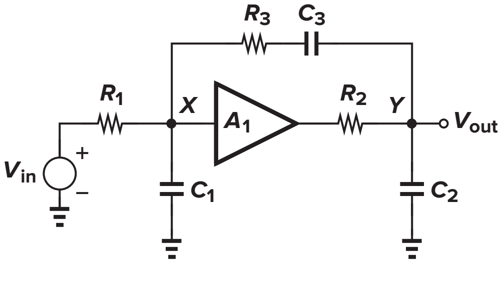
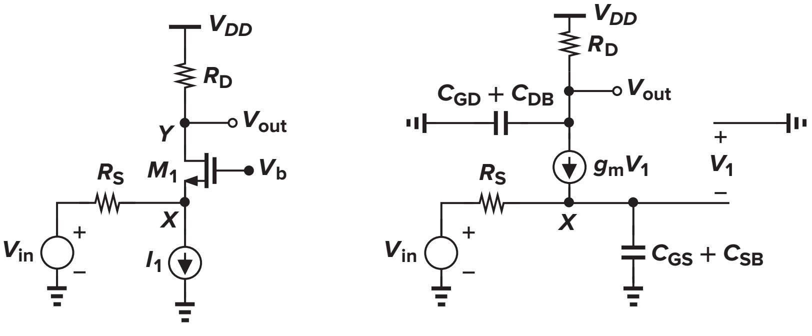


Figure 6.11 Example of interaction between nodes.

▶ **Example 6.5**

Neglecting channel-length modulation, compute the transfer function of the common-gate stage shown in Fig. 6.12(a).



(a) (b)

Figure 6.12 Common-gate stage with parasitic capacitances.

Solution

In this circuit, the capacitances contributed by M_1 are connected from the input and output nodes to ground [Fig. 6.12(b)]. At node X, $C_S = C_{GS} + C_{SB}$, giving a pole frequency

$$\omega_{in} = \left[(C_{GS} + C_{SB}) \left(R_S \middle\| \frac{1}{g_m + g_{mb}} \right) \right]^{-1} \tag{6.18}$$

Similarly, at node Y, $C_D = C_{DG} + C_{DB}$, yielding a pole frequency

$$\omega_{out} = [(C_{DG} + C_{DB})R_D]^{-1} \tag{6.19}$$

The overall transfer function is thus given by

$$\frac{V_{out}}{V_{in}}(s) = \frac{(g_m + g_{mb})R_D}{1 + (g_m + g_{mb})R_S} \cdot \frac{1}{\left(1 + \dfrac{s}{\omega_{in}}\right)\left(1 + \dfrac{s}{\omega_{out}}\right)} \tag{6.20}$$

where the first fraction represents the low-frequency gain of the circuit. Note that if we do not neglect r_{O1}, the input and output nodes interact, making it difficult to calculate the poles.

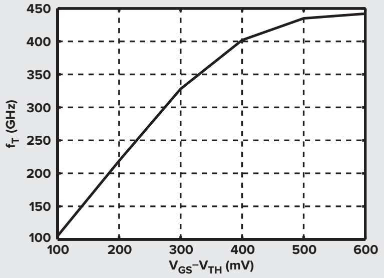
As observed in Example 6.4, Miller's approximation converts a floating impedance to two grounded impedances, allowing us to associate one pole with each node. We apply this technique to various amplifier topologies in this chapter—but cautiously and retrospectively to avoid its pitfalls. It is also helpful to bear in mind that the f_T of a MOS transistor is roughly equal to $g_m/(2\pi C_{GS})$ and can exceed 300 GHz in today's technologies. (However, since $f_T \propto V_{GS} - V_{TH}$, as we push the devices for low-voltage operation, we tend to reduce their f_T's.)

6.2 ■ Common-Source Stage

The common-source topology exhibits a relatively high input impedance while providing voltage gain and requiring a minimal voltage headroom. As such, it finds wide application in analog circuits and its frequency response is of interest.

Nanometer Design Notes

Defined as the frequency at which the small-signal current gain of a device is unity, the transit frequency, f_T, of MOSFETs increases with the overdrive, but flattens out as the vertical electric field reduces the mobility. Plotted below is the f_T for an NMOS device with $W/L = 5$ μm/40 nm and $V_{DS} = 0.8$ V.



Shown in Fig. 6.13(a) is a common-source stage driven by a finite source resistance, R_S.[2] We identify all of the capacitances in the circuit, noting that C_{GS} and C_{DB} are "grounded" capacitances while C_{GD} appears between the input and the output. In reality, the circuit also drives a load capacitance, which can be merged with C_{DB}.

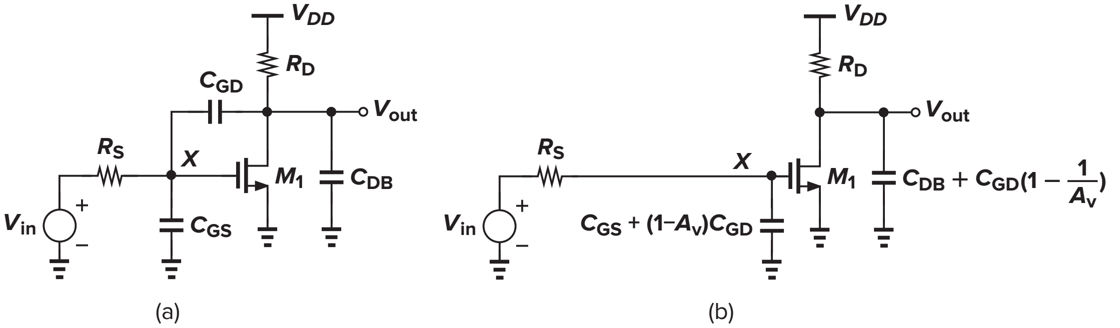


Figure 6.13 (a) High-frequency model of a common-source stage, and (b) simplified circuit using Miller's approximation.

Miller's Approximation Assuming that $\lambda = 0$ and M_1 operates in saturation, let us first estimate the transfer function by associating one pole with each node. The total capacitance seen from X to ground is equal to C_{GS} plus the Miller multiplication of C_{GD}, namely, $C_{GS} + (1 - A_v)C_{GD}$, where $A_v = -g_m R_D$ [Fig. 6.13(b)]. The magnitude of the "input" pole is therefore given by

$$\omega_{in} = \frac{1}{R_S[C_{GS} + (1 + g_m R_D)C_{GD}]} \tag{6.21}$$

At the output node, the total capacitance seen to ground is equal to C_{DB} plus the Miller effect of C_{GD}, i.e., $C_{DB} + (1 - A_v^{-1})C_{GD} \approx C_{DB} + C_{GD}$ (if $A_v \gg 1$). Thus,

$$\omega_{out} = \frac{1}{R_D(C_{DB} + C_{GD})} \tag{6.22}$$

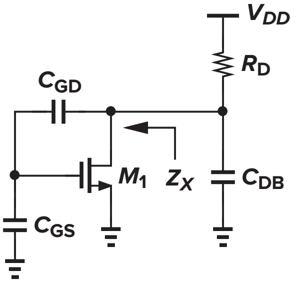


Figure 6.14 Model for calculation of output impedance.

Another approximation of the output pole can be obtained if R_S is relatively large. Simplifying the circuit as shown in Fig. 6.14, where the effect of R_S is neglected, the reader can prove that

$$Z_X = \frac{1}{C_{eq}s} \left\| \left(\frac{C_{GD} + C_{GS}}{C_{GD}} \cdot \frac{1}{g_{m1}} \right) \right. \tag{6.23}$$

[2] Note that R_S is not deliberately added to the circuit. Rather, it models the output resistance of the preceding stage.

where $C_{eq} = C_{GD}C_{GS}/(C_{GD} + C_{GS})$. Thus, the output pole is roughly equal to

$$\omega_{out} = \frac{1}{\left[R_D \middle\| \left(\frac{C_{GD} + C_{GS}}{C_{GD}} \cdot \frac{1}{g_{m1}}\right)\right](C_{eq} + C_{DB})} \tag{6.24}$$

We should point out that the sign of ω_{in} and ω_{out} in the above equations is positive because we eventually write the denominator of the transfer function in the form of $(1 + s/\omega_{in})(1 + s/\omega_{out})$; i.e., the denominator vanishes at $s = -\omega_{in}$ and $s = -\omega_{out}$. Alternatively, we could express the values of ω_{in} and ω_{out} with a negative sign and hence write the denominator as $(1 - s/\omega_{in})(1 - s/\omega_{out})$. We adopt the former notation in this book. We then surmise that the transfer function is

$$\frac{V_{out}}{V_{in}}(s) = \frac{-g_m R_D}{\left(1 + \frac{s}{\omega_{in}}\right)\left(1 + \frac{s}{\omega_{out}}\right)} \tag{6.25}$$

Note that r_{O1} and any load capacitance can easily be included here.

The primary error in this estimation is that we have not considered the existence of zeros in the circuit. Another concern stems from approximating the gain of the amplifier by $-g_m R_D$ whereas in reality the gain varies with frequency (for example, due to the capacitance at the output node).

Direct Analysis We now obtain the exact transfer function, investigating the validity of the above approach. Using the equivalent circuit depicted in Fig. 6.15, we can sum the currents at each node:

$$\frac{V_X - V_{in}}{R_S} + V_X C_{GS}s + (V_X - V_{out})C_{GD}s = 0 \tag{6.26}$$

$$(V_{out} - V_X)C_{GD}s + g_m V_X + V_{out}\left(\frac{1}{R_D} + C_{DB}s\right) = 0 \tag{6.27}$$

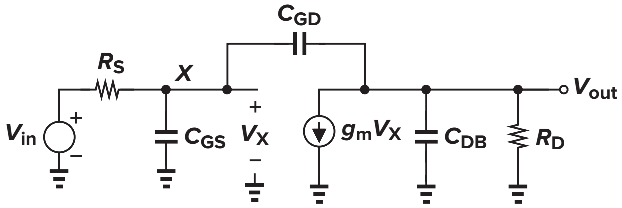


Figure 6.15 Equivalent circuit of Fig. 6.13.

From (6.27), V_X is obtained as

$$V_X = -\frac{V_{out}\left(C_{GD}s + \frac{1}{R_D} + C_{DB}s\right)}{g_m - C_{GD}s} \tag{6.28}$$

which, upon substitution in (6.26), yields

$$-V_{out}\frac{[R_S^{-1} + (C_{GS} + C_{GD})s][R_D^{-1} + (C_{GD} + C_{DB})s]}{g_m - C_{GD}s} - V_{out}C_{GD}s = \frac{V_{in}}{R_S} \tag{6.29}$$

That is

$$\frac{V_{out}}{V_{in}}(s) = \frac{(C_{GD}s - g_m)R_D}{R_S R_D \xi s^2 + [R_S(1 + g_m R_D)C_{GD} + R_S C_{GS} + R_D(C_{GD} + C_{DB})]s + 1} \tag{6.30}$$

where $\xi = C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB}$. Note that the transfer function is of second order even though the circuit contains three capacitors. This is because the capacitors form a "loop," allowing only *two* independent initial conditions in the circuit and hence yielding a second-order differential equation for the time response.

▶ Example 6.6

A student considers only C_{GD} in Fig. 6.13(a) so as to obtain a one-pole response, reasons that the voltage gain drops by 3 dB (by a factor of $= \sqrt{2}$) at the pole frequency, and concludes that a better approximation of the Miller effect should multiply C_{GD} by $1 + g_m R_D \sqrt{2}$. Explain the flaw in this reasoning.

Solution

Setting C_{GS} and C_{DB} to zero, we obtain

$$\frac{V_{out}}{V_{in}}(s) = \frac{(C_{GD}s - g_m)R_D}{\dfrac{s}{\omega_0} + 1} \tag{6.31}$$

where $\omega_0 = R_S(1 + g_m R_D)C_{GD} + R_D C_{GD}$. We note that C_{GD} is multiplied by $1 + g_m R_D$ in this exact analysis. So where is the flaw in the student's argument? It is true that the voltage gain in Fig. 6.13(a) falls by $\sqrt{2}$ at ω_0, but this gain would be from V_{in} to V_{out} and *not* the gain seen by C_{GD}. The reader can readily express the transfer function from node X to V_{out} as

$$\frac{V_{out}}{V_X}(s) = \frac{(C_{GD}s - g_m)R_D}{R_D C_{GD} + 1} \tag{6.32}$$

observing that this gain begins to roll off at a *higher* frequency, namely, at $1/(R_D C_{GD})$. Thus, the multiplication of C_{GD} by $1 + g_m R_D$ is still justified. ◀

Special Cases If manipulated judiciously, Eq. (6.30) reveals several interesting points about the circuit. While the denominator appears rather complicated, it can yield intuitive expressions for the two poles, ω_{p1} and ω_{p2}, if we assume that $|\omega_{p1}| \ll |\omega_{p2}|$. This is called the "dominant pole" approximation. Writing the denominator as

$$D = \left(\frac{s}{\omega_{p1}} + 1\right)\left(\frac{s}{\omega_{p2}} + 1\right) \tag{6.33}$$

$$= \frac{s^2}{\omega_{p1}\omega_{p2}} + \left(\frac{1}{\omega_{p1}} + \frac{1}{\omega_{p2}}\right)s + 1 \tag{6.34}$$

we recognize that the coefficient of s is approximately equal to $1/\omega_{p1}$ if ω_{p2} is much farther from the origin. It follows from (6.30) that the dominant pole is given by

$$\omega_{p1} = \frac{1}{R_S(1 + g_m R_D)C_{GD} + R_S C_{GS} + R_D(C_{GD} + C_{DB})} \tag{6.35}$$

How does this compare with the "input" pole given by (6.21)? The only difference results from the term $R_D(C_{GD}+C_{DB})$, which may be negligible in some cases. The key point here is that the intuitive approach of associating a pole with the input node provides a rough estimate with much less effort. We also note that the Miller multiplication of C_{GD} by the low-frequency gain of the amplifier is relatively accurate in this case. Of course, for a given set of values, we must check to ensure that $\omega_{p1} \ll \omega_{p2}$.

Other special cases are also of interest. We consider the case of $C_{GD} = 0$ in Problem 6.26 and the case of $R_D = \infty$ below.

▶ **Example 6.7**

The circuit shown in Fig. 6.16(a) is a special case where $R_D \to \infty$. Calculate the transfer function (with $\lambda = 0$) and explain why the Miller effect vanishes as C_{DB} (or the load capacitance) increases.

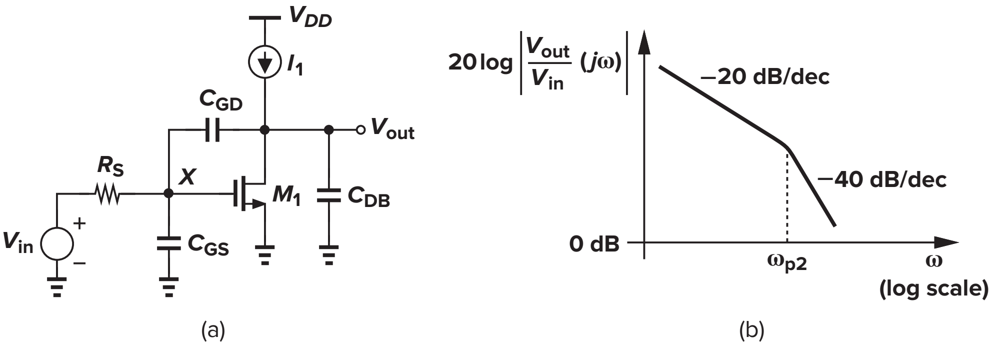


Figure 6.16

Solution

Using (6.30) and letting R_D approach infinity, we have

$$\frac{V_{out}}{V_{in}}(s) = \frac{C_{GD}s - g_m}{R_S\xi s^2 + [g_m R_S C_{GD} + (C_{GD} + C_{DB})]s} \tag{6.36}$$

$$= \frac{C_{GD}s - g_m}{s[R_S(C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB})s + (g_m R_S + 1)C_{GD} + C_{DB}]}$$

As expected, the circuit exhibits two poles—one at the origin because the dc gain is infinity [Fig. 6.16(b)]. The magnitude of the other pole is given by

$$\omega_{p2} \approx \frac{(1 + g_m R_S)C_{GD} + C_{DB}}{R_S(C_{GD}C_{GS} + C_{GS}C_{DB} + C_{GD}C_{DB})} \tag{6.37}$$

For a large C_{DB} or load capacitance, this expression reduces to

$$\omega_{p2} \approx \frac{1}{R_S(C_{GS} + C_{GD})} \tag{6.38}$$

indicating that C_{GD} experiences no Miller multiplication. This can be explained by noting that, for a large C_{DB}, the voltage gain from node X to the output begins to drop even at low frequencies. As a result, for frequencies close to $[R_S(C_{GS} + C_{GD})]^{-1}$, the effective gain is quite small and $C_{GD}(1 - A_v) \approx C_{GD}$. Such a case is an example where the application of the Miller effect using low-frequency gain does not provide a reasonable estimate. ◀

From (6.30) and applying the dominant pole approximation, we can also estimate the second pole of the CS stage of Fig. 6.13(a). Since the coefficient of s^2 is equal to $(\omega_{p1}\omega_{p2})^{-1}$, we have

$$\omega_{p2} = \frac{1}{\omega_{p1}} \cdot \frac{1}{R_S R_D (C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB})} \tag{6.39}$$

$$= \frac{R_S(1+g_m R_D)C_{GD} + R_S C_{GS} + R_D(C_{GD}+C_{DB})}{R_S R_D (C_{GS}C_{GD} + C_{GS}C_{DB} + C_{GD}C_{DB})} \tag{6.40}$$

We emphasize that these results hold only if $\omega_{p1} \ll \omega_{p2}$.

As a special case, if $C_{GS} \gg (1+g_m R_D)C_{GD} + R_D(C_{GD}+C_{DB})/R_S$, then

$$\omega_{p2} \approx \frac{R_S C_{GS}}{R_S R_D (C_{GS}C_{GD} + C_{GS}C_{DB})} \tag{6.41}$$

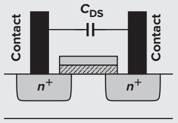
$$= \frac{1}{R_D(C_{GD}+C_{DB})} \tag{6.42}$$

the same as (6.22). Thus, the "output" pole approach is valid only if C_{GS} dominates the response.

The transfer function of (6.30) exhibits a zero given by $\omega_z = +g_m/C_{GD}$, an effect not predicted by Miller's approximation and (6.25). Located in the *right* half plane, the zero arises from direct coupling of the input to the output through C_{GD}. As illustrated in Fig. 6.17, C_{GD} provides a feedthrough path that conducts the input signal to the output at very high frequencies, resulting in a slope in the frequency response that is less negative than -40 dB/dec. Note that $g_m/C_{GD} > g_m/C_{GS}$ because $C_{GD} < C_{GS}$, implying that the zero lies beyond the transistor's f_T. However, as explained in Chapter 10, this zero falls to lower frequencies in cases where we deliberately add a capacitor between the gate and the drain, introducing other difficulties.

Nanometer Design Notes

The high-frequency MOS model developed in Chapter 2 does not contain a drain-source capacitance. In reality, however, the metal contact stacks touching the source and drain areas form two "columns" that create a capacitance between the drain and the source. This effect has become more pronounced in modern CMOS technologies because of the shorter channel length, i.e., less spacing between the columns, and the ability to stack many contacts, i.e., taller columns. The reader is encouraged to analyze a C_G stage while including C_{DS}.



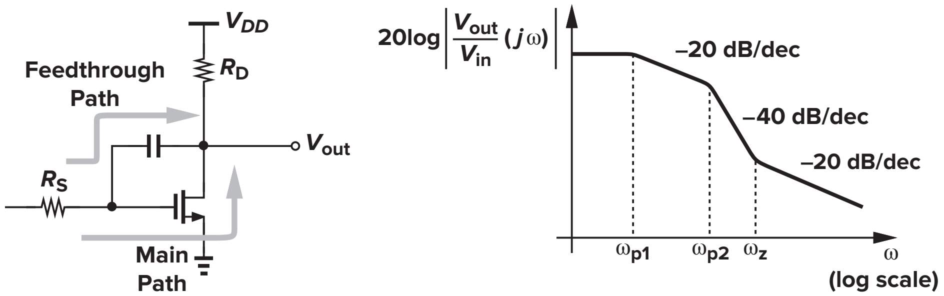


Figure 6.17 Feedforward path through C_{GD} (log-log scale).

The zero, s_z, can also be computed by noting that the transfer function $V_{out}(s)/V_{in}(s)$ must drop to zero for $s = s_z$. For a finite V_{in}, this means that $V_{out}(s_z) = 0$, and hence the output can be *shorted* to ground at this (possibly complex) frequency with no current flowing through R_D or the short (Fig. 6.18). Therefore, the currents through C_{GD} and M_1 are equal and opposite:

$$V_1 C_{GD} s_z = g_m V_1 \tag{6.43}$$

That is, $s_z = +g_m/C_{GD}$.[3]

[3] This approach is similar to expressing the transfer function as $G_m Z_{out}$ and finding the zeros of G_m and Z_{out}.

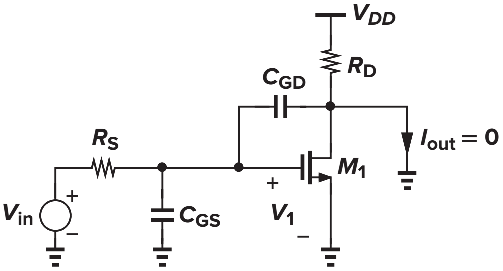


Figure 6.18 Calculation of the zero in a CS stage.

▶ Example 6.8

We have seen that the signals traveling through two paths within an amplifier may cancel each other at one frequency, creating a zero in the transfer function (Fig. 6.19). Can this occur if $H_1(s)$ and $H_2(s)$ are first-order low-pass circuits?

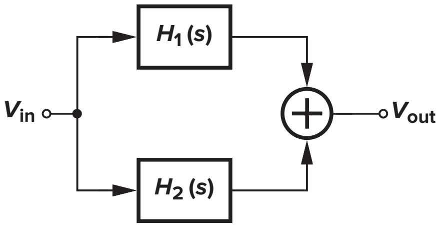


Figure 6.19

Solution

Modeling $H_1(s)$ by $A_1/(1+s/\omega_{p1})$ and $H_2(s)$ by $A_2/(1+s/\omega_{p2})$, we have

$$\frac{V_{out}}{V_{in}}(s) = \frac{\left(\dfrac{A_1}{\omega_{p2}} + \dfrac{A_2}{\omega_{p1}}\right)s + A_1 + A_2}{\left(1+\dfrac{s}{\omega_{p1}}\right)\left(1+\dfrac{s}{\omega_{p2}}\right)} \tag{6.44}$$

Indeed, the overall transfer function contains a zero.

▶ Example 6.9

Determine the transfer function of the complementary CS stage shown in Fig. 6.20(a).

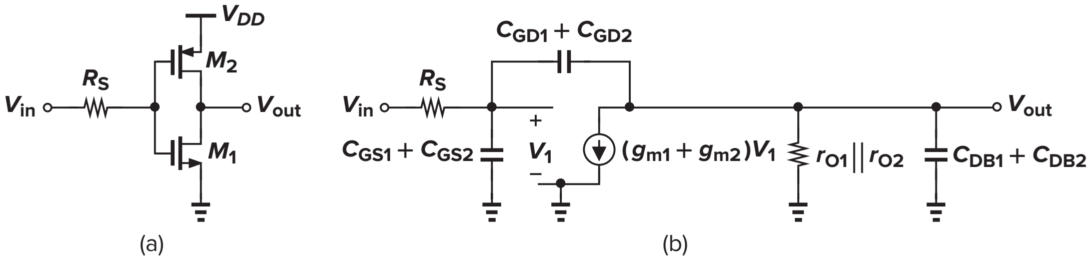


Figure 6.20

Solution

Since the corresponding terminals of M_1 and M_2 are shorted to one another in the small-signal model, we merge the two transistors, drawing the equivalent circuit as shown in Fig. 6.20(b). The circuit thus has the same transfer function as the simple CS stage studied above.

In high-speed applications, the input impedance of the common-source stage is also important. With the aid of Miller's approximation, we have from Fig. 6.21(a)

$$Z_{in} = \frac{1}{[C_{GS} + (1 + g_m R_D)C_{GD}]s} \tag{6.45}$$

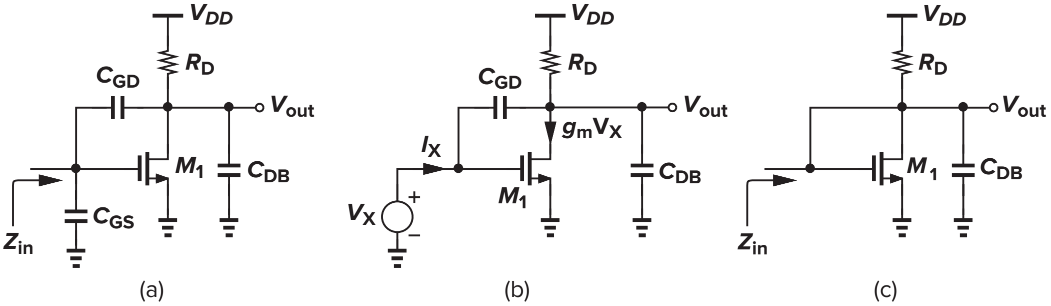


Figure 6.21 Calculation of input impedance of a CS stage.

But at high frequencies, the effect of the output node capacitance must be taken into account. Ignoring C_{GS} for the moment and using the circuit of Fig. 6.21(b), we add the voltage drops across $R_D||(C_{DB}s)^{-1}$ and C_{GD}, equating the result to V_X:

$$(I_X - g_m V_X)\frac{R_D}{1 + R_D C_{DB}s} + \frac{I_X}{C_{GD}s} = V_X \tag{6.46}$$

and hence

$$\frac{V_X}{I_X} = \frac{1 + R_D(C_{GD} + C_{DB})s}{C_{GD}s(1 + g_m R_D + R_D C_{DB}s)} \tag{6.47}$$

The actual input impedance consists of the parallel combination of (6.47) and $1/(C_{GS}s)$.

As a special case, suppose that at the frequency of interest, $|R_D(C_{GD}+C_{DB})s| \ll 1$ and $|R_D C_{DB}s| \ll 1+g_m R_D$. Then, (6.47) reduces to $[(1+g_m R_D)C_{GD}s]^{-1}$ (as expected), indicating that the input impedance is primarily capacitive. At higher frequencies, however, (6.47) contains both real and imaginary parts. In fact, if C_{GD} is large, it provides a low-impedance path between the gate and the drain of M_1, yielding the equivalent circuit of Fig. 6.21(c) and suggesting that $1/g_{m1}$ and R_D appear in parallel with the input.

▶ Example 6.10

Explain what happens to Eq. (6.47) if the circuit drives a large load capacitance.

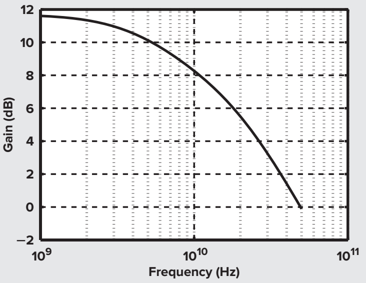
Solution

Merged with C_{DB}, the large load capacitance reduces the numerator to $R_D C_{DB}s$ and the denominator to $C_{GD}s(R_D C_{DB}s)$, yielding $V_X/I_X \approx 1/(C_{GD}s)$. In a manner similar to that in Example 6.7, the large load capacitance lowers the gain at high frequencies, suppressing Miller multiplication of C_{GD}.

6.3 ■ Source Followers

Nanometer Design Notes

While characterizing the intrinsic MOSFET, the f_T overestimates how fast typical circuits can operate. A more "circuit-centric" measure is the gain-bandwidth product of a simple CS amplifier that is loaded by an identical stage. Shown below is the frequency response of such an amplifier, revealing a gain-bandwidth product of about 34 GHz if $W/L = 5\ \mu$m/40 nm, $R_D = 5$ kΩ, and the bias current is 130 μA.



Source followers are occasionally employed as level shifters or buffers, affecting the overall frequency response. Consider the circuit depicted in Fig. 6.22(a), where C_L represents the total capacitance seen at the output node to ground, including C_{SB1}. The strong interaction between nodes X and Y through C_{GS} in Fig. 6.22(a) makes it difficult to associate a pole with each node in a source follower. Neglecting channel-length modulation and body effect for simplicity and using the equivalent circuit shown in Fig. 6.22(b), we sum the currents at the output node:

$$V_1 C_{GS} s + g_m V_1 = V_{out} C_L s \tag{6.48}$$

obtaining

$$V_1 = \frac{C_L s}{g_m + C_{GS} s} V_{out} \tag{6.49}$$

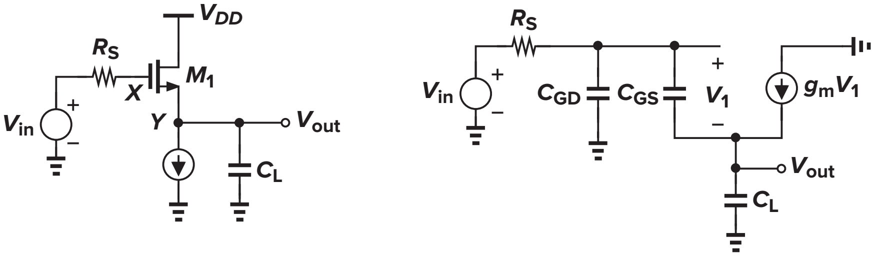
Also, noting that the voltage across C_{GD} is equal to $V_1 + V_{out}$ and beginning from V_{in}, we add the voltage across R_S to V_1 and V_{out}:

$$V_{in} = R_S[V_1 C_{GS} s + (V_1 + V_{out}) C_{GD} s] + V_1 + V_{out} \tag{6.50}$$

Substituting for V_1 from (6.49), we have

$$\frac{V_{out}}{V_{in}}(s) = \frac{g_m + C_{GS} s}{R_S(C_{GS} C_L + C_{GS} C_{GD} + C_{GD} C_L) s^2 + (g_m R_S C_{GD} + C_L + C_{GS}) s + g_m} \tag{6.51}$$

Interestingly, the transfer function contains a zero in the *left* half plane (and near the f_T). This is because the signal conducted by C_{GS} at high frequencies adds with the same polarity to the signal produced by the intrinsic transistor. We study some special cases below.



(a) (b)

Figure 6.22 (a) Source follower; (b) high-frequency equivalent circuit.

▶ **Example 6.11**

Examine the source follower transfer function if $C_L = 0$.

Solution

We have

$$\frac{V_{out}}{V_{in}} = \frac{g_m + C_{GS}s}{R_S C_{GS} C_{GD} s^2 + (g_m R_S C_{GD} + C_{GS})s + g_m} \tag{6.52}$$

$$= \frac{g_m + C_{GS}s}{(1 + R_S C_{GD}s)(g_m + C_{GS}s)} \tag{6.53}$$

$$= \frac{1}{1 + R_S C_{GD}s} \tag{6.54}$$

The circuit now has only one pole at the input. Why does C_{GS} disappear here? This is because, in the absence of channel-length modulation and body effect, the voltage gain from the gate to the source is equal to unity. Since a change of ΔV at the gate translates to an equal change at the source (Fig. 6.23), no current flows through C_{GS}. Consequently, C_{GS} contributes neither a zero nor a pole. We say C_{GS} is "bootstrapped" by the source follower. With $\lambda, \gamma > 0$, the output change is less than ΔV, requiring some change in the voltage across C_{GS}.

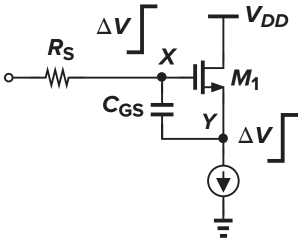


Figure 6.23 Bootstrapping of C_{GS} in a source follower.

If the two poles of (6.51) are assumed far apart, then the lower one has a magnitude of

$$\omega_{p1} \approx \frac{g_m}{g_m R_S C_{GD} + C_L + C_{GS}} \tag{6.55}$$

$$= \frac{1}{R_S C_{GD} + \dfrac{C_L + C_{GS}}{g_m}} \tag{6.56}$$

Also, if $R_S = 0$, then $\omega_{p1} = g_m/(C_L + C_{GS})$—as expected.

Let us now calculate the input impedance of the circuit, noting that C_{GD} simply shunts the input and can be ignored initially. Shown in Fig. 6.24, the equivalent circuit includes body effect, but channel-length modulation can also be added by replacing $1/g_{mb}$ with $(1/g_{mb})||r_O$. The small-signal gate-source voltage of M_1 is equal to $I_X/(C_{GS}s)$, giving a source current of $g_m I_X/(C_{GS}s)$. Starting from the input and adding

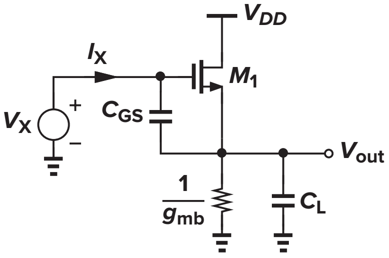


Figure 6.24 Calculation of source follower input impedance.

the voltages, we have

$$V_X = \frac{I_X}{C_{GS}s} + \left(I_X + \frac{g_m I_X}{C_{GS}s}\right)\left(\frac{1}{g_{mb}} \middle\| \frac{1}{C_L s}\right) \tag{6.57}$$

that is

$$Z_{in} = \frac{1}{C_{GS}s} + \left(1 + \frac{g_m}{C_{GS}s}\right)\frac{1}{g_{mb} + C_L s} \tag{6.58}$$

We consider some special cases. First, if $g_{mb} = 0$ and $C_L = 0$, then $Z_{in} = \infty$, because C_{GS} is entirely bootstrapped by the source follower and draws no current from the input. Second, at relatively low frequencies, $g_{mb} \gg |C_L s|$ and

$$Z_{in} \approx \frac{1}{C_{GS}s}\left(1 + \frac{g_m}{g_{mb}}\right) + \frac{1}{g_{mb}} \tag{6.59}$$

indicating that the equivalent input capacitance is equal to $C_{GS}g_{mb}/(g_m + g_{mb})$ and hence quite less than C_{GS}. In other words, the overall input capacitance is equal to C_{GD} plus a *fraction* of C_{GS}—again because of bootstrapping.

▶ **Example 6.12**

Apply Miller's approximation to the above circuit if $C_L = 0$.

Solution

As illustrated in Fig. 6.25, the low-frequency gain from the gate to the source is equal to $(1/g_{mb})/[(1/g_m) + (1/g_{mb})] = g_m/(g_m+g_{mb})$. The Miller multiplication of C_{GS} at the input is thus equal to $C_{GS}[1-g_m/(g_m+g_{mb})] = C_{GS}g_{mb}/(g_m + g_{mb})$.

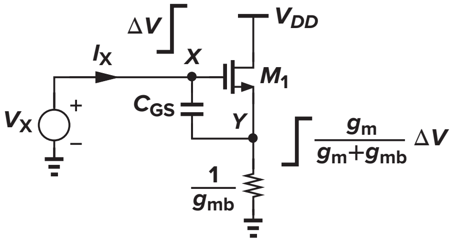


Figure 6.25

◀

At high frequencies, $g_{mb} \ll |C_L s|$ and

$$Z_{in} \approx \frac{1}{C_{GS}s} + \frac{1}{C_L s} + \frac{g_m}{C_{GS}C_L s^2} \tag{6.60}$$

For a given $s = j\omega$, the input impedance consists of the series combination of capacitors C_{GS} and C_L and a *negative* resistance equal to $-g_m/(C_{GS}C_L\omega^2)$ (Fig. 6.26). The negative resistance property can be utilized in oscillators (Chapter 15). It is important to bear in mind that a source follower driving a load capacitance exhibits a negative input resistance, possibly causing instability.

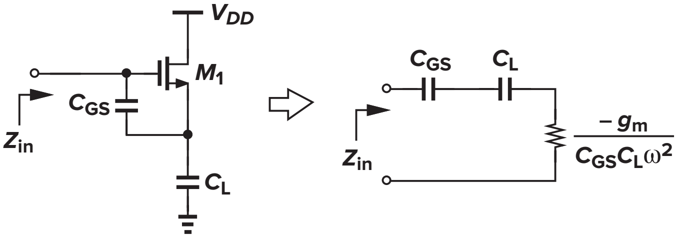


Figure 6.26 Negative resistance seen at the input of a source follower.

▶ **Example 6.13**

Neglecting channel-length modulation and body effect, calculate the transfer function of the circuit shown in Fig. 6.27(a).

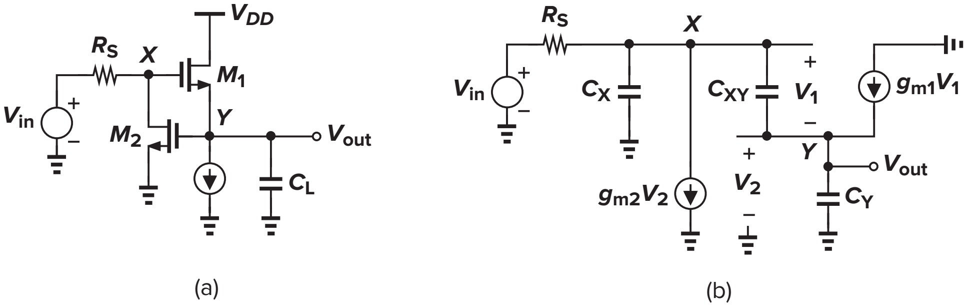


Figure 6.27

Solution

Let us first identify all of the capacitances in the circuit. At node X, C_{GD1} and C_{DB2} are connected to ground and C_{GS1} and C_{GD2} to Y. At node Y, C_{SB1}, C_{GS2}, and C_L are connected to ground. Similar to the source follower of Fig. 6.22(b), this circuit has three capacitances in a loop and hence a second-order transfer function. Using the equivalent circuit shown in Fig. 6.27(b), where $C_X = C_{GD1} + C_{DB2}$, $C_{XY} = C_{GS1} + C_{GD2}$, and $C_Y = C_{SB1} + C_{GS2} + C_L$, we have $V_1 C_{XY} s + g_{m1} V_1 = V_{out} C_Y s$, and hence $V_1 = V_{out} C_Y s/(C_{XY} s + g_{m1})$. Also, since $V_2 = V_{out}$, the summation of currents at node X gives

$$(V_1 + V_{out})C_X s + g_{m2} V_{out} + V_1 C_{XY} s = \frac{V_{in} - V_1 - V_{out}}{R_S} \tag{6.61}$$

Substituting for V_1 and simplifying the result, we obtain

$$\frac{V_{out}}{V_{in}}(s) = \frac{g_{m1} + C_{XY} s}{R_S \xi s^2 + [C_Y + g_{m1} R_S C_X + (1 + g_{m2} R_S) C_{XY}]s + g_{m1}(1 + g_{m2} R_S)} \tag{6.62}$$

where $\xi = C_X C_Y + C_X C_{XY} + C_Y C_{XY}$. As expected, (6.62) reduces to a form similar to (6.51) for $g_{m2} = 0$. ◀

The output impedance of source followers is also of interest. In Fig. 6.22(a), the body effect and C_{SB} simply yield an impedance in parallel with the output. Ignoring this impedance and neglecting C_{GD}, we note from the equivalent circuit of Fig. 6.28(a) that $V_1 C_{GS} s + g_m V_1 = -I_X$. Also, $V_1 C_{GS} s R_S + V_1 = -V_X$.

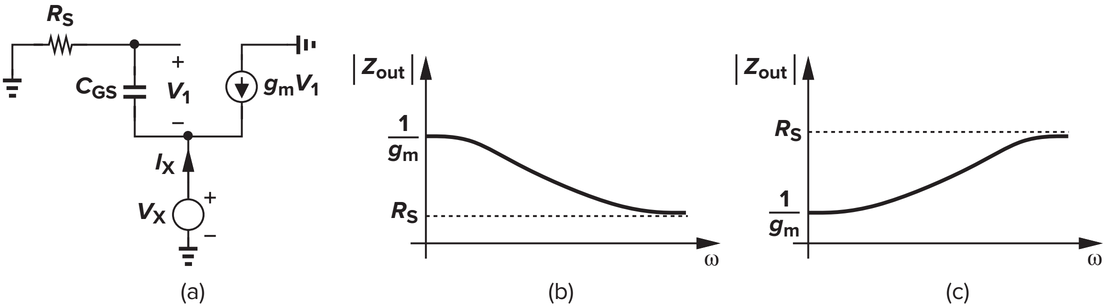


Figure 6.28 Calculation of source follower output impedance.

Dividing both sides of these equations gives

$$Z_{out} = \frac{V_X}{I_X} \tag{6.63}$$

$$= \frac{R_S C_{GS} s + 1}{g_m + C_{GS} s} \tag{6.64}$$

It is instructive to examine the magnitude of this impedance as a function of frequency. At low frequencies, $Z_{out} \approx 1/g_m$, as expected. At very high frequencies, $Z_{out} \approx R_S$ (because C_{GS} shorts the gate and the source). We therefore surmise that $|Z_{out}|$ varies as shown in Figs. 6.28(b) or (c). Which one of these variations is more realistic? Operating as buffers, source followers must lower the output impedance, i.e., $1/g_m < R_S$. For this reason, the characteristic shown in Fig. 6.28(c) occurs more commonly than that in Fig. 6.28(b).

The behavior illustrated in Fig. 6.28(c) reveals an important attribute of source followers. Since the output impedance *increases* with frequency, we postulate that it contains an *inductive* component. To confirm this guess, we represent Z_{out} by a first-order passive network, noting that Z_{out} equals $1/g_m$ at $\omega = 0$ and R_S at $\omega = \infty$. The network can therefore be realized as shown in Fig. 6.29 because Z_1 equals R_2 at $\omega = 0$ and $R_1 + R_2$ at $\omega = \infty$. In other words, $Z_1 = Z_{out}$ if three conditions hold: $R_2 = 1/g_m$, $R_1 = R_S - 1/g_m$, and L is chosen properly.

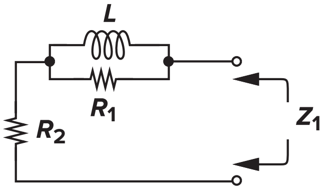


Figure 6.29 Equivalent output impedance of a source follower.

To calculate L, we can simply obtain an expression for Z_1 in terms of the three components in Fig. 6.29 and equate the result to Z_{out} found above. Alternatively, since R_2 is a series component of Z_1, we can subtract its value from Z_{out}, thereby obtaining an expression for the parallel combination of R_1 and L:

$$Z_{out} - \frac{1}{g_m} = \frac{C_{GS} s \left(R_S - \frac{1}{g_m} \right)}{g_m + C_{GS} s} \tag{6.65}$$

Inverting the result to obtain the admittance of the parallel circuit, we have

$$\frac{1}{Z_{out} - \dfrac{1}{g_m}} = \frac{1}{R_S - \dfrac{1}{g_m}} + \frac{1}{\dfrac{C_{GS}s}{g_m}\left(R_S - \dfrac{1}{g_m}\right)} \tag{6.66}$$

We can thus identify the first term on the right-hand side as the inverse of R_1 and the second term as the inverse of an impedance equal to $(C_{GS}s/g_m)(R_S - 1/g_m)$, i.e., an inductor with the value

$$L = \frac{C_{GS}}{g_m}\left(R_S - \frac{1}{g_m}\right) \tag{6.67}$$

Note that C_{GS}/g_m is approximately equal to $\omega_T = 2\pi f_T$.

▶ **Example 6.14**

Can we construct a (two-terminal) inductor from a source follower?

Solution

Yes, we can. Called an "active inductor," such a structure is shown in Fig. 6.30(a), providing an inductance of $(C_{GS2}/g_{m2})(R_S - 1/g_{m2})$. But the inductor is not ideal because it also incurs a parallel resistance equal to $R_1 = R_S = 1/g_{m2}$ and a series resistance equal to $1/g_{m2}$. Figure 6.30(b) depicts an application of active inductors: the inductance can partially cancel the load capacitance, C_L, at high frequencies, thus extending the *bandwidth*. However, the voltage headroom consumed by M_2 ($= V_{GS2}$) limits the gain. Also, C_{GD2}, which has been neglected in our analysis, limits the bandwidth enhancement.

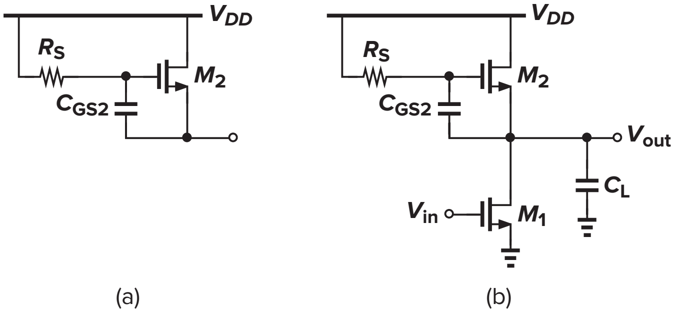


Figure 6.30

◀

6.4 ■ Common-Gate Stage

As explained in Example 6.5, in a common-gate stage, the input and output nodes are "isolated" if channel-length modulation is neglected. For a common-gate stage such as that in Fig. 6.31, the calculation of Example 6.5 suggested a transfer function

$$\frac{V_{out}}{V_{in}}(s) = \frac{(g_m + g_{mb})R_D}{1 + (g_m + g_{mb})R_S} \frac{1}{\left(1 + \dfrac{C_S}{g_m + g_{mb} + R_S^{-1}}s\right)(1 + R_D C_D s)} \tag{6.68}$$

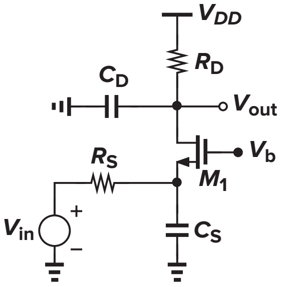


Figure 6.31 Common-gate stage at high frequencies.

An important property of this circuit is that it exhibits no Miller multiplication of capacitances, potentially achieving a wide band. Note, however, that the low input impedance may load the preceding stage. Furthermore, since the voltage drop across R_D is typically maximized to obtain a reasonable gain, the dc level of the input signal must be quite low. For these reasons, the CG stage finds two principal applications: as an amplifier in cases where a low input impedance is required (Chapter 3) and in cascode stages.

If channel-length modulation is not negligible, the calculations become quite complex. Recall from Chapter 3 that the input impedance of a common-gate topology does depend on the drain load if $\lambda \neq 0$. From Eq. (3.117), we can express the impedance seen looking into the source of M_1 in Fig. 6.31 as

$$Z_{in} \approx \frac{Z_L}{(g_m + g_{mb})r_O} + \frac{1}{g_m + g_{mb}} \tag{6.69}$$

where $Z_L = R_D \| [1/(C_D s)]$. Since Z_{in} now depends on Z_L, it is difficult to associate a pole with the input node.

▶ Example 6.15

For the common-gate stage shown in Fig. 6.32(a), calculate the transfer function and the input impedance, Z_{in}. Explain why Z_{in} becomes independent of C_L as this capacitance increases.

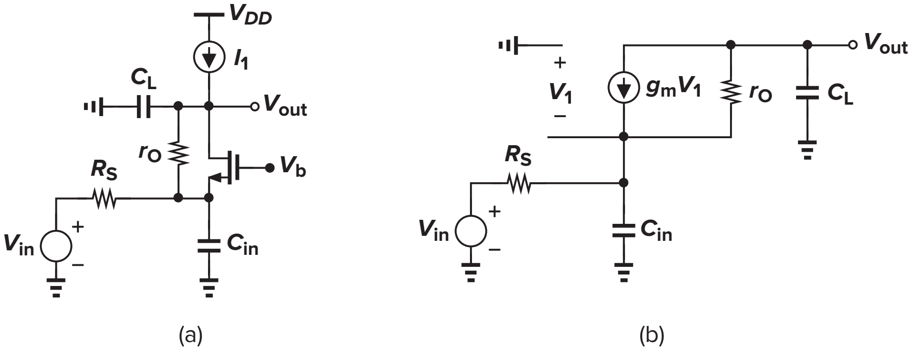


Figure 6.32

Solution

Using the equivalent circuit shown in Fig. 6.32(b), we can write the current through R_S as $-V_{out}C_L s + V_1 C_{in} s$. Noting that the voltage across R_S plus V_{in} must equal $-V_1$, we have

$$(-V_{out}C_L s + V_1 C_{in} s)R_S + V_{in} = -V_1 \tag{6.70}$$

That is

$$V_1 = -\frac{-V_{out}C_L s R_S + V_{in}}{1 + C_{in}R_S s} \tag{6.71}$$

We also observe that the voltage across r_O minus V_1 equals V_{out}:

$$r_O(-V_{out}C_L s - g_m V_1) - V_1 = V_{out} \tag{6.72}$$

Substituting for V_1 from (6.71), we obtain the transfer function:

$$\frac{V_{out}}{V_{in}}(s) = \frac{1 + g_m r_O}{r_O C_L C_{in} R_S s^2 + [r_O C_L + C_{in} R_S + (1 + g_m r_O) C_L R_S]s + 1} \tag{6.73}$$

The reader can prove that body effect can be included by simply replacing g_m with $g_m + g_{mb}$. As expected, the gain at very low frequencies is equal to $1 + g_m r_O$. For Z_{in}, we can use (6.69) by replacing Z_L with $1/(C_L s)$, obtaining

$$Z_{in} = \frac{1}{g_m + g_{mb}} + \frac{1}{C_L s} \cdot \frac{1}{(g_m + g_{mb}) r_O} \tag{6.74}$$

We note that as C_L or s increases, Z_{in} approaches $1/(g_m + g_{mb})$, and hence the input pole can be defined as

$$\omega_{p,in} = \frac{1}{\left(R_S \middle\| \frac{1}{g_m + g_{mb}}\right) C_{in}} \tag{6.75}$$

Why does Z_{in} become independent of C_L at high frequencies? This is because C_L lowers the voltage gain of the circuit, thereby suppressing the effect of the negative resistance introduced by the Miller effect through r_O (Fig. 6.7). In the limit, C_L shorts the output node to ground, and r_O affects the input impedance negligibly.

Our analysis of the CG frequency response has assumed a zero impedance in series with the gate. In practice, the bias network providing the gate voltage exhibits a finite impedance, altering the frequency response. Shown in Fig. 6.33(a) is an example in which this impedance is modeled by a resistor, R_G. If all of the device capacitances are included here, the circuit's transfer function is of third order. For simplicity, we consider only C_{GS} here and only C_{GD} in Appendix B. From the equivalent circuit in Fig. 6.33(b),[4] we have $g_m V_1 = -V_{out}/R_D$, and hence $V_1 = -V_{out}/(g_m R_D)$. The current flowing through R_S is equal to $V_1 C_{GS} s + g_m V_1 = -(C_{GS} s + g_m)V_{out}/(g_m R_D)$, and that through R_G equal to $V_1 C_{GS} s = -C_{GS} s V_{out}/(g_m R_D)$. Writing a KVL around the input network, we have

$$V_{in} - (C_{GS} + g_m)\frac{V_{out}}{g_m R_D} R_S + \frac{V_{out}}{g_m R_D} - C_{GS} s \frac{V_{out}}{g_m R_D} R_G = 0 \tag{6.76}$$

It follows that

$$\frac{V_{out}}{V_{in}} = \frac{g_m R_D}{(R_G + R_S) C_{GS} s + 1 + g_m R_S} \tag{6.77}$$

yielding a pole at

$$\omega_p = \frac{1 + g_m R_S}{(R_G + R_S) C_{GS}} \tag{6.78}$$

Thus, R_G directly adds to R_S in this case, lowering the pole magnitude.

[4] Channel-length modulation and body effect are neglected here.

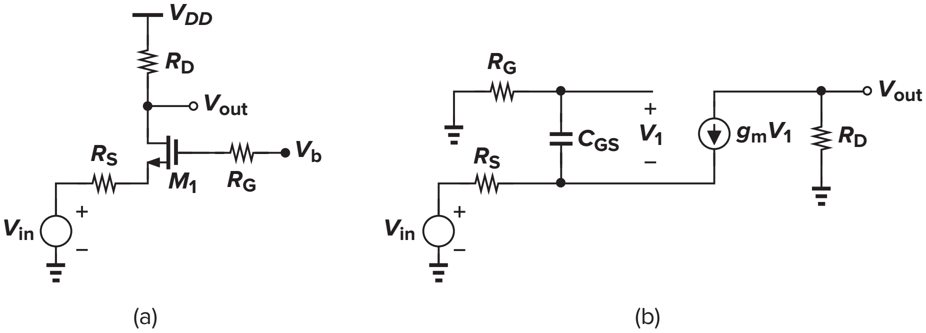


Figure 6.33 (a) CG stage with resistance in series with gate, and (b) equivalent circuit.

If a common-gate stage is driven by a relatively large source impedance, then the output impedance of the circuit drops at high frequencies. This effect is better described in the context of cascode circuits.

6.5 ■ Cascode Stage

As explained in Chapter 3, cascoding proves beneficial in increasing the voltage gain of amplifiers and the output impedance of current sources while providing shielding as well. The invention of the cascode (in the vacuum tube era), however, was motivated by the need for high-frequency amplifiers with a relatively high input impedance. Viewed as a cascade of a common-source stage and a common-gate stage, a cascode circuit offers the speed of the latter—by suppressing the Miller effect—and the input impedance of the former.

Let us consider the cascode shown in Fig. 6.34, first identifying all of the device capacitances. At node A, C_{GS1} is connected to ground and C_{GD1} to node X. At node X, C_{DB1}, C_{SB2}, and C_{GS2} are tied to ground, and at node Y, C_{DB2}, C_{GD2}, and C_L are connected to ground. The Miller effect of C_{GD1} is determined by the gain from A to X. As an approximation, we use the low-frequency value of this gain, which for low values of R_D (or negligible channel-length modulation) is equal to $-g_{m1}/(g_{m2}+g_{mb2})$. Thus, if M_1 and M_2 have roughly equal dimensions, C_{GD1} is multiplied by approximately 2 rather than the large

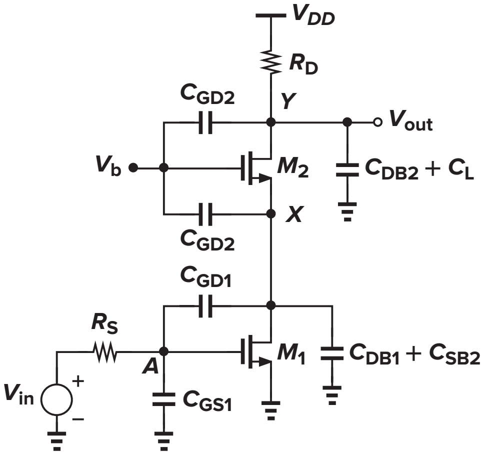


Figure 6.34 High-frequency model of a cascode stage.

voltage gain in a simple common-source stage. We therefore say that the Miller effect is less significant in cascode amplifiers than in common-source stages. The pole associated with node A is estimated as

$$\omega_{p,A} = \frac{1}{R_S\left[C_{GS1} + \left(1 + \frac{g_{m1}}{g_{m2} + g_{mb2}}\right)C_{GD1}\right]} \tag{6.79}$$

We can also attribute a pole to node X. The total capacitance at this node is roughly equal to $2C_{GD1} + C_{DB1} + C_{SB2} + C_{GS2}$, giving a pole

$$\omega_{p,X} = \frac{g_{m2} + g_{mb2}}{2C_{GD1} + C_{DB1} + C_{SB2} + C_{GS2}} \tag{6.80}$$

How does this pole compare with $2\pi f_T \approx g_{m2}/C_{GS2}$? The other capacitances in the denominator reduce the magnitude of $\omega_{p,X}$ to roughly $2\pi f_T/2$. Finally, the output node yields a third pole:

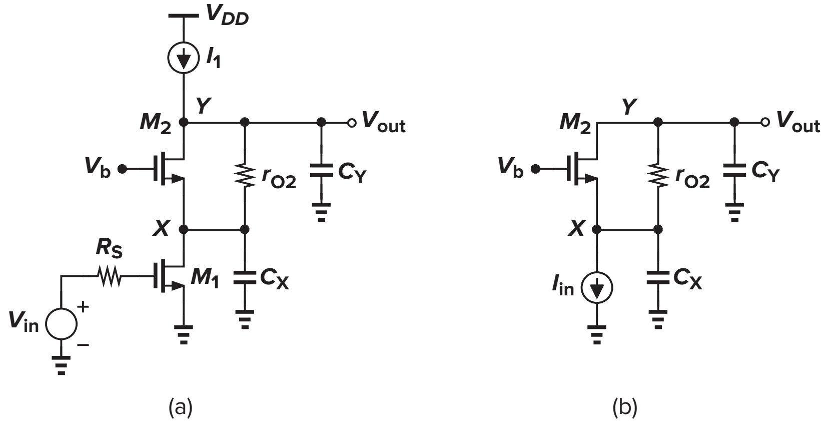
$$\omega_{p,Y} = \frac{1}{R_D(C_{DB2} + C_L + C_{GD2})} \tag{6.81}$$

The relative magnitudes of the three poles in a cascode circuit depend on the actual design parameters, but $\omega_{p,X}$ is typically quite a lot higher than the other two.

But what if R_D in Fig. 6.34 is replaced by a current source so as to achieve a higher dc gain? We know from Chapter 3 that the impedance seen at node X reaches high values if the load impedance at the drain of M_2 is large. For example, Eq. (3.117) predicts that the pole at node X may be quite a lot lower than $(g_{m2} + g_{mb2})/C_X$ if R_D itself is the output impedance of a PMOS cascode current source. Interestingly, however, the overall transfer function is negligibly affected by this phenomenon. This can be better seen by an example.

▶ **Example 6.16**

Consider the cascode stage shown in Fig. 6.35(a), where the load resistor is replaced by an ideal current source. Neglecting the capacitances associated with M_1, representing V_{in} and M_1 by a Norton equivalent as in Fig. 6.35(b), and assuming $\gamma = 0$, compute the transfer function.



(a) (b)

Figure 6.35 Simplified model of a cascode stage.

Solution

Since the current through C_X is equal to $-V_{out}C_Y s - I_{in}$, we have $V_X = -(V_{out}C_Y s + I_{in})/(C_X s)$, and the small-signal drain current of M_2 is $-g_{m2}(-V_{out}C_Y s - I_{in})/(C_X s)$. The current through r_{O2} is then equal to $-V_{out}C_Y s - g_{m2}(V_{out}C_Y s + I_{in})/(C_X s)$. Noting that V_X plus the voltage drop across r_{O2} is equal to V_{out}, we write

$$-r_{O2}\left[(V_{out}C_Y s + I_{in})\frac{g_{m2}}{C_X s} + V_{out}C_Y s\right] - (V_{out}C_Y s + I_{in})\frac{1}{C_X s} = V_{out} \tag{6.82}$$

That is

$$\frac{V_{out}}{I_{in}} = -\frac{g_{m2}r_{O2} + 1}{C_X s} \cdot \frac{1}{1 + (1 + g_{m2}r_{O2})\dfrac{C_Y}{C_X} + C_Y r_{O2} s} \tag{6.83}$$

which, for $g_{m2}r_{O2} \gg 1$ and $g_{m2}r_{O2}C_Y/C_X \gg 1$ (i.e., $C_Y > C_X$), reduces to

$$\frac{V_{out}}{I_{in}} \approx -\frac{g_{m2}}{C_X s}\frac{1}{\dfrac{C_Y}{C_X}g_{m2} + C_Y s} \tag{6.84}$$

and hence

$$\frac{V_{out}}{V_{in}} = -\frac{g_{m1}g_{m2}}{C_Y C_X s}\frac{1}{g_{m2}/C_X + s} \tag{6.85}$$

The magnitude of the pole at node X is still given by g_{m2}/C_X. This is because at high frequencies (as we approach this pole), C_Y shunts the output node, dropping the gain and suppressing the Miller effect of r_{O2}.

If a cascode structure is used as a current source, then the variation of its output impedance with frequency is of interest. Neglecting C_{GD1} and C_Y in Fig. 6.35(a), we have

$$Z_{out} = (1 + g_{m2}r_{O2})Z_X + r_{O2} \tag{6.86}$$

where $Z_X = r_{O1}||(C_X s)^{-1}$. Thus, Z_{out} contains a pole at $(r_{O1}C_X)^{-1}$ and falls at frequencies higher than this value.

6.6 ■ Differential Pair

The versatility of differential pairs and their extensive use in analog systems motivate us to characterize their frequency response for both differential and common-mode signals.

6.6.1 Differential Pair with Passive Loads

Consider the simple differential pair shown in Fig. 6.36(a), with the differential half circuit and the common-mode equivalent circuit depicted in Figs. 6.36(b) and (c), respectively. For differential signals, the response is identical to that of a common-source stage, exhibiting Miller multiplication of C_{GD}. Note that since $+V_{in2}/2$ and $-V_{in2}/2$ are multiplied by the same transfer function, the number of poles in V_{out}/V_{in} is equal to that of each path (rather than the sum of the number of poles in the two paths).

For common-mode signals, the total capacitance at node P in Fig. 6.36(c) determines the high-frequency gain. Arising from C_{GD3}, C_{DB3}, C_{SB1}, and C_{SB2}, this capacitance can be quite substantial if M_1–M_3 are wide transistors. For example, limited voltage headroom often necessitates that W_3 be so large

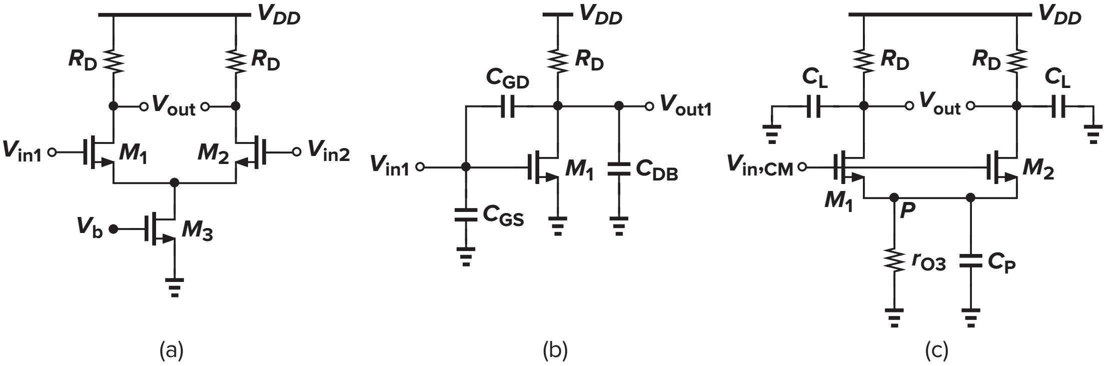


Figure 6.36 (a) Differential pair; (b) half-circuit equivalent; (c) equivalent circuit for common-mode inputs.

that M_3 does not require a large drain-source voltage for operating in the saturation region. If only the mismatch between M_1 and M_2 is considered, the high-frequency common-mode gain can be calculated with the aid of Eq. (4.53). We replace r_{O3} with $r_{O3}\|[1/(C_P s)]$ and R_D by $R_D\|[1/(C_L s)]$, where C_L denotes the total capacitance seen at each output node.[5] Thus,

$$A_{v,CM} = -\frac{\Delta g_m \left[R_D \middle\| \left(\frac{1}{C_L s}\right)\right]}{(g_{m1}+g_{m2})\left[r_{O3}\middle\|\left(\frac{1}{C_P s}\right)\right]+1} \tag{6.87}$$

This result suggests that the common-mode rejection of the circuit degrades considerably at high frequencies. In fact, writing the CMRR from Chapter 4 for this case gives

$$\text{CMRR} \approx \frac{g_m}{\Delta g_m}\left[1+2g_m\left(r_{O3}||\frac{1}{C_P s}\right)\right] \tag{6.88}$$

$$\approx \frac{g_m}{\Delta g_m}\frac{r_{O3}C_P s+1+2g_m r_{O3}}{r_{O3}C_P s+1} \tag{6.89}$$

where $g_m = (g_{m1}+g_{m2})/2$. We observe that this transfer function contains a zero at $(1+2g_{m3}r_{O3})/(r_{O3}C_P)$ and a pole at $1/(r_{O3}C_P)$. Since $2g_{m3}r_{O3} \gg 1$, the magnitude of the zero is much greater than the pole and approximately equal to $2g_{m3}/C_P$. The CMRR response thus appears as shown in Fig. 6.37.

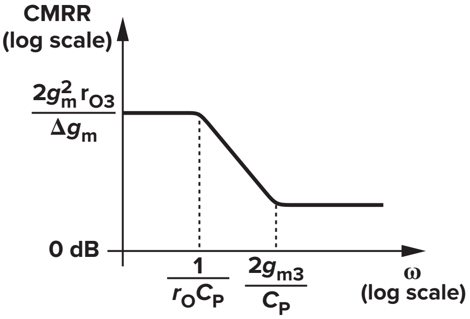


Figure 6.37 CMRR for a differential pair vs. frequency.

[5] For simplicity, channel-length modulation, body effect, and other capacitances are neglected.

As illustrated in Fig. 6.38, if the supply voltage contains high-frequency noise and the circuit exhibits mismatches, the resulting common-mode disturbance at node P translates to a differential noise component at the output. This effect becomes more pronounced as the noise frequency exceeds $1/(2\pi r_{O3} C_P)$.

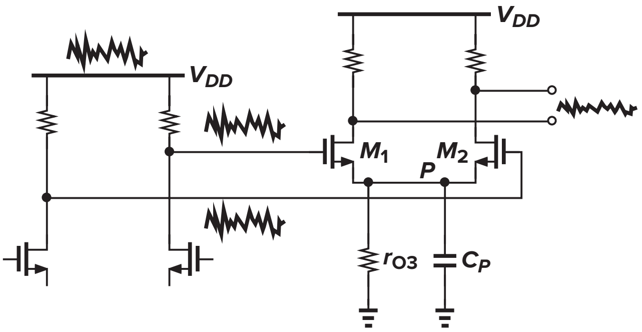


Figure 6.38 Effect of high-frequency supply noise in differential pairs.

We should emphasize that the circuit of Fig. 6.36(a) suffers from a trade-off between voltage headroom and CMRR. To minimize the headroom consumed by M_3, its width is maximized, introducing substantial capacitance at the sources of M_1 and M_2 and degrading the high-frequency CMRR. The issue becomes more serious at low supply voltages.

We now study the frequency response of differential pairs with high-impedance loads. Shown in Fig. 6.39(a) is a fully differential implementation. As with the topology of Fig. 6.36, this circuit can be analyzed for differential and common-mode signals separately. Note that here C_L includes the drain junction capacitance and the gate-drain overlap capacitance of each PMOS transistor as well. Also,

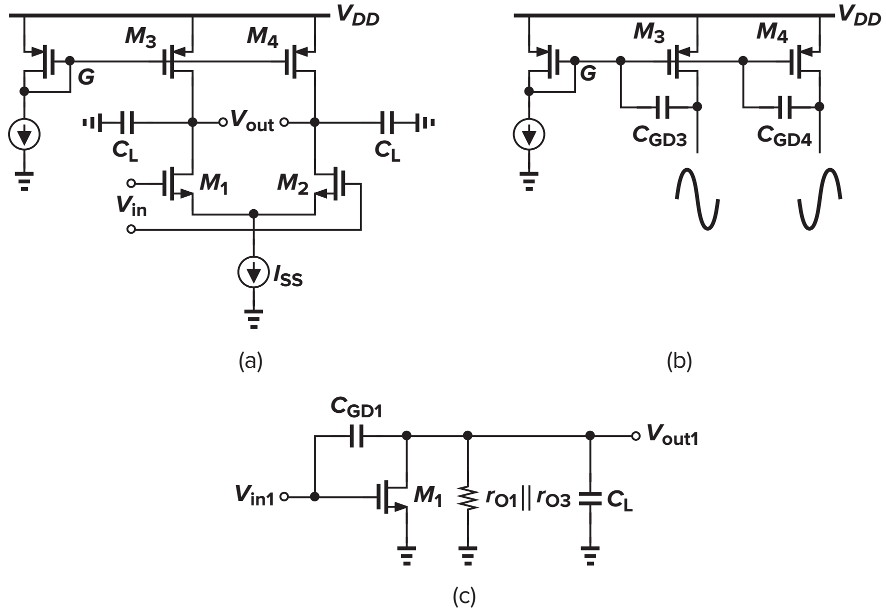


Figure 6.39 (a) Differential pair with current-source loads; (b) effect of differential swings at node G; (c) half-circuit equivalent.

as depicted in Fig. 6.39(b) for differential output signals, C_{GD3} and C_{GD4} conduct equal and opposite currents to node G, making this node an ac ground. (In practice, node G is still bypassed to ground by means of a capacitor.)

The differential half circuit is depicted in Fig. 6.39(c), with the output resistance of M_1 and M_3 shown explicitly. This topology implies that Eq. (6.30) can be applied to this circuit if R_L is replaced by $r_{O1}\|r_{O3}$. In practice, the relatively high value of this resistance makes the output pole, given by $[(r_{O1}\|r_{O3})C_L]^{-1}$, the "dominant" pole. We return to this observation in Chapter 10. The common-mode behavior of the circuit is similar to that of Fig. 6.36(c).

6.6.2 Differential Pair with Active Load

Let us now consider a differential pair with an active current mirror (Fig. 6.40). How many poles does this circuit have? In contrast to the fully differential configuration of Fig. 6.39(a), this topology contains two signal paths with *different* transfer functions. The path consisting of M_3 and M_4 includes a pole at node E, approximately given by g_{m3}/C_E, where C_E denotes the total capacitance from E to ground. This capacitance arises from C_{GS3}, C_{GS4}, C_{DB3}, C_{DB1}, and the Miller effect of C_{GD1} and C_{GD4}. Even if only C_{GS3} and C_{GS4} are considered, the severe trade-off between g_m and C_{GS} of PMOS devices results in a pole that impacts the performance of the circuit. The pole associated with node E is called a "mirror pole." Note that, as with the circuit of Fig. 6.39(a), both signal paths shown in Fig. 6.40 contain a pole at the output node.

In order to estimate the frequency response of the differential pair with an active current mirror, we construct the simplified model depicted in Fig. 6.41(a), where all other capacitances are neglected.

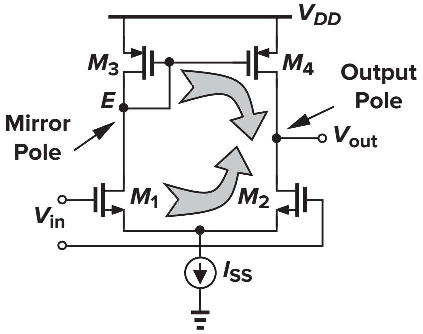


Figure 6.40 High-frequency behavior of differential pair with active current mirror.

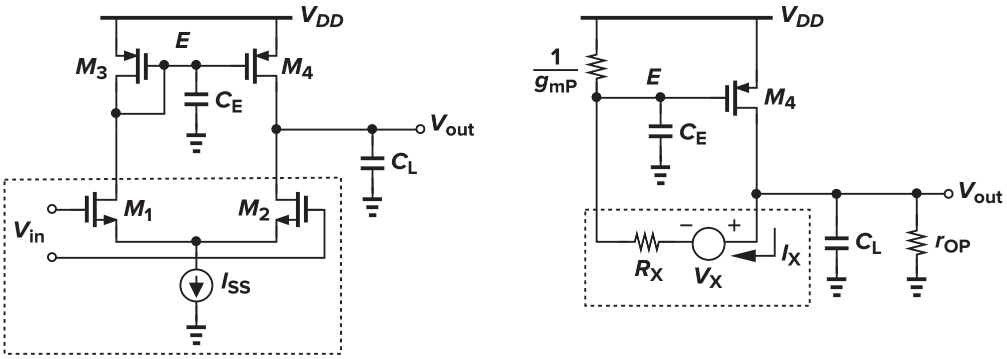


Figure 6.41 (a) Simplified high-frequency model of differential pair with active current mirror; (b) circuit of (a) with a Thevenin equivalent.

Replacing V_{in}, M_1, and M_2 by a Thevenin equivalent, we arrive at the circuit of Fig. 6.41(b), where $V_X = g_{mN}r_{ON}V_{in}$ and $R_X = 2r_{ON}$ (why?). Here, the subscripts P and N refer to PMOS and NMOS devices, respectively, and we have assumed that $1/g_{mP} \ll r_{OP}$. The small-signal voltage at E is equal to

$$V_E = (V_{out} - V_X)\frac{\dfrac{1}{C_E s + g_{mP}}}{\dfrac{1}{C_E s + g_{mP}} + R_X} \tag{6.90}$$

and the small-signal drain current of M_4 is $g_{m4}V_E$. Noting that $-g_{m4}V_E - I_X = V_{out}(C_L s + r_{OP}^{-1})$, we have

$$\frac{V_{out}}{V_{in}} = \frac{g_{mN}r_{ON}(2g_{mP} + C_E s)r_{OP}}{2r_{OP}r_{ON}C_E C_L s^2 + [(2r_{ON} + r_{OP})C_E + r_{OP}(1 + 2g_{mP}r_{ON})C_L]s + 2g_{mP}(r_{ON} + r_{OP})} \tag{6.91}$$

Since the mirror pole is typically much higher in magnitude than the output pole, we can utilize the results of Eq. (6.34) to write

$$\omega_{p1} \approx \frac{2g_{mP}(r_{ON} + r_{OP})}{(2r_{ON} + r_{OP})C_E + r_{OP}(1 + 2g_{mP}r_{ON})C_L} \tag{6.92}$$

Neglecting the first term in the denominator and assuming that $2g_{mP}r_{ON} \gg 1$, we have

$$\omega_{p1} \approx \frac{1}{(r_{ON}\|r_{OP})C_L} \tag{6.93}$$

an expected result. The second pole is then given by

$$\omega_{p2} \approx \frac{g_{mP}}{C_E} \tag{6.94}$$

which is also expected.

An interesting point revealed by Eq. (6.91) is a zero with a magnitude of $2g_{mP}/C_E$ in the left half plane. The appearance of such a zero can be understood by noting that the circuit consists of a "slow path" (M_1, M_3, and M_4) in parallel with a "fast path" (M_1 and M_2). Representing the two by $A_0/[(1 + s/\omega_{p1})(1 + s/\omega_{p2})]$ and $A_0/(1 + s/\omega_{p1})$, respectively, we have

$$\frac{V_{out}}{V_{in}} = \frac{A_0}{1 + s/\omega_{p1}}\left(\frac{1}{1 + s/\omega_{p2}} + 1\right) \tag{6.95}$$

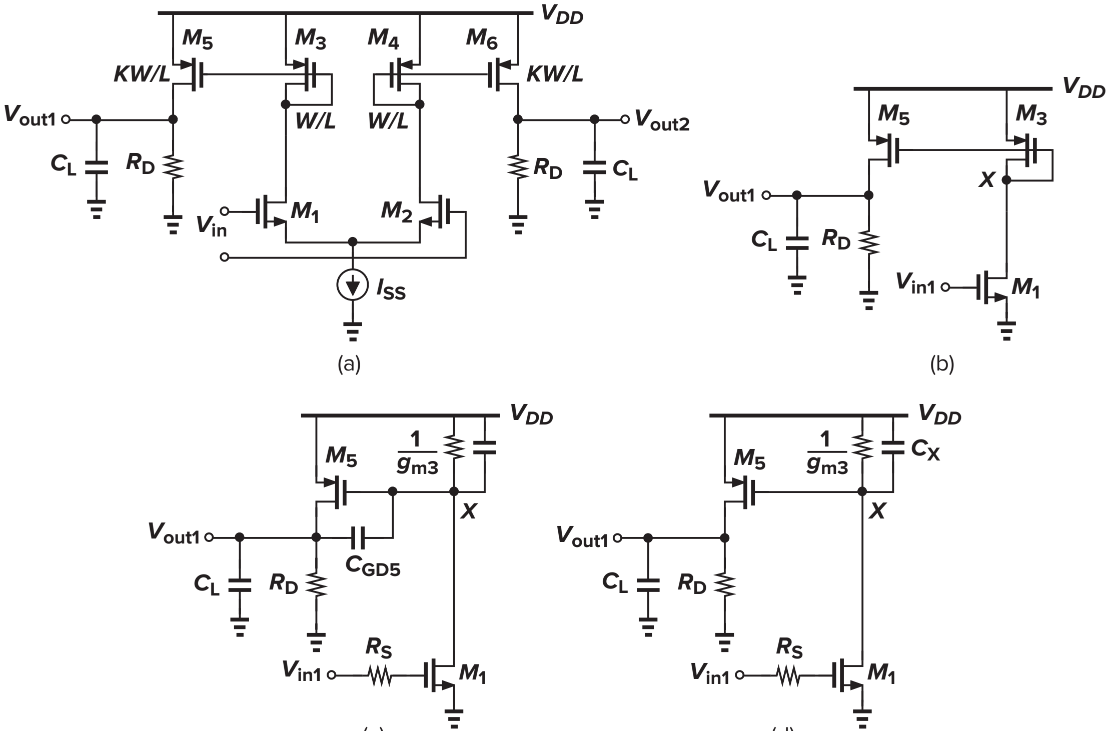
$$= \frac{A_0(2 + s/\omega_{p2})}{(1 + s/\omega_{p1})(1 + s/\omega_{p2})} \tag{6.96}$$

That is, the system exhibits a zero at $2\omega_{p2}$. The zero can also be obtained by the method of Fig. 6.18 (Problem 6.15).

Comparing the circuits of Figs. 6.39(a) and 6.40, we conclude that the former entails no mirror pole, another advantage of fully differential circuits over single-ended topologies.

▶ Example 6.17

Not all fully differential circuits are free from mirror poles. Figure 6.42(a) illustrates an example where current mirrors M_3–M_5 and M_4–M_6 "fold" the signal current. Estimate the low-frequency gain and the transfer function of this circuit.



(a) (b)

(c) (d)

Figure 6.42

Solution

Neglecting channel-length modulation and using the differential half circuit shown in Fig. 6.42(b), we observe that M_5 multiplies the drain current of M_3 by K, yielding an overall low-frequency voltage gain $A_v = g_{m1} K R_D$.

To obtain the transfer function, we utilize the equivalent circuit depicted in Fig. 6.42(c), including a source resistance R_S for completeness. To simplify calculations, we assume that $R_D C_L$ is relatively small so that the Miller multiplication of C_{GD5} can be approximated as $C_{GD5}(1 + g_{m5} R_D)$. The circuit thus reduces to that in Fig. 6.42(d), where $C_X \approx C_{GS3} + C_{GS5} + C_{DB3} + C_{GD5}(1 + g_{m5} R_D) + C_{DB1}$. The overall transfer function is then equal to V_X / V_{in1} multiplied by V_{out1}/V_X. The former is readily obtained from (6.30) by replacing R_D with $1/g_{m3}$ and C_{DB} with C_X, while the latter is

$$\frac{V_{out1}}{V_X}(s) = -g_{m5} R_D \frac{1}{1 + R_D C_L s} \tag{6.97}$$

Note that we have neglected the zero due to C_{GD5}.

6.7 ■ Gain-Bandwidth Trade-Offs

In many applications, we wish to maximize both the gain and the bandwidth of amplifiers. For example, optical communication receivers employ an amplifier that must achieve a high gain and a wide bandwidth. This section deals with gain-bandwidth trade-offs encountered in high-speed design. As shown in Fig. 6.43, we are interested in both the −3-dB bandwidth, ω_{-3dB}, and the "unity-gain" bandwidth, ω_u.

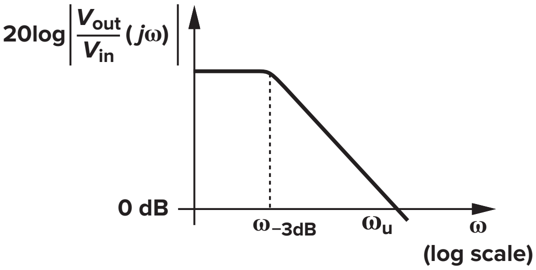


Figure 6.43 Frequency response showing −3-dB and unity-gain bandwidths.

6.7.1 One-Pole Circuits

In some circuits, the load capacitance seen at the output node produces a dominant pole, allowing a one-pole approximation. That is, we can say that the −3-dB bandwidth is equal to the pole frequency. For example, the CS stage of Fig. 6.44 exhibits an output pole given by $\omega_p = [(r_{O1}||r_{O2})C_L]^{-1}$ if other capacitances are neglected. Noting that the low-frequency gain is equal to $|A_0| = g_{m1}(r_{O1}||r_{O2})$, we define the "gain-bandwidth" product (GBW) as

$$\text{GBW} = A_0\omega_p \tag{6.98}$$

$$= g_{m1}(r_{O1}||r_{O2})\frac{1}{2\pi(r_{O1}||r_{O2})C_L} \tag{6.99}$$

$$= \frac{g_{m1}}{2\pi C_L} \tag{6.100}$$

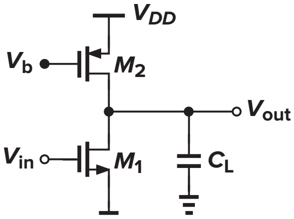


Figure 6.44 CS stage with one pole.

As an example, if $g_{m1} = (100\ \Omega)^{-1}$ and $C_L = 50$ fF, then GBW = 31.8 GHz. For a one-pole system, the gain-bandwidth product is approximately equal to the unity-gain bandwidth; this can be seen by writing

$$\frac{A_0}{\sqrt{1+(\frac{\omega_u}{\omega_p})^2}} = 1 \tag{6.101}$$

and hence

$$\omega_u = \sqrt{A_0^2 - 1}\omega_p \tag{6.102}$$

$$\approx A_0\omega_p \tag{6.103}$$

if $A_0^2 \gg 1$.

▶ **Example 6.18**

Does cascoding increase the GBW product? Assume that the output pole is dominant.

Solution

No, it does not. Equation (6.100) suggests that the GBW product is independent of the output resistance. More specifically, if cascoding in Fig. 6.44 raises the output impedance by a factor of K, then $|A_0|$ ($= G_m R_{out}$) rises and ω_p falls, both by a factor of K, yielding a constant GBW product. ◀

6.7.2 Multi-Pole Circuits

It is possible to increase the GBW product by cascading two or more gain stages. Consider the amplifier shown in Fig. 6.45, where, for simplicity, we assume that the two stages are identical and neglect other capacitances. Associating one pole with each node, we write the transfer function as $(V_{out}/V_X)(V_X/V_{in})$:

$$\frac{V_{out}}{V_{in}} = \frac{A_0^2}{(1+\dfrac{s}{\omega_p})^2} \tag{6.104}$$

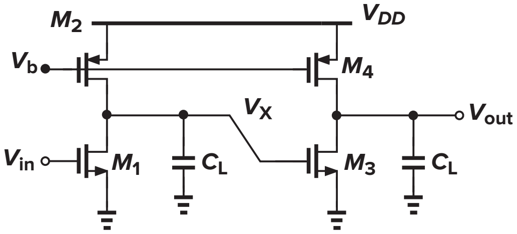


Figure 6.45 Cascaded CS stages.

where $A_0 = g_{mN}(r_{ON}||r_{OP})$ and $\omega_p = [(r_{ON}||r_{OP})C_L]^{-1}$. To obtain the -3-dB bandwidth, we equate the magnitude of V_{out}/V_{in} to $A_0^2/\sqrt{2}$:

$$\frac{A_0^2}{1+\dfrac{\omega_{-3dB}^2}{\omega_p^2}} = \frac{A_0^2}{\sqrt{2}} \tag{6.105}$$

and

$$\omega_{-3dB} = \sqrt{\sqrt{2}-1}\omega_p \tag{6.106}$$

$$\approx 0.64\omega_p \tag{6.107}$$

The GBW product thus rises to

$$\text{GBW} = \sqrt{\sqrt{2}-1}A_0^2\omega_p \tag{6.108}$$

a factor of $0.64A_0$ greater than that in Eq. (6.103). Of course, the power consumption is doubled.

While raising the GBW product, cascading *reduces* the bandwidth, as evidenced by Eq. (6.107). In fact, we prove in Problem 6.25 that for N identical stages,

$$\omega_{-3dB} = \sqrt{\sqrt[N]{2}-1}\omega_p \tag{6.109}$$

observing a steady decline in the bandwidth as N increases. Another disadvantage of cascading is that the resulting multiple poles lead to instability if the circuit is placed in a negative-feedback loop (Chapter 10).

6.8 ■ Appendix A: Extra Element Theorem

Introduced by Middlebrook [1], the extra element theorem (EET) proves useful in calculating some transfer functions. Suppose the transfer function of a circuit is known and denoted by $H(s)$. Now, as shown in Fig. 6.46(a), we add an extra impedance Z_1 between two nodes of the circuit. We wish to determine the new transfer function, $G(s)$. Middlebrook proves that

$$G(s) = H(s)\frac{1 + \dfrac{Z_{out,0}}{Z_1}}{1 + \dfrac{Z_{in,0}}{Z_1}} \tag{6.110}$$

i.e., the original transfer function is multiplied by a "correction factor." The terms $Z_{out,0}$ and $Z_{in,0}$ are quantities measured between nodes A and B in the absence of Z_1. The former is computed as depicted in Fig. 6.46(b): we apply a voltage source between A and B while V_{in} is present and choose their values so that $V_{out} = 0$; then $Z_{out,0} = V_1/I_1$. This calculation appears rather complex and unintuitive, but, as shown below, it is in fact quite simple. We should also remark that $Z_{out,0}$ is not an impedance in the standard sense because it is obtained with a finite V_{in}. The latter, $Z_{in,0}$, is simply equal to the impedance seen between A and B when $V_{in} = 0$ [Fig. 6.46(c)].

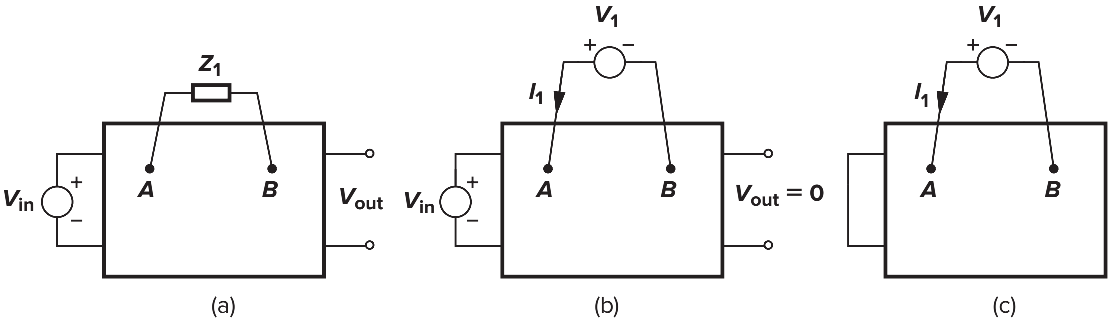


Figure 6.46 (a) Circuit with extra parallel element, Z_1, (b) $Z_{out,0}$ calculation, and (c) $Z_{in,0}$ calculation.

This theorem is particularly useful for frequency-response analysis because we can begin with no capacitances in the circuit, find $H(s)$ as the low-frequency gain, add the capacitors one by one, and calculate the correction factors. Note that $H(s)$ cannot be zero or infinity because the EET's proof relies on division by $H(s)$.

▶ **Example 6.19**

Using the EET, find the transfer function of the circuit in Fig. 6.47(a).

Solution

We first consider the circuit without C_F and write $H(s) = -g_m(R_D||r_O)$. Next, we find $Z_{out,0}$ using the setup shown in Fig. 6.47(b), exploiting the condition that V_{out} is zero and so is the current through R_D. Since $V_{out} = 0$, we have $V_{GS} = V_1$ and $I_1 = -g_m V_{GS} = -g_m V_1$. That is, $Z_{out,0} = -1/g_m$. Note that we resisted the temptation to

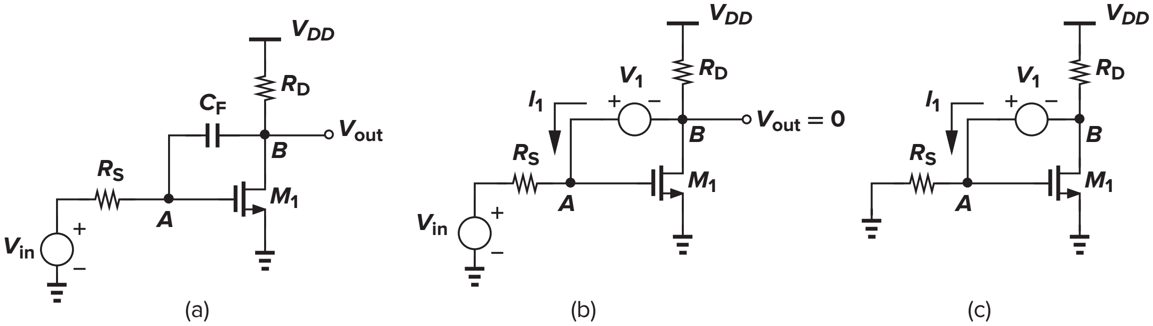


Figure 6.47

write equations involving V_{in}. Also, the negative sign of $Z_{out,0}$ does not imply a negative impedance between A and B because $V_{in} \neq 0$.

For $Z_{in,0}$, we have from Fig. 6.47(c), $V_A = I_1 R_S = V_{GS}$. A KCL at node B gives the current through R_D as $g_m I_1 R_S + I_1$, and a KVL across R_D, V_1, and R_S leads to $I_1 R_D(1 + g_m R_S) - V_1 + I_1 R_S = 0$. It follows that $Z_{in,0} = (1 + g_m R_S)R_D + R_S = (1 + g_m R_D)R_S + R_D$ and

$$G(s) = -g_m(R_D||r_O)\frac{1 - \dfrac{1}{g_m}C_F s}{1 + [(1 + g_m R_D)R_S + R_D]C_F s} \tag{6.111}$$

We see that the EET beautifully predicts the zero and the pole produced by C_F.

▶ Example 6.20

Repeat the above example while including both C_F and a capacitor, C_B, from node B to ground.

Solution

Since we have already obtained the transfer function with C_F present, we must seek the $Z_{out,0}$ and $Z_{in,0}$ corresponding to C_B. The arrangement depicted in Fig. 6.48(a) suggests that $Z_{out,0} = 0$ because the drain voltage must be zero while V_1 is not, requiring an infinite current to flow through V_1.

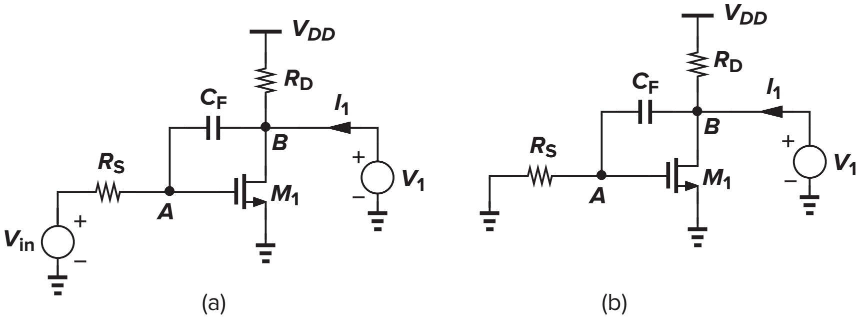


Figure 6.48

For $Z_{in,0}$, we note from Fig. 6.48(b) that $V_{GS} = V_1 R_S C_F s/(R_S C_F s + 1)$ and the current flowing through C_F is equal to $V_1/[(C_F s)^{-1} + R_S]$. A KCL at the drain node gives

$$\frac{V_1}{R_D} + \frac{V_1 C_F s}{R_S C_F s + 1} + g_m V_1 \frac{R_S C_F s}{R_S C_F s + 1} = I_1 \tag{6.112}$$

Thus,

$$Z_{in,0} = \frac{R_D(R_S C_F s + 1)}{[R_S(1 + g_m R_D) + R_D]C_F s + 1} \tag{6.113}$$

Using Eq. (6.111), we write the new transfer function as

$$\begin{aligned} G(s) &= -g_m(R_D||r_O)\frac{1 - \frac{C_F}{g_m}s}{1 + [(1 + g_m R_D)R_S + R_D]C_F s}\frac{1}{1 + \frac{R_D(R_S C_F s + 1)C_B s}{[R_S(1 + g_m R_D) + R_D]C_F s + 1}} \\ &= -g_m(R_D||r_O)\frac{1 - \frac{C_F}{g_m}s}{[R_S(1 + g_m R_D) + R_D]C_F s + R_D(R_S C_F s + 1)C_B s + 1} \end{aligned} \tag{6.114}$$

The EET can also be expressed for series elements [1]. That is, if the transfer function of a circuit is $H(s)$ before we insert an element, Z_1, in series with a branch, then the new transfer function is given by [1]

$$G(s) = H(s)\frac{1 + \dfrac{Z_1}{Z_{out,0}}}{1 + \dfrac{Z_1}{Z_{in,0}}} \tag{6.115}$$

6.9 ■ Appendix B: Zero-Value Time Constant Method

Our analysis of frequency response in this chapter reveals considerable mathematical labor when the number of poles exceeds two. In some cases, we are content with estimating the dominant pole—if one exists—or the −3-dB bandwidth of the circuit. The "zero-value time constant" (ZVTC) method provides an approximation of these quantities. It also proves useful as an additional analysis tool.

Before delving into the ZVTC method, let us make an observation. Suppose a circuit contains one capacitor and no other storage elements and we wish to determine the pole of the system [Fig. 6.49(a)]. We can derive the transfer function $V_{out}(s)/V_{in}(s)$ and examine its denominator, $D(s)$. Alternatively, as shown in Fig. 6.49(b), we can set the input to zero, compute the resistance, R_1, seen by C_1, and express

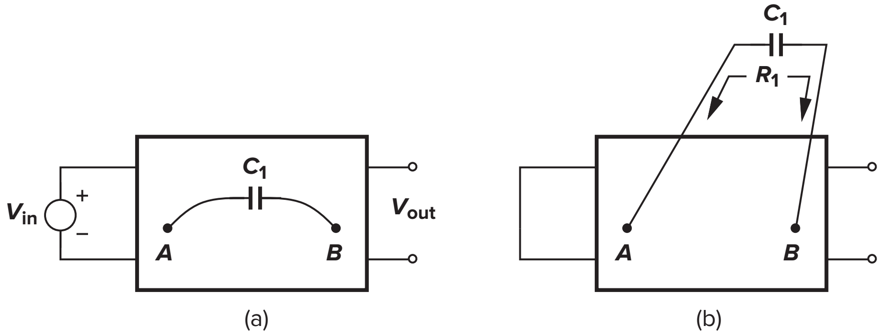


Figure 6.49 (a) General circuit containing one capacitor, and (b) resistance seen by C_1.

the pole as $1/(R_1C_1)$. In Problem 6.23, we prove why this is true, but the important point here is that this method often simplifies the analysis.

▶ **Example 6.21**

A CG stage contains a resistance R_G in series with the gate [Fig. 6.50(a)]. If only C_{GD} is considered, determine the pole frequency.

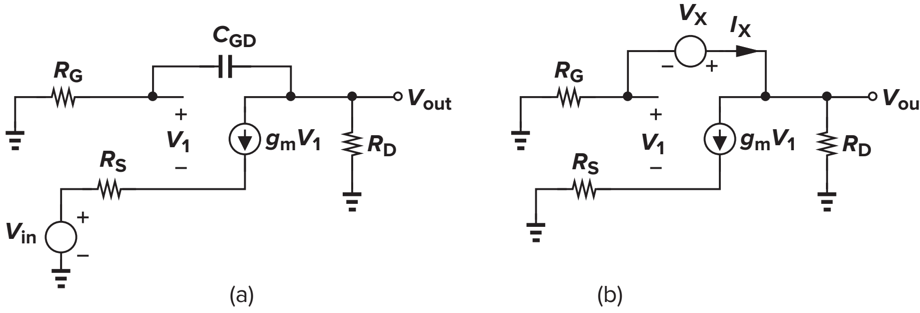


Figure 6.50

Solution

As illustrated in Fig. 6.50(b), we remove C_{GD}, set V_{in} to zero, and apply a voltage (or current) source to measure the resistance seen by this capacitor. The voltage across R_S is equal to $g_mV_1R_S$, yielding

$$g_mV_1R_S + V_1 = -I_XR_G \tag{6.116}$$

and hence $V_1 = -I_XR_G/(1+g_mR_S)$. Since the current flowing through R_D is equal to $I_X - g_mV_1$, we have

$$-I_XR_G + V_X = (I_X - g_mV_1)R_D \tag{6.117}$$

Substituting for V_1, we obtain

$$\frac{V_X}{I_X} = R_D + \left(\frac{g_mR_D}{1+g_mR_S} + 1\right)R_G = R_{eq} \tag{6.118}$$

The pole is given by $1/(R_{eq}C_{GD})$. The reader is encouraged to determine the circuit's transfer function directly and compare the mathematical labor.

Interestingly, as a result of R_G, the resistance seen by C_{GD} rises from R_D to R_D plus a multiple of R_G, the multiple given by the low-frequency gain of the CG stage plus 1. It is also interesting to note that the circuit of Fig. 6.50(a) does not lend itself to Miller's approximation (why?).

◀

As our first step toward developing the ZVTC method, let us determine the transfer function of the simple second-order circuit shown in Fig. 6.51. Since the current through R_2 is equal to $V_{out}C_2s$, and hence $V_X = R_2V_{out}C_2s + V_{out}$, we obtain the current through C_1 as $V_XC_1s = (1+R_2C_2s)C_1sV_{out}$. This current and that through R_2 flow through R_1, producing a voltage drop equal to $R_1(1+R_2C_2s)C_1sV_{out} + R_1V_{out}C_2s$. Writing a KVL around V_{in}, R_1, R_2, and V_{out} gives

$$V_{in} = R_1(1+R_2C_2s)C_1sV_{out} + R_1C_2sV_{out} + R_2V_{out}C_2s + V_{out} \tag{6.119}$$

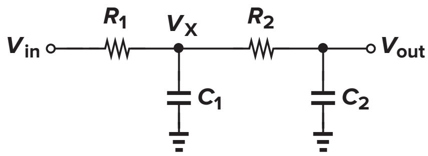


Figure 6.51 Second-order RC circuit.

It follows that

$$\frac{V_{out}}{V_{in}}(s) = \frac{1}{R_1R_2C_1C_2s^2 + [R_1C_1 + (R_1 + R_2)C_2]s + 1} \tag{6.120}$$

Recall from Sec. 6.2 that, if a dominant pole exists, then it is given by the inverse of the coefficient of s, B_s. We now focus on this coefficient, noting that it must have a *time* dimension and is therefore the sum of time constants. The first time constant, R_1C_1, contains a resistance equal to the resistance seen by C_1 *as if C_2 were zero*.[6] Similarly, the second time constant, $(R_1 + R_2)C_2$, arises from the resistance seen by C_2 as if C_1 were zero. We call R_1C_1 and $(R_1 + R_2)C_2$ “zero-value” time constants because each is obtained by setting the other capacitor to zero.

Can we generalize this result? That is, can we say that the dominant pole is given by the inverse of the sum of all of the zero-value time constants? We must first prove that, even for higher-order systems, the dominant pole is equal to the inverse of the coefficient of s in the denominator. Writing the denominator as

$$D(s) = \left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right)\cdots\left(1 + \frac{s}{\omega_{pn}}\right) \tag{6.121}$$

we recognize that the coefficient of s, B_s, is equal to $\omega_{p1}^{-1} + \omega_{p2}^{-1} + \cdots + \omega_{pn}^{-1}$, which reduces to ω_{p1}^{-1} if this pole is dominant.

Next, we must prove that B_s is equal to the sum of the zero-value time constants of the circuit. Assuming that the circuit contains only capacitors as storage elements,[7] we note that, since B_s has a time dimension, it can be expressed as

$$B_s = R_1C_1 + R_2C_2 + \cdots + R_nC_n \tag{6.122}$$

where R_1–R_n are unknown. Note that C_1–C_n denote the capacitors in the circuit, but R_1–R_n may represent physical resistors or equivalent resistances (e.g., $1/g_m$). How do we obtain R_1–R_n? If C_2–C_n are set to zero, the order of the system falls to 1, i.e., $D(s) = B_ss + 1 = R_1C_1s + 1$, where R_1 is the resistance seen by C_1. Similarly, if $C_1 = C_3 = \cdots = C_n = 0$, we have $D(s) = R_2C_2s + 1$, where R_2 is the resistance seen by C_2. Thus, the dominant pole is indeed equal to the inverse of the sum of the zero-value time constants. The reader is cautioned that, even though, $B_s = \omega_{p1}^{-1} + \omega_{p2}^{-1} + \cdots + \omega_{p1}^{-n} = R_1C_1 + R_2C_2 + \cdots + R_nC_n$, we cannot conclude that $\omega_{p1}^{-1} = R_1C_1$, $\omega_{p2}^{-1} = R_2C_2$, etc. Also, note that this method neglects the effect of zeros.

The ZVTC method proves useful if we wish to estimate the −3-dB bandwidth of a circuit. Depicted in Fig. 6.52, the idea is to approximate the frequency response by a one-pole system, and hence the time response by a single exponential. The following example illustrates this point.

[6] With $V_{in} = 0$.

[7] The analysis can be repeated for other types of storage elements as well.

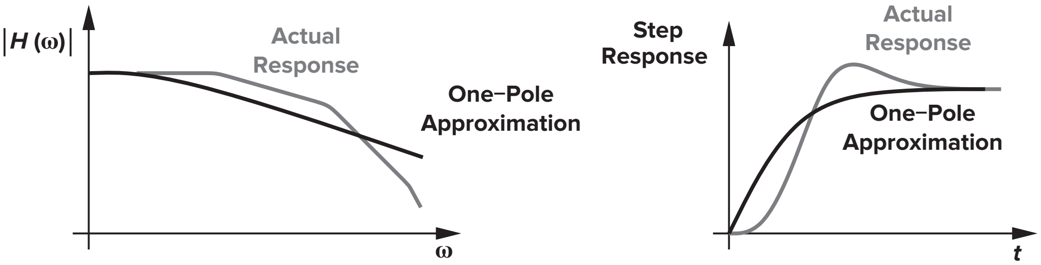


Figure 6.52 Approximation of the frequency and time responses by one-pole counterparts.

▶ Example 6.22

Estimate the -3-dB bandwidth of a resistively-degenerated common-source stage. Assume $\lambda = \gamma = 0$.

Solution

Shown in Fig. 6.53(a), the small-signal model is of third order,[8] providing little intuition. The zero-value time constant method can give a rough estimate of the circuit's bandwidth, thereby revealing the contribution of each capacitor.

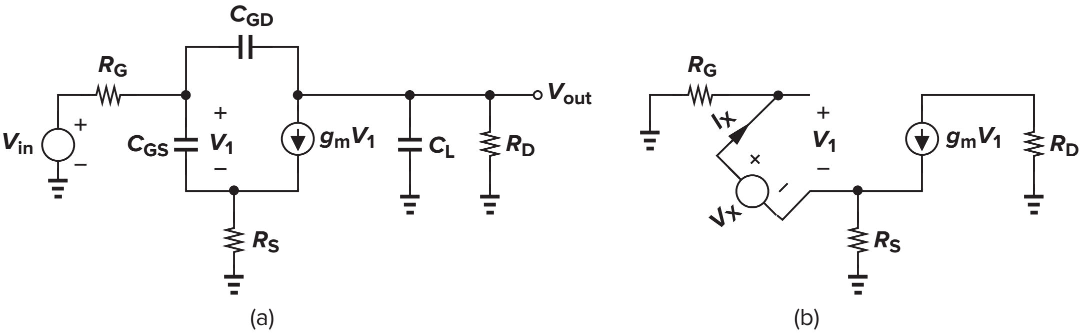


Figure 6.53

We begin with the time constant associated with C_{GS} and set C_{GD} and C_L to zero. As depicted in Fig. 6.53(b), the resistance seen by C_{GS} is V_X/I_X. We denote this resistance by R_{CGS}. Since $V_1 = V_X$ and the current flowing through R_S is equal to $g_mV_1 - I_X = g_mV_X - I_X$, we write a KVL as follows:

$$I_XR_G = V_X + (g_mV_X - I_X)R_S \tag{6.123}$$

obtaining

$$R_{CGS} = \frac{R_G + R_S}{1 + g_mR_S} \tag{6.124}$$

For the resistance seen by C_{GD}, we have from Example 6.21

$$R_{CGD} = R_D + \left(\frac{g_mR_D}{1 + g_mR_S} + 1\right)R_G \tag{6.125}$$

[8] This can be seen by observing that it is possible to impose three independent initial conditions across the three capacitors without violating KVL.

Finally, the resistance seen by C_L is simply equal to R_D. It follows that the −3-dB bandwidth is given by

$$\omega_{-3dB}^{-1} = \frac{R_G + R_S}{1 + g_m R_S} C_{GS} + \left[R_D + \left(\frac{g_m R_D}{1 + g_m R_S} + 1 \right) R_G \right] C_{GD} + R_D C_L \tag{6.126}$$

With no degeneration, this result reduces to Eq. (6.35). With a finite R_S, the effect of C_{GS} and R_G is reduced by a factor of $1 + g_m R_S$, albeit at the cost of voltage gain.

Example 6.23

Repeat the above example for a common-gate stage containing a gate resistance of R_G and a source resistance of R_S.

Solution

We draw the small-signal circuit as shown in Fig. 6.54. For the computation of zero-value time constants, the main input is set to zero. Thus, the resulting equivalent circuits are *identical* for CS and CG stages, yielding the same time constants and hence the same bandwidth. After all, the circuits in Figs. 6.53(a) and 6.54 are topologically identical and contain the same poles.

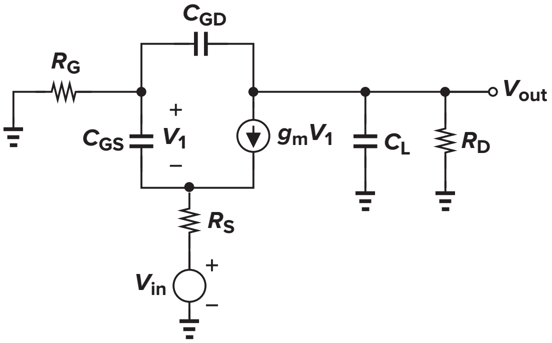
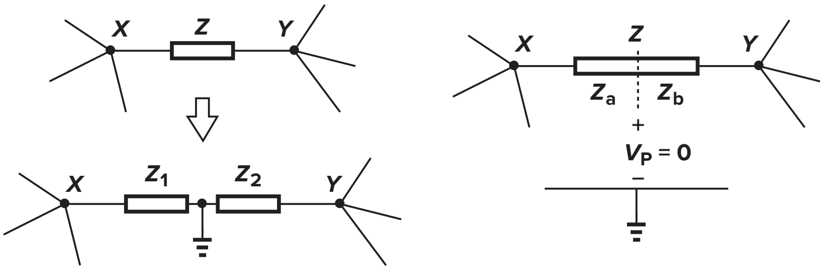


Figure 6.54

Does this result contradict our earlier assertion that the CG stage is free from the Miller effect? No, it does not. In a CG stage, we strive to avoid R_G, whereas in a CS stage, R_G represents the preceding circuit's output resistance and is inevitable.

6.10 ■ Appendix C: Dual of Miller's Theorem

In Miller's theorem (Fig. 6.2), we readily observe that $Z_1 + Z_2 = Z$. This is no coincidence, and it has interesting implications. Redrawing Fig. 6.2 as shown in Fig. 6.55(a), we surmise that since the point between Z_1 and Z_2 can be grounded, then if we "walk" from X toward Y along the impedance Z, the



(a) (b)

Figure 6.55 Illustration of Miller's theorem, identifying a local zero potential a long Z.

local potential drops to zero at some intermediate point [Fig. 6.55(b)]. Indeed, for $V_P = 0$, we have

$$\frac{Z_a}{Z_a + Z_b}(V_Y - V_X) + V_X = 0 \tag{6.127}$$

and, since $Z_a + Z_b = Z$,

$$Z_a = \frac{Z}{1 - V_Y/V_X} \tag{6.128}$$

Similarly,

$$Z_b = \frac{Z}{1 - V_X/V_Y} \tag{6.129}$$

In other words, $Z_1 (= Z_a)$ and $Z_2 (= Z_b)$ are such decompositions of Z that provide an intermediate node having a zero potential. For example, since in the common-source stage of Fig. 6.13, V_X and V_Y have opposite polarities, the potential falls to zero at some point "inside" C_{GD}.

The above observation explains the difficulty with the transformation depicted in Fig. 6.5. Drawing Fig. 6.55(b) for this case as in Fig. 6.56(a), we recognize that the circuit is still valid before node P is grounded because the current through $R_1 + R_2$ must equal that through $-R_2$. However, if, as shown in Fig. 6.56(b), node P is tied to ground, then the only current path between X and Y vanishes.

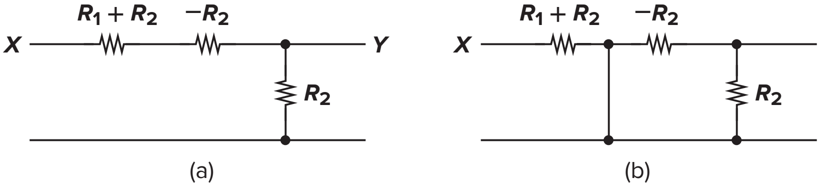


Figure 6.56 Resistive divider with decomposition of R_1.

The concept of a local zero potential along the floating impedance Z also allows us to develop the "dual" of Miller's theorem, i.e., decomposition in terms of admittances and current ratios. Suppose two loops carrying currents I_1 and I_2 share an admittance Y [Fig. 6.57(a)]. Then, if Y is properly decomposed into two *parallel* admittances Y_1 and Y_2, the *current* flowing between the two is zero [Fig. 6.57(b)], and the connection can be broken [Fig. 6.57(c)]. In Fig. 6.57(a), the voltage across Y is equal to $(I_1 - I_2)/Y$, and in Fig. 6.57(c), the voltage across Y_1 is I_1/Y_1. For the two circuits to be equivalent,

$$\frac{I_1 - I_2}{Y} = \frac{I_1}{Y_1} \tag{6.130}$$

and

$$Y_1 = \frac{Y}{1 - I_2/I_1} \tag{6.131}$$

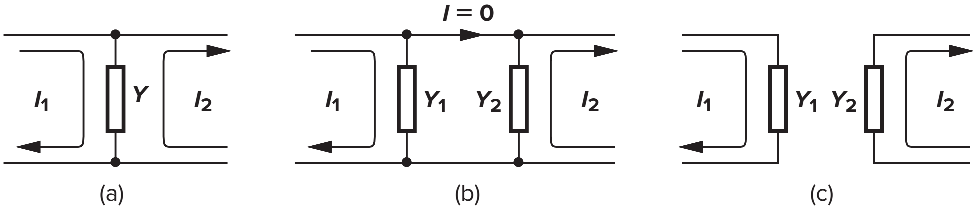


Figure 6.57 (a) Two loops sharing admittance Y, (b) decomposition of Y into Y_1 and Y_2 such that $I = 0$, (c) equivalent circuit.

Note the duality between this expression and $Z_1 = (1 - V_Y/V_X)Z$. We also have

$$Y_2 = \frac{Y}{1 - I_1/I_2} \tag{6.132}$$

References

[1] R. D. Middlebrook, "Null Double Injection and the Extra Element Theorem," *IEEE Trans. Circuits and Systems,* vol. 32, pp. 167–180, Aug. 1989.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation. All device dimensions are effective values and in microns.

6.1. In the circuit of Fig. 6.3(c), suppose the amplifier has a finite output resistance R_{out}.
(a) Explain why the output jumps *up* by ΔV before it begins to go down. This indicates the existence of a zero in the transfer function.
(b) Determine the transfer function and the step response without using Miller's theorem.

6.2. Repeat Problem 6.1 if the amplifier has an output resistance R_{out} and the circuit drives a load capacitance C_L.

6.3. The CS stage of Fig. 6.13 is designed with $(W/L)_1 = 50/0.5$, $R_S = 1$ kΩ, and $R_D = 2$ kΩ. If $I_{D1} = 1$ mA, determine the poles and the zero of the circuit.

6.4. Consider the CS stage of Fig. 6.16, where I_1 is realized by a PMOS device operating in saturation. Assume that $(W/L)_1 = 50/0.5$, $I_{D1} = 1$ mA, and $R_S = 1$ kΩ.
(a) Determine the aspect ratio of the PMOS transistor such that the maximum allowable output level is 2.6 V. What is the maximum peak-to-peak swing?
(b) Determine the poles and the zero.

6.5. A source follower employing an NFET with $W/L = 50/0.5$ and a bias current of 1 mA is driven by a source impedance of 10 kΩ. Calculate the equivalent inductance seen at the output.

6.6. Neglecting other capacitances, calculate the input impedance of each circuit shown in Fig. 6.58.

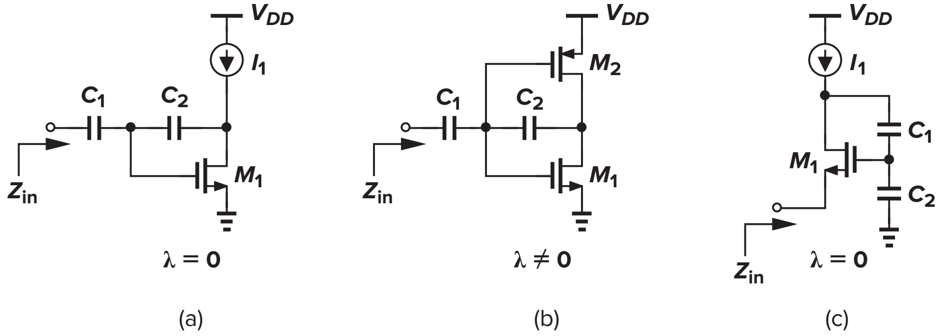


Figure 6.58

6.7. Estimate the poles of each circuit in Fig. 6.59.

6.8. Calculate the input impedance and the transfer function of each circuit in Fig. 6.60.

6.9. Calculate the gain of each circuit in Fig. 6.61 at very low and very high frequencies. Neglect all other capacitances and assume that $\lambda = 0$ for circuits (a) and (b) and $\gamma = 0$ for all of the circuits.

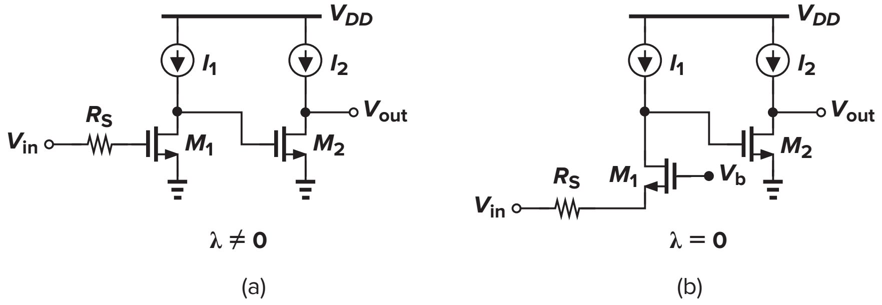


Figure 6.59

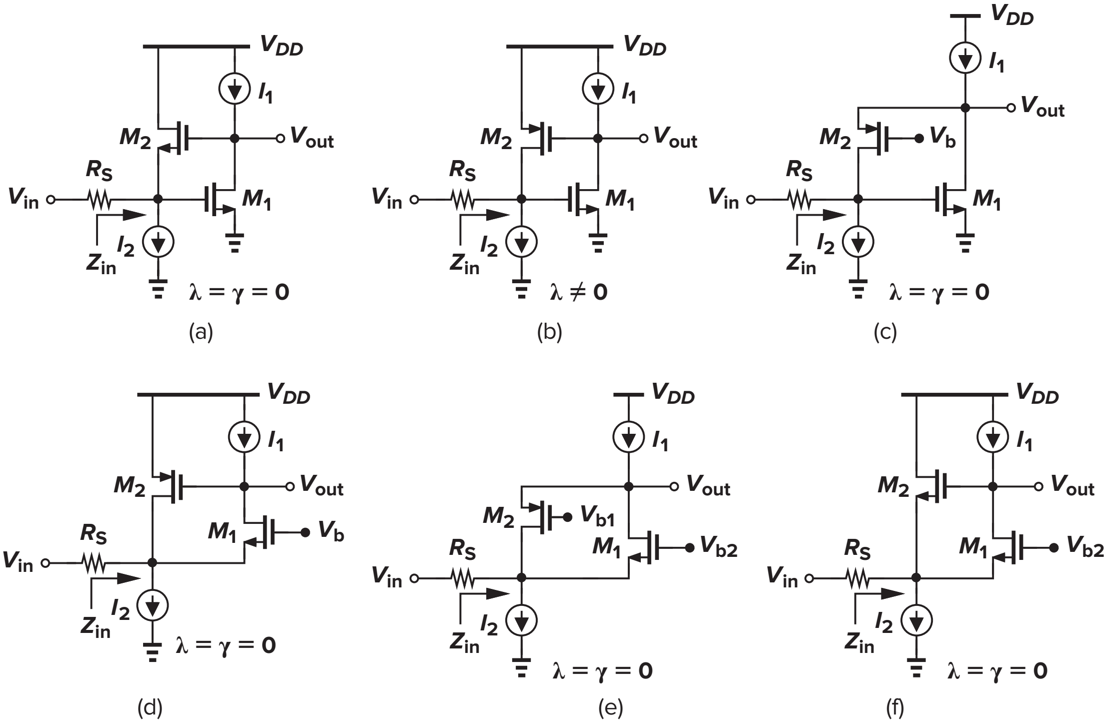


Figure 6.60

6.10. Calculate the gain of each circuit in Fig. 6.62 at very low and very high frequencies. Neglect all other capacitances and assume that $\lambda = \gamma = 0$.

6.11. Consider the cascode stage shown in Fig. 6.63. In our analysis of the frequency response of a cascode stage, we assumed that the gate-drain overlap capacitance of M_1 is multiplied by $g_{m1}/(g_{m2} + g_{mb2})$. Recall from Chapter 3, however, that with a high resistance loading the drain of M_2, the resistance seen looking into the source of M_2 can be quite high, suggesting a much higher Miller multiplication factor for C_{GD1}. Explain why C_{GD1} is still multiplied by $1 + g_{m1}/(g_{m2} + g_{mb2})$ if C_L is relatively large.

6.12. Neglecting other capacitances, calculate Z_X in the circuits of Fig. 6.64. Sketch $|Z_X|$ versus frequency.

6.13. The common-gate stage of Fig. 6.31 is designed with $(W/L)_1 = 50/0.5$, $I_{D1} = 1$ mA, $R_D = 2$ kΩ, and $R_S = 1$ kΩ. Assuming $\lambda = 0$, determine the poles and the low-frequency gain. How do these results compare with those obtained in Problem 6.9?

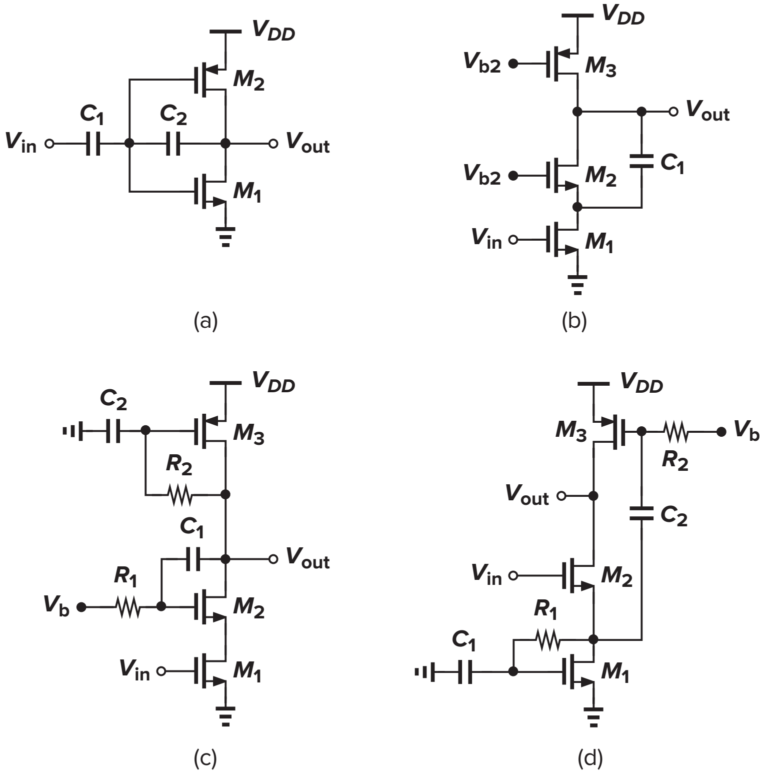


Figure 6.61

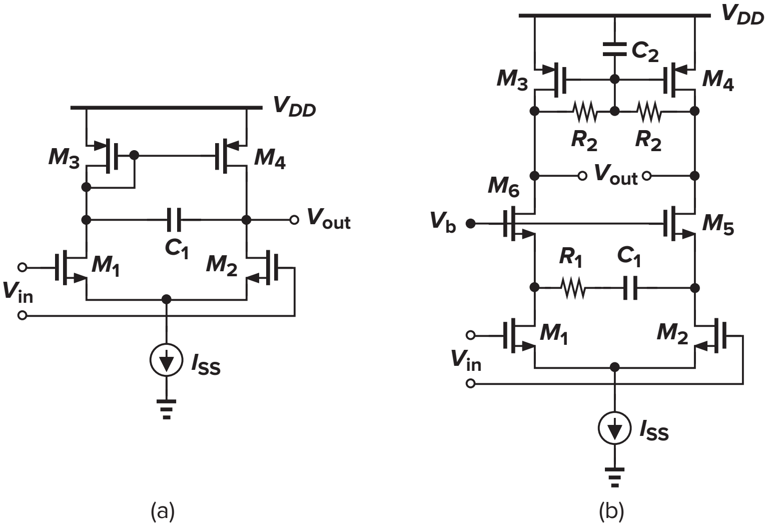


Figure 6.62

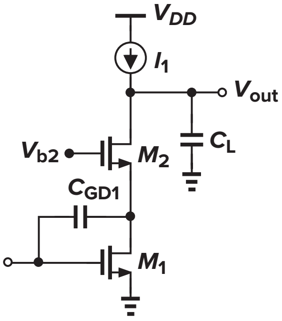


Figure 6.63

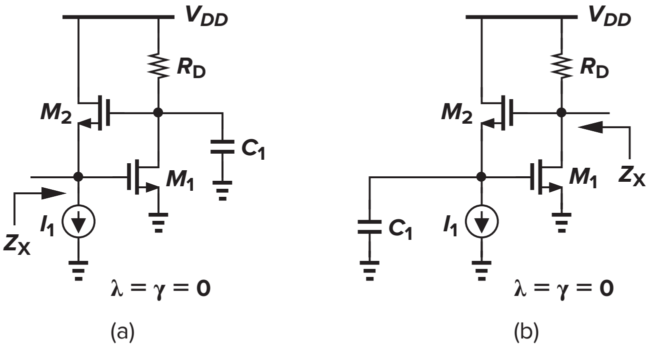


Figure 6.64

6.14. Suppose that in the cascode stage of Fig. 6.34, a resistor R_G appears in series with the gate of M_2. Including only C_{GS2}, neglecting other capacitances, and assuming $\lambda = \gamma = 0$, determine the transfer function.

6.15. Apply the method of Fig. 6.18 to the circuit of Fig. 6.41(b) to determine the zero of the transfer function.

6.16. The circuit of Fig. 6.42(a) is designed with $(W/L)_{1,2} = 50/0.5$ and $(W/L)_{3,4} = 10/0.5$. If $I_{SS} = 100\ \mu$A, $K = 2$, $C_L = 0$, and R_D is implemented by an NFET having $W/L = 50/0.5$, estimate the poles and zeros of the circuit. Assume the amplifier is driven by an ideal voltage source.

6.17. A differential pair driven by an ideal voltage source is required to have a total phase shift of 135° at the frequency where its gain drops to unity.

(a) Explain why a topology in which the load is realized by diode-connected devices or current sources does not satisfy this condition.

(b) Consider the circuit shown in Fig. 6.65. Neglecting other capacitances, determine the transfer function. Explain under what conditions the load exhibits an inductive behavior. Can this circuit provide a total phase shift of 135° at the frequency where its gain drops to unity?

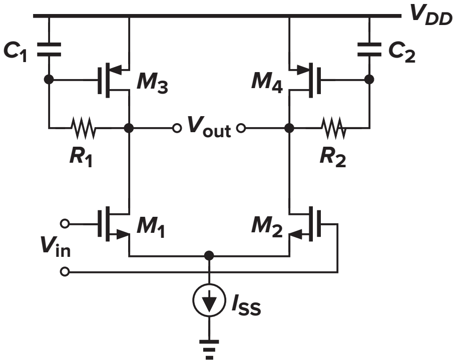


Figure 6.65

6.18. Repeat Example 6.3, but assume that I_1 is replaced with a resistor R_1.

6.19. A resistively-degenerated common-source stage bootstraps C_{GS} in a manner similar to a source follower. Estimate the input capacitance of such a stage.

6.20. Determine the transfer function of a CG stage with a resistance R_G in series with the gate, including only C_{GS} and C_{GD}. Assume $\lambda = \gamma = 0$.

6.21. Determine the transfer function of a CG stage with a resistance R_G in series with the gate, including only C_{GD} and C_{DB}. Assume $\lambda = \gamma = 0$.

6.22. Determine the transfer function of a differential pair with current-source loads for differential signals. Assume that each input is driven by a series resistance of R_S.

6.23. Consider a circuit containing only one capacitor, C_1. We set the main input to zero and apply a current source, I_X, in parallel with C_1, obtaining the voltage across it, V_X, and hence $V_X(s)/I_X(s)$ (Fig. 6.66). This impedance has the same pole as the main transfer function. Prove that the pole is given by $1/(R_1C_1)$, where R_1 is the resistance seen by C_1.

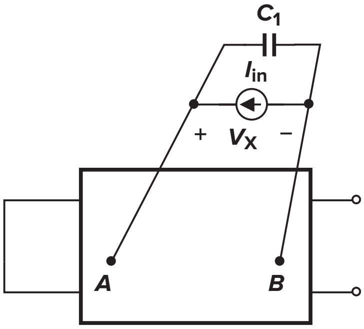


Figure 6.66

6.24. Repeat Example 6.22, but with $\lambda > 0$ and $\gamma > 0$.

6.25. Prove that the -3-dB bandwidth of N first-order identical gain stages is given by $\sqrt{\sqrt[N]{2} - 1}\omega_p$, where ω_p denotes the pole of one stage.

6.26. Prove that if $C_{GD} = 0$, then Eq. (6.30) reduces to the product of two transfer functions that can simply be obtained by association of poles with the input and output nodes.

CHAPTER

7 Noise

Noise limits the minimum signal level that a circuit can process with acceptable quality. Today's analog designers constantly deal with the problem of noise because it trades with power dissipation, speed, and linearity.

In this chapter, we describe the phenomenon of noise and its effect on analog circuits. The objective is to provide sufficient understanding of the problem so that further developments of analog circuits in the following chapters take noise into account as naturally as they do other circuit parameters, such as gain, input and output impedances, etc. Seemingly a complex subject, noise is introduced at this early stage so as to accompany the reader for the remainder of the book and become more intuitive through various examples.

Following a general description of noise characteristics in the frequency and time domains, we introduce thermal and flicker noise. Next, we consider methods of representing noise in circuits. Finally, we describe the effect of noise in single-stage and differential amplifiers along with trade-offs with other performance parameters.

7.1 ■ Statistical Characteristics of Noise

Noise is a random process. For our purposes in this book, this statement means that the value of noise cannot be predicted at any time even if the past values are known. Compare the output of a sine-wave generator with that of a microphone picking up the sound of water flow in a river (Fig. 7.1). While the value of $x_1(t)$ at $t = t_1$ can be predicted from the observed waveform, the value of $x_2(t)$ at $t = t_2$ cannot. This is the principal difference between deterministic and random phenomena.

If the instantaneous value of noise in the time domain cannot be predicted, how can we incorporate noise in circuit analysis? This is accomplished by observing the noise for a long time and using the measured results to construct a "statistical model" for the noise. While the instantaneous *amplitude* of noise cannot be predicted, a statistical model provides knowledge about some other important properties of the noise that prove useful and adequate in circuit analysis.

Which properties of noise *can* be predicted? In many cases, the average power of noise is predictable. For example, if a microphone picking up the sound of a river is brought closer to the river, the resulting electrical signal displays, on the average, larger excursions and hence higher power (Fig. 7.2). The reader may wonder if a random process can be so random that even its average power is unpredictable. Such processes do exist, but we are fortunate that most sources of noise in circuits exhibit a constant average power.

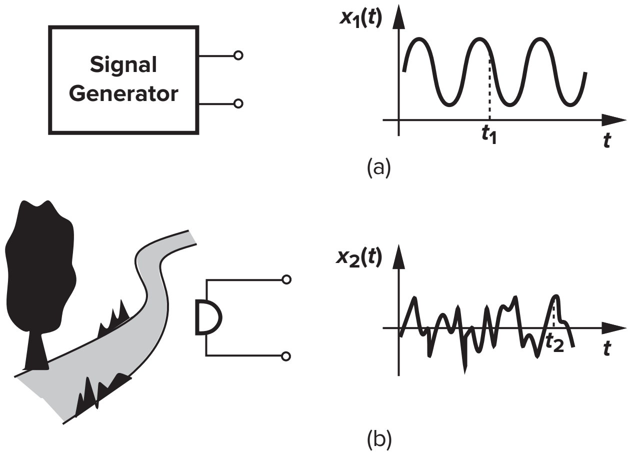


Figure 7.1 (a) The output of a generator, and (b) the sound of a river.

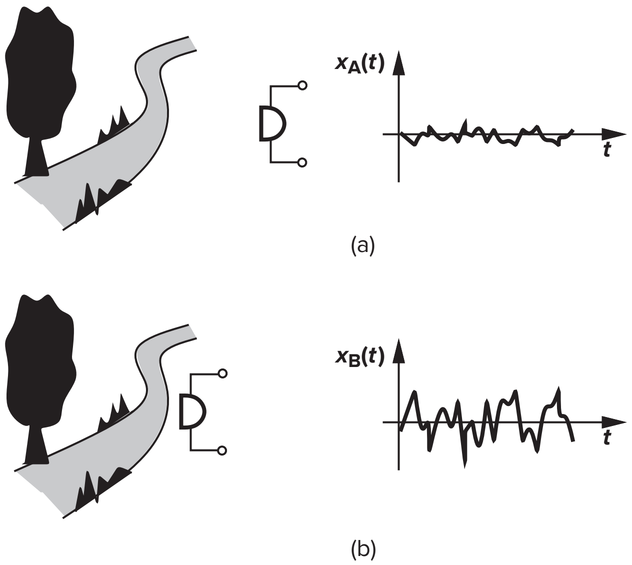


Figure 7.2 Illustration of the average power of a random signal.

The concept of average power proves essential in our analysis and must be defined carefully. Recall from basic circuit theory that the average power delivered by a periodic voltage $v(t)$ to a load resistance R_L is given by

$$P_{av} = \frac{1}{T}\int_{-T/2}^{+T/2} \frac{v^2(t)}{R_L}dt \tag{7.1}$$

where T denotes the period.[1] Measured in watts, this quantity can be visualized as the average heat produced in R_L by $v(t)$.

[1] To be more rigorous, $v^2(t)$ should be replaced by $v(t) \cdot v^*(t)$, where $v^*(t)$ is the complex conjugate waveform.

How do we define P_{av} for a random signal? In the example of Fig. 7.2, we expect that $x_B(t)$ generates more heat than $x_A(t)$ if the microphone drives a resistive load. However, since the signals are not periodic, the measurement must be carried out over a long time:

$$P_{av} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{+T/2} \frac{x^2(t)}{R_L} dt \tag{7.2}$$

where $x(t)$ is a voltage quantity. Figure 7.3 illustrates the operation on $x(t)$: the signal is squared, the area under the resulting waveform is calculated for a long time T, and the average power is obtained by normalizing the area to T.[2]

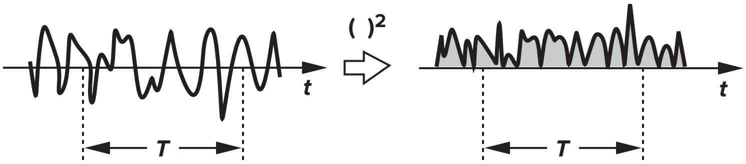


Figure 7.3 Average noise power.

To simplify calculations, we write the definition of P_{av} as

$$P_{av} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{+T/2} x^2(t) dt \tag{7.3}$$

where P_{av} is expressed in V^2 rather than W. The idea is that if we know P_{av} from (7.3), then the actual power delivered to a load R_L can be readily calculated as P_{av}/R_L. In analogy with deterministic signals, we can also define a root-mean-square (rms) voltage for noise as $\sqrt{P_{av}}$, where P_{av} is given by (7.3).

7.1.1 Noise Spectrum

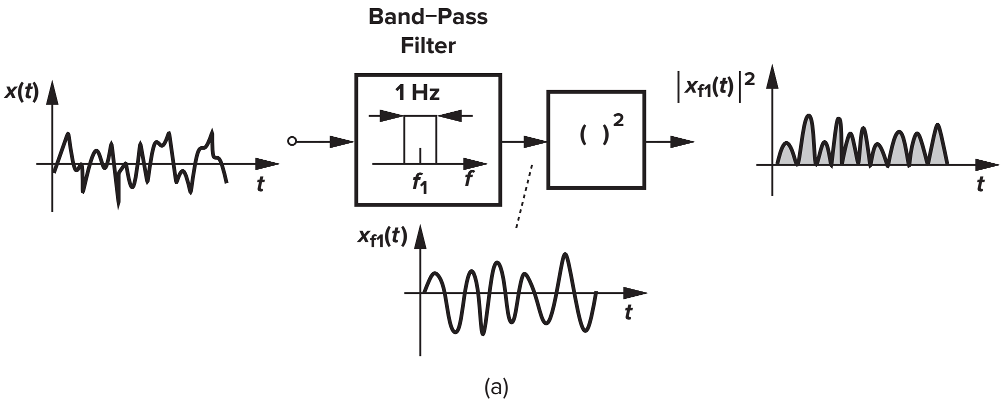
The concept of average power becomes more versatile if defined with regard to the *frequency content* of noise. The noise made by a group of men contains weaker high-frequency components than that made by a group of women, a difference observable from the "spectrum" of each type of noise. Also called the "power spectral density" (PSD), the spectrum shows how much power the signal carries at each frequency. More specifically, the PSD, $S_x(f)$, of a noise waveform $x(t)$ is defined as the average power carried by $x(t)$ in a one-hertz bandwidth around f. That is, as illustrated in Fig. 7.4(a), we apply $x(t)$ to a bandpass filter with a center frequency of f_1 and a 1-Hz bandwidth, square the output, $x_{f1}(t)$, and calculate the average over a long time to obtain $S_x(f_1)$. Repeating the procedure with bandpass filters having different center frequencies, we arrive at the overall shape of $S_x(f)$ [Fig. 7.4(b)].[3] Generally, $S_x(f)$ is measured in watts per hertz. The total area under $S_x(f)$ represents the power carried by the signal (or the noise) at all frequencies; i.e., the total power.

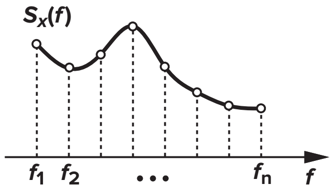
▶ **Example 7.1**

(a) Sketch the spectrum of voice for men and women. What does the difference imply about their time-domain waveforms?
(b) Estimate the averaging time, T, in Eq. (7.3) for voice signals.

[2] Strictly speaking, this definition holds only for "stationary" processes [1].

[3] In signal processing theory, the PSD is defined as the Fourier transform of the autocorrelation function of the noise. The two definitions are equivalent in most cases of interest to us.





(b)

Figure 7.4 Calculation of noise spectrum.

Solution

(a) The human voice exhibits frequencies from 20 Hz to 20 kHz. Since women's voice contain stronger high-frequency components, we expect the two spectra to differ as shown in Fig. 7.5(a). In the time domain, we observe faster changes in women's voice [Fig. 7.5(b)].

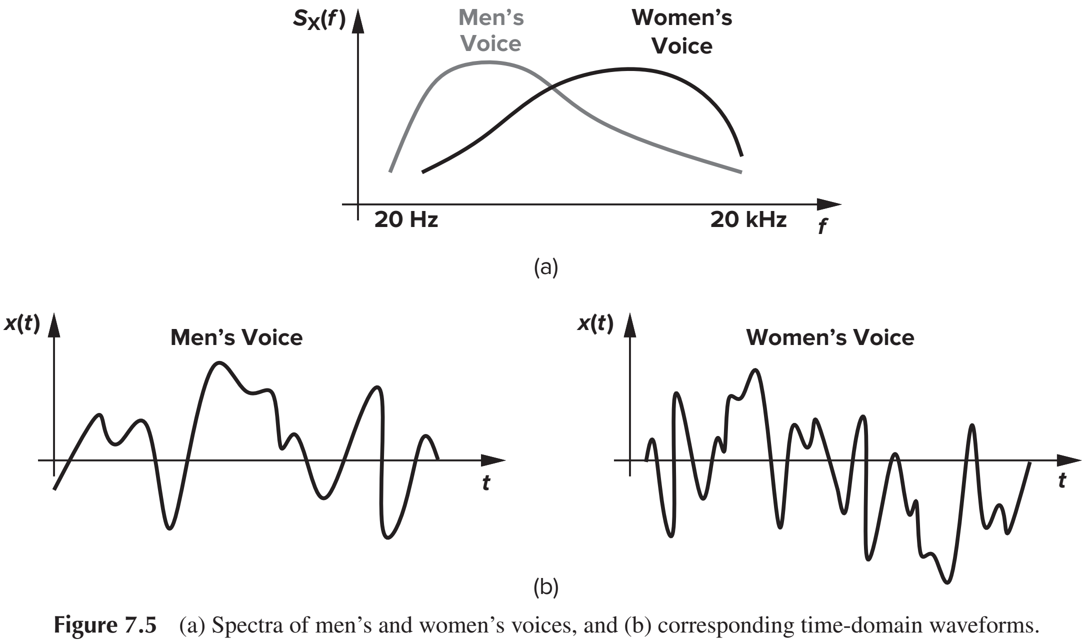


Figure 7.5 (a) Spectra of men's and women's voices, and (b) corresponding time-domain waveforms.

(b) The averaging time must be long enough to include a sufficient number of cycles of the *lowest* frequencies. That is, the averaging must capture the slowest dynamics in the signal. We must therefore choose T to be at least about 10 cycles of 20 Hz, i.e., roughly 500 ms.

As with the definition of P_{av} in (7.3), it is customary to eliminate R_L from $S_x(f)$. Thus, since each value on the plot in Fig. 7.4(b) is measured for a 1-Hz bandwidth, $S_x(f)$ is expressed in V^2/Hz rather than W/Hz. It is also common to take the square root of $S_x(f)$, expressing the result in V/$\sqrt{\text{Hz}}$. For example, we say that the input noise voltage of an amplifier at 100 MHz is equal to 3 nV/$\sqrt{\text{Hz}}$, simply to mean that the average power in a 1-Hz bandwidth at 100 MHz is equal to $(3 \times 10^{-9})^2$ V^2.

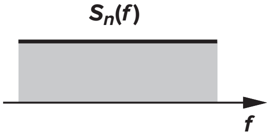


Figure 7.6 White spectrum.

An example of a common type of noise PSD is the "white spectrum," also called white noise. Shown in Fig. 7.6, such a PSD displays the same value at all frequencies (similar to white light). Strictly speaking, we note that white noise does not exist because the total area under the power spectral density, i.e., the total power carried by the noise, is infinite. In practice, however, any noise spectrum that is flat *in the band of interest* is usually called white.

The PSD is a powerful tool in analyzing the effect of noise in circuits, especially in conjunction with the following theorem.

Theorem If a signal with spectrum $S_x(f)$ is applied to a linear time-invariant system with transfer function $H(s)$, then the output spectrum is given by

$$S_Y(f) = S_x(f)|H(f)|^2 \tag{7.4}$$

where $H(f) = H(s = 2\pi jf)$. The proof can be found in textbooks on signal processing or communications, e.g., [1].

Somewhat similar to the relation $Y(s) = X(s)H(s)$, this theorem agrees with our intuition that the spectrum of the signal should be "shaped" by the transfer function of the system (Fig. 7.7). For example, as illustrated in Fig. 7.8, since regular telephones have a bandwidth of approximately 4 kHz, they suppress the high-frequency components of the caller's voice. Note that, owing to its limited bandwidth, $x_{out}(t)$ exhibits slower changes than does $x_{in}(t)$. This bandwidth limitation sometimes makes it difficult to recognize the caller's voice.

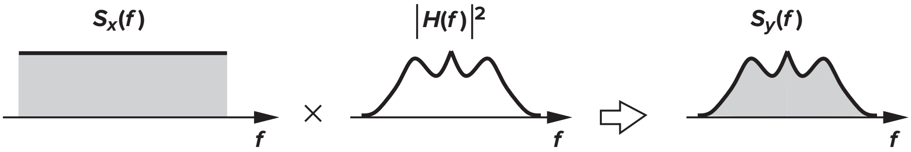


Figure 7.7 Noise shaping by a transfer function.

Since $S_x(f)$ is an even function of f for real $x(t)$ [1], as depicted in Fig. 7.9, the total power carried by $x(t)$ in the frequency range $[f_1\ f_2]$ is equal to

$$P_{f1,f2} = \int_{-f_2}^{-f_1} S_x(f)df + \int_{+f_1}^{+f_2} S_x(f)df \tag{7.5}$$

$$= \int_{+f_1}^{+f_2} 2S_x(f)df \tag{7.6}$$

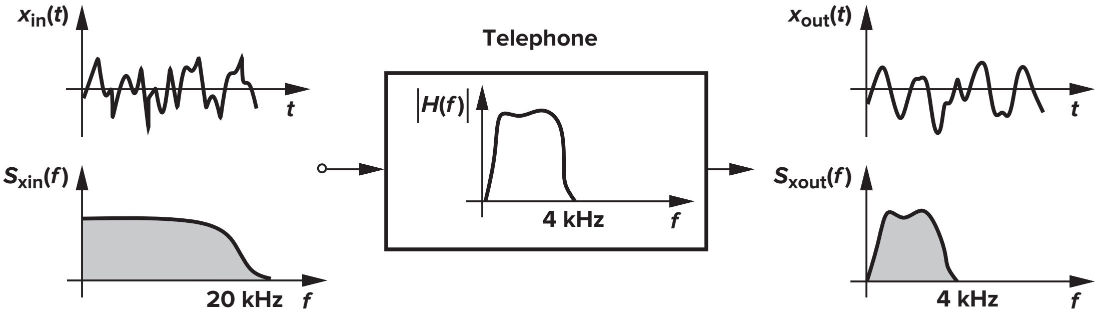


Figure 7.8 Spectral shaping by telephone bandwidth.

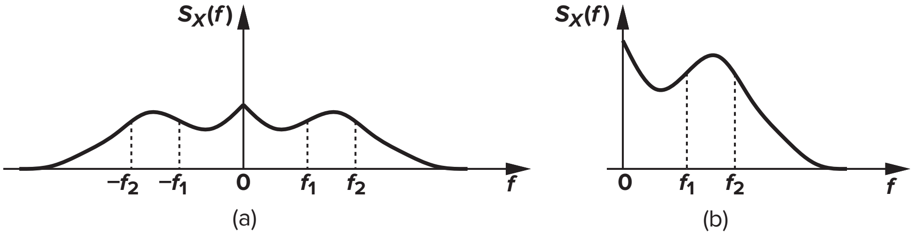


Figure 7.9 (a) Two-sided and (b) one-sided noise spectra.

In fact, the integral in (7.6) is the quantity measured by a power meter sensing the output of a bandpass filter between f_1 and f_2. That is, the negative-frequency part of the spectrum is folded around the vertical axis and added to the positive-frequency part. We call the representation of Fig. 7.9(a) the "two-sided" spectrum and that of Fig. 7.9(b) the "one-sided" spectrum. For example, the two-sided white spectrum of Fig. 7.6 has the one-sided counterpart shown in Fig. 7.10.

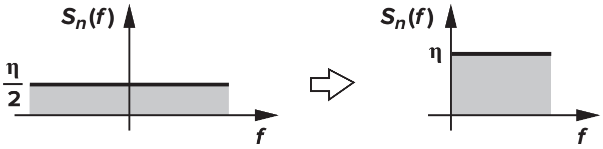


Figure 7.10 Folded white spectrum.

In summary, the spectrum shows the power carried in a small bandwidth at each frequency, revealing how *fast* the waveform is expected to vary in the time domain.

7.1.2 Amplitude Distribution

As mentioned earlier, the instantaneous amplitude of noise is usually unpredictable. However, by observing the noise waveform for a long time, we can construct a "distribution" of the amplitude, indicating how *often* each value occurs. Also called the "probability density function" (PDF), the distribution of $x(t)$ is defined as

$$p_X(x)dx = \text{probability of } x < X < x + dx \tag{7.7}$$

where X is the measured value of $x(t)$ at some point in time.

As illustrated in Fig. 7.11, to estimate the distribution, we sample $x(t)$ at many points, construct bins of small width, choose the bin height equal to the number of samples whose value falls between the two edges of the bin, and normalize the bin heights to the total number of samples. Note that the PDF provides no information as to how fast $x(t)$ varies in the time domain. For example, the sound generated by a violin may have the same amplitude distribution as that produced by a drum even though their frequency contents are vastly different.

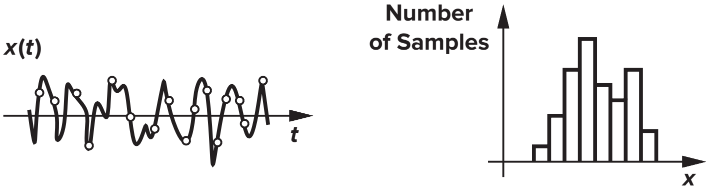


Figure 7.11 Amplitude distribution of noise.

An important example of PDFs is the Gaussian (or normal) distribution. The central limit theorem states that if many independent random processes with arbitrary PDFs are added, the PDF of the sum approaches a Gaussian distribution [1]. It is therefore not surprising that many natural phenomena exhibit Gaussian statistics. For example, since the noise of a resistor results from the random "walk" of a very large number of electrons, each having relatively independent statistics, the overall amplitude follows a Gaussian PDF.

In this book, we employ the spectrum and average power of noise to a much greater extent than the amplitude distribution. For completeness, however, we note that the Gaussian PDF is defined as

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\frac{-(x-m)^2}{2\sigma^2} \tag{7.8}$$

where σ and m are the standard deviation and mean of the distribution, respectively. For Gaussian distribution, σ is equal to the rms value of the noise.

7.1.3 Correlated and Uncorrelated Sources

In analyzing circuits, we often need to add the effect of several sources of noise to obtain the total noise. While for deterministic voltages and currents, we simply use the superposition principle, the procedure is somewhat different for random noise because we are ultimately interested in the average noise *power*. Let us add two noise waveforms and take the average of the resulting power:

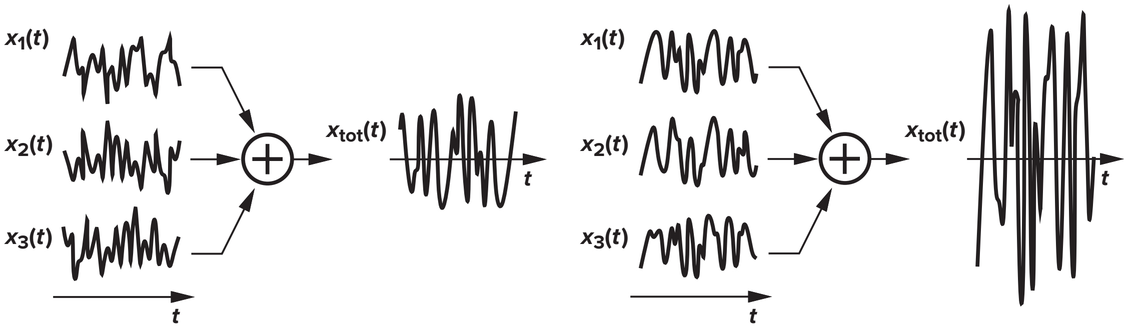
$$P_{av} = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{+T/2}[x_1(t)+x_2(t)]^2 dt \tag{7.9}$$

$$\begin{aligned} &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{+T/2}x_1^2(t)dt + \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{+T/2}x_2^2(t)dt \\ &\quad + \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{+T/2}2x_1(t)x_2(t)dt \end{aligned} \tag{7.10}$$

$$= P_{av1} + P_{av2} + \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{+T/2}2x_1(t)x_2(t)dt \tag{7.11}$$

where P_{av1} and P_{av2} denote the average power of $x_1(t)$ and $x_2(t)$, respectively. Called the "correlation" between $x_1(t)$ and $x_2(t)$,[4] the third term in (7.11) indicates how "similar" these two waveforms are. If generated by independent devices, the noise waveforms are usually "uncorrelated" and the integral in (7.11) vanishes. For example, the noise produced by a resistor has no correlation with that generated by a transistor. In such a case, $P_{av} = P_{av1} + P_{av2}$. From this observation, we say that superposition holds for the *power* of uncorrelated noise sources. Of course, superposition also holds for noise voltages and currents, but this does not help us in most cases.

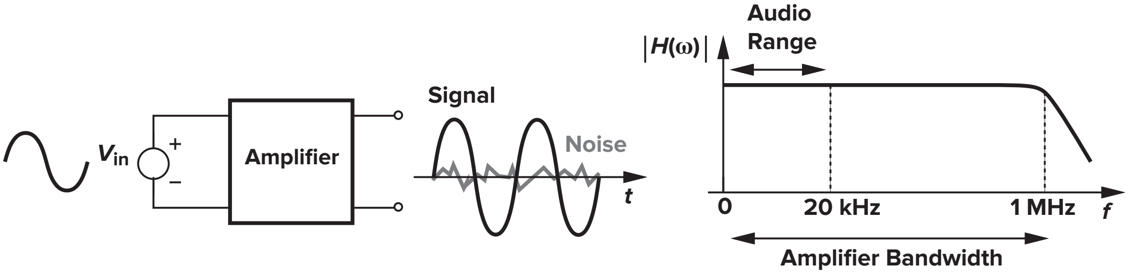
A familiar analogy is that of the spectators in a sports stadium. Before the game begins, many conversations are in progress, generating uncorrelated noise components [Fig. 7.12(a)]. During the game, the spectators applaud (or scream) simultaneously, producing correlated noise at a much higher power level arising from the third term in Eq. (7.11). [Fig. 7.12(b)].



(a) (b)

Figure 7.12 (a) Uncorrelated noise and (b) correlated noise generated in a stadium.

In most cases studied in this book, the noise sources are uncorrelated. One exception is studied in Section 7.3.



(a) (b)

Figure 7.13 (a) Output noise produced by a circuit, and (b) additional noise if bandwidth is excessively wide.

7.1.4 Signal-to-Noise Ratio

Suppose an amplifier receives a sinusoidal signal as shown in Fig. 7.13. The output contains both the amplified signal and the noise generated by the circuit. For the output signal to be intelligible, its power, P_{sig}, must be sufficiently higher than that of the noise, P_{noise}. We therefore define the "signal-to-noise ratio" (SNR) as

$$\mathrm{SNR} = \frac{P_{sig}}{P_{noise}} \tag{7.12}$$

[4] This terminology applies only to stationary signals.

For example, audio signals require a minimum SNR of about 20 dB (i.e., $P_{sig}/P_{noise} = 100$).[5] For a sinusoid having a peak amplitude of A, $P_{sig} = A^2/2$, but how do we calculate P_{noise}? The total average power carried by noise is equal to the area under its spectrum:

$$P_{noise} = \int_{-\infty}^{+\infty} S_{noise}(f)df \tag{7.13}$$

Does this mean that P_{noise} can be very large if $S_{noise}(f)$ spans a wide frequency range? Yes, indeed. As an example, suppose the above amplifier provides a bandwidth of 1 MHz while sensing an audio signal [Fig. 7.13(b)]. Then, the signal is corrupted by all of the noise components in the 1-MHz bandwidth. For this reason, the bandwidth of the circuit must always be limited to the minimum acceptable value so as to minimize the integrated noise power. The bandwidth can be reduced within the amplifier or by a low-pass filter placed thereafter.

▶ **Example 7.2**

An amplifier produces a one-sided noise spectrum given by $S_{noise}(f) = 5 \times 10^{-16}$ V^2/Hz. Determine the total output noise in a bandwidth of 1 MHz.

Solution

We have

$$P_{noise} = \int_{0}^{1\text{ MHz}} S_{noise}(f)df \tag{7.14}$$

$$= 5 \times 10^{-10}\text{ V}^2 \tag{7.15}$$

Note that the total integrated noise is measured in V^2 and not in V^2/Hz. This noise power corresponds to an rms voltage of $\sqrt{5 \times 10^{-10}\text{ V}^2} = 22.4\ \mu$V. ◀

7.1.5 Noise Analysis Procedure

With the tools developed in previous sections, we can now outline a methodology for the analysis of noise in circuits. The output signal of a given circuit is corrupted by the noise sources within the circuit. We are therefore interested in the noise observed at the output. Our procedure consists of four steps:

1. Identify the sources of noise (e.g., resistors and transistors) and write down the spectrum of each.
2. Find the transfer function from each noise source to the output (as if the source were a deterministic signal).
3. Utilize the theorem $S_Y(f) = S_x(f)|H(f)|^2$ to calculate the output noise spectrum contributed by each noise source. (The input signal is set to zero.)
4. Add all of the output spectra, paying attention to correlated and uncorrelated sources.

This procedure gives the output noise spectrum, which must then be integrated from $-\infty$ to $+\infty$ so as to yield the total output noise. To carry out the first step, we need the noise representation of various electronic devices, to be described in the next section.

[5] Since P_{sig} and P_{noise} are power quantities, 20 dB $= 10\log(P_{sig}/P_{noise})$.

7.2 ■ Types of Noise

Analog signals processed by integrated circuits are corrupted by two different types of noise: device electronic noise and "environmental" noise. The latter refers to (seemingly) random disturbances that a circuit experiences through the supply or ground lines or through the substrate. We focus on device electronic noise here and defer the study of environmental noise to Chapter 19.

7.2.1 Thermal Noise

Resistor Thermal Noise The random motion of electrons in a conductor introduces fluctuations in the voltage measured across the conductor, even if the average current is zero. Thus, the spectrum of thermal noise is proportional to the absolute temperature. As shown in Fig. 7.14, the thermal noise of a resistor R can be modeled by a series voltage source, with the one-sided spectral density

$$S_v(f) = 4kTR, \quad f \geq 0 \tag{7.16}$$

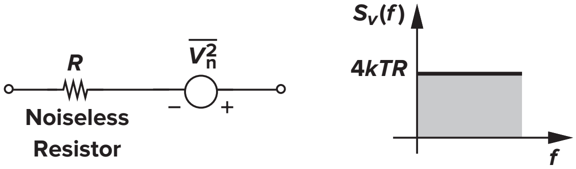


Figure 7.14 Thermal noise of a resistor.

where $k = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant. Note that $S_v(f)$ is expressed in V^2/Hz. Thus, we also write $\overline{V_n^2} = 4kTR$, where the overline indicates averaging.[6] We may even say that the noise "voltage" is given by $4kTR$ even though this quantity is in fact the noise voltage squared. For example, a 50-Ω resistor held at $T = 300$ K exhibits 8.28×10^{-19} V^2/Hz of thermal noise. To convert this number to a more familiar voltage quantity, we take the square root, obtaining 0.91 nV/$\sqrt{\text{Hz}}$. While the square root of hertz may appear strange, it is helpful to remember that 0.91 nV/$\sqrt{\text{Hz}}$ has little significance per se and simply means that the power in a 1-Hz bandwidth is equal to $(0.91 \times 10^{-9})^2$ V^2.

The equation $S_v(f) = 4kTR$ suggests that thermal noise is white. In reality, $S_v(f)$ is flat for up to roughly 100 THz, dropping at higher frequencies. For our purposes, the white spectrum is quite accurate.

Since noise is a random quantity, the polarity used for the voltage source in Fig. 7.14 is unimportant. Nevertheless, once a polarity is chosen, it must be retained throughout the analysis of the circuit so as to obtain consistent results.

▶ **Example 7.3**

Consider the RC circuit shown in Fig. 7.15. Calculate the noise spectrum and the total noise power in V_{out}.

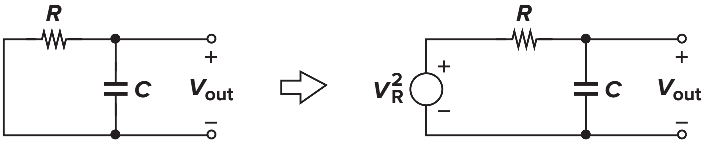


Figure 7.15 Noise generated in a low-pass filter.

[6] Some books write $\overline{V_n^2} = 4kTR\Delta f$ to emphasize that $4kTR$ is the noise power per unit bandwidth. To simplify the notation, we assume that $\Delta f = 1$ Hz, unless otherwise stated. In other words, we use $S_v(f)$ and $\overline{V_n^2}$ interchangeably.

Solution

We follow the four steps described in Section 7.1.5. The noise spectrum of R is given by $S_v(f) = 4kTR$. Next, modeling the noise of R by a series voltage source V_R, we compute the transfer function from V_R to V_{out}:

$$\frac{V_{out}}{V_R}(s) = \frac{1}{RCs+1} \tag{7.17}$$

From the theorem in Section 7.1.1, we have

$$S_{out}(f) = S_v(f)\left|\frac{V_{out}}{V_R}(j\omega)\right|^2 \tag{7.18}$$

$$= 4kTR\frac{1}{4\pi^2R^2C^2f^2+1} \tag{7.19}$$

Thus, the white noise spectrum of the resistor is shaped by a low-pass characteristic (Fig. 7.16). To calculate the total noise power at the output, we write

$$P_{n,out} = \int_0^\infty \frac{4kTR}{4\pi^2R^2C^2f^2+1}df \tag{7.20}$$

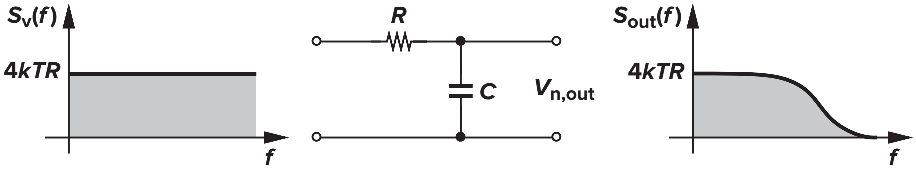


Figure 7.16 Noise spectrum shaping by a low-pass filter.

Note that the integration must be with respect to f rather than ω (why?). Since

$$\int \frac{dx}{x^2+1} = \tan^{-1}x \tag{7.21}$$

the integral reduces to

$$P_{n,out} = \frac{2kT}{\pi C}\tan^{-1}u|_{u=0}^{u=\infty} \tag{7.22}$$

$$= \frac{kT}{C} \tag{7.23}$$

Note that the unit of kT/C is V^2. We may also consider $\sqrt{kT/C}$ as the total rms noise voltage measured at the output. For example, with a 1-pF capacitor, the total noise voltage is equal to 64.3 μV_{rms} at $T = 300$ K.

Equation (7.23) implies that the total noise at the output of the circuit shown in Fig. 7.15 is independent of the value of R. Intuitively, this is because for larger values of R, the associated noise per unit bandwidth increases while the overall bandwidth of the circuit decreases. The fact that kT/C noise can be decreased only by increasing C (if T is fixed) introduces many difficulties in the design of analog circuits (Chapter 13).

The thermal noise of a resistor can be represented by a parallel current source as well (Fig. 7.17). For the representations of Figs. 7.14 and 7.17 to be equivalent, we have $\overline{V_n^2}/R^2 = \overline{I_n^2}$, that is, $\overline{I_n^2} = 4kT/R$. Note that $\overline{I_n^2}$ is expressed in A^2/Hz. Depending on the circuit topology, one model may lead to simpler calculations than the other.

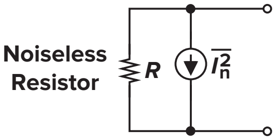


Figure 7.17 Representation of resistor thermal noise by a current source.

▶ **Example 7.4**

Calculate the equivalent noise voltage of two parallel resistors R_1 and R_2 [Fig. 7.18(a)].

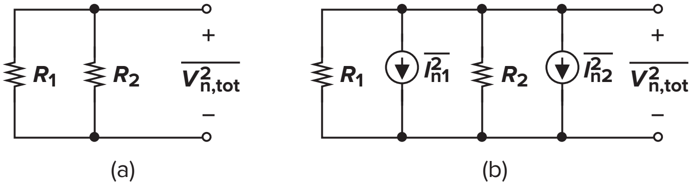


Figure 7.18

Solution

As shown in Fig. 7.18(b), each resistor exhibits an equivalent noise current with the spectral density $4kT/R$. Since the two noise sources are uncorrelated, we add the *powers*:

$$\overline{I_{n,tot}^2} = \overline{I_{n1}^2} + \overline{I_{n2}^2} \tag{7.24}$$

$$= 4kT\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{7.25}$$

Thus, the equivalent noise voltage is given by

$$\overline{V_{n,tot}^2} = \overline{I_{n,tot}^2}(R_1\|R_2)^2 \tag{7.26}$$

$$= 4kT(R_1\|R_2) \tag{7.27}$$

as intuitively expected. Note that our notation assumes a 1-Hz bandwidth.

◀

The dependence of thermal noise (and some other types of noise) upon T suggests that low-temperature operation can decrease the noise in analog circuits. This approach becomes more attractive with the observation that the mobility of charge carriers in MOS devices increases at low temperatures [2].[7] Nonetheless, the required cooling equipment limits the practicality of low-temperature circuits.

MOSFETs MOS transistors also exhibit thermal noise. The most significant source is the noise generated in the channel. It can be proved [4] that for long-channel MOS devices operating in saturation, the channel noise can be modeled by a current source connected between the drain and source terminals (Fig. 7.19) with a spectral density:[8]

$$\overline{I_n^2} = 4kT\gamma g_m \tag{7.28}$$

[7] At extremely low temperatures, the mobility drops due to "carrier freezeout" [2].

[8] The actual equation reads $\overline{I_n^2} = 4kT\gamma g_{ds}$, where g_{ds} is the drain-source conductance with $V_{DS} = 0$, i.e., the same as R_{on}^{-1}. For long-channel devices, g_{ds} with $V_{DS} = 0$ is equal to g_m in saturation.

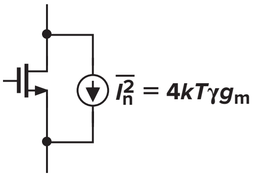


Figure 7.19 Thermal noise of a MOSFET.

The coefficient γ (not to be confused with the body effect coefficient!) is derived to be equal to 2/3 for long-channel transistors and may need to be replaced by a larger value for submicron MOSFETs [5]. It also varies to some extent with the drain-source voltage. As a rule of thumb, we assume $\gamma \approx 1$.

▶ Example 7.5

Find the maximum noise voltage that a single MOSFET can generate.

Solution

As shown in Fig. 7.20, the maximum output noise occurs if the transistor sees only its own output impedance as the load, i.e., if the external load is an ideal current source. The output noise voltage spectrum is then given by $S_{out}(f) = S_{in}(f)|H(f)|^2$, i.e.,

$$\overline{V_n^2} = \overline{I_n^2} r_O^2 \tag{7.29}$$

$$= (4kT\gamma g_m) r_O^2 \tag{7.30}$$

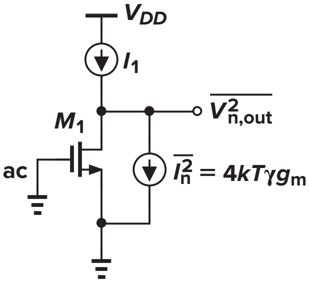


Figure 7.20

Let us make three observations. First, (7.30) suggests that the noise *current* of a MOS transistor decreases if the transconductance drops. For example, if the transistor operates as a constant current source, it is desirable to *minimize* its transconductance.

Second, the noise measured at the output of the circuit does not depend on where the input terminal is because for output noise calculation, the input is set to zero.[9] For example, the circuit of Fig. 7.20 may be a common-source or a common-gate stage, exhibiting the same output noise.

Third, the output resistance, r_O, does not produce noise because it is not a physical resistor. ◀

The ohmic sections of a MOSFET also contribute thermal noise. As conceptually illustrated in the top view of Fig. 7.21(a), the gate, source, and drain materials exhibit finite resistivity, thereby introducing noise. For a relatively wide transistor, the source and drain resistance is typically negligible whereas the gate distributed resistance may become noticeable.

[9] Of course, if the input voltage or current source has an output impedance that generates noise, this statement must be interpreted carefully.

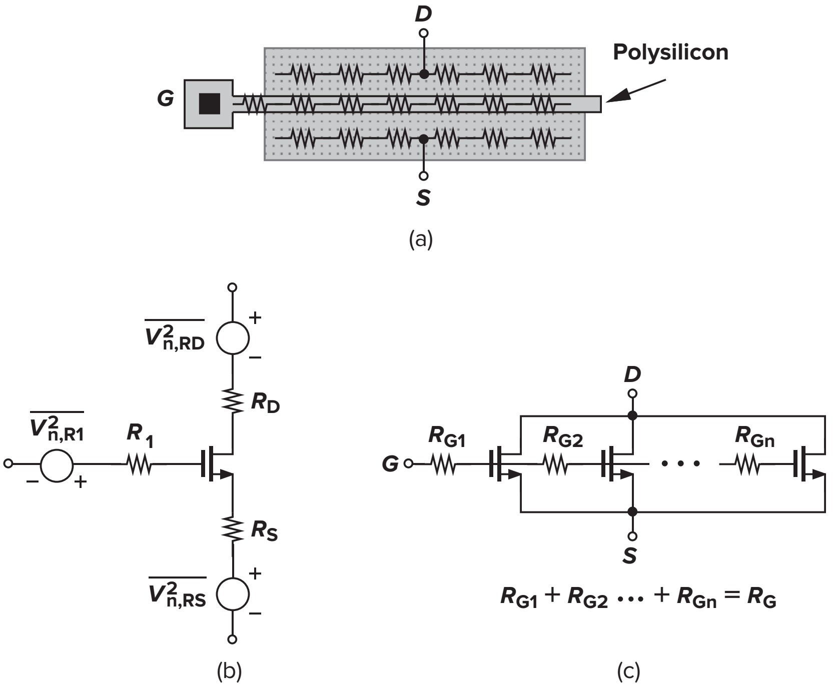
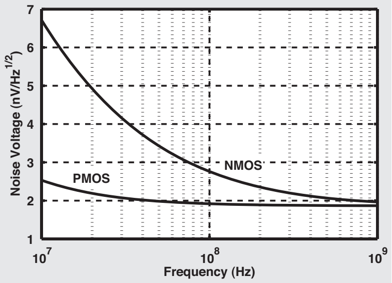


Figure 7.21 (a) Layout of a MOSFET indicating the terminal resistances; (b) circuit model; (c) distributed gate resistance.

Nanometer Design Notes

The small dimensions of nanometer devices lead to considerable flicker noise. Plotted below are the gate-referred noise spectra for PMOS and NMOS devices with $W/L = 5\ \mu$m/40 nm and $I_D = 250\ \mu$A. We observe that PMOS devices exhibit less noise, and the NMOS flicker noise corner is as high as several hundred megahertz. For low flicker noise, therefore, the transistor areas must be increased substantially.



In the noise model of Fig. 7.21(b), a lumped resistor R_1 represents the distributed gate resistance. Viewing the overall transistor as the distributed structure shown in Fig. 7.21(c), we observe that the unit transistors near the left end see the noise of only a fraction of R_G whereas those near the right end see the noise of most of R_G. We therefore expect the lumped resistor in the noise model to be *less* than R_G. In fact, it can be proved that $R_1 = R_G/3$ (Problem 7.3) [3], and hence the noise generated by the gate resistance is given by $\overline{V^2_{nRG}} = 4kTR_G/3$.

While the thermal noise generated in the channel is controlled by only the transconductance of the device, the effect of R_G can be reduced by proper layout. Shown in Fig. 7.22 are two examples. In Fig. 7.22(a), the two ends of the gate are shorted by a metal line, thus reducing the distributed resistance from R_G to $R_G/4$ (why?). Alternatively, the transistor can be folded as described in Chapter 19 [Fig. 7.22(b)] so that each gate "finger" exhibits a resistance of $R_G/2$, yielding a total distributed resistance of $R_G/4$ for the composite transistor.

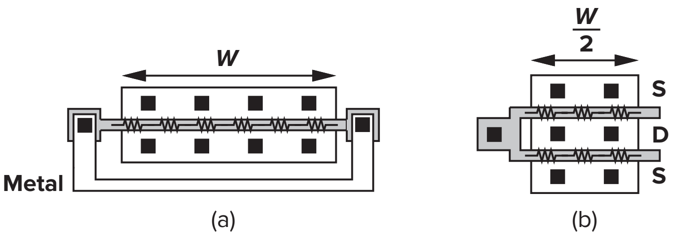


Figure 7.22 Reduction of gate resistance by (a) adding contacts to both sides or (b) folding.

▶ Example 7.6

A transistor of width W is laid out with one gate finger and exhibits a total gate resistance of R_G [Fig. 7.23(a)]. Now, we reconfigure the device into four equal gate fingers [Fig. 7.23(b)]. Determine the total gate resistance thermal noise spectrum of the new structure.

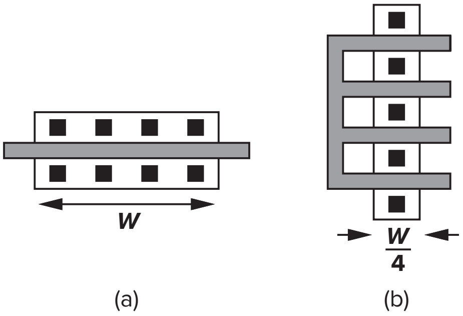


Figure 7.23

Solution

With a width of $W/4$, each gate finger now has a distributed resistance of $R_G/4$ and hence a lumped-model resistance of $R_G/12$. Since the four fingers are in parallel, the net resistance is given by $R_G/48$, yielding a noise spectrum of

$$\overline{V_{nRG}^2} = 4kT\frac{R_G}{48} \tag{7.31}$$

(In general, if the gate is decomposed into N parallel fingers, the distributed resistance falls by a factor of N^2.)

▶ Example 7.7

Find the maximum thermal noise voltage that the gate resistance of a single MOSFET can generate. Neglect the device capacitances.

Solution

If the total distributed gate resistance is R_G, then from Fig. 7.24, the output noise voltage due to R_G is given by

$$\overline{V_{n,out}^2} = 4kT\frac{R_G}{3}(g_m r_O)^2 \tag{7.32}$$

An important observation here is that, for the gate resistance noise to be negligible, we must ensure that (7.32) is much less than (7.30), and thus

$$\frac{R_G}{3} \ll \frac{\gamma}{g_m} \tag{7.33}$$

The number of gate fingers is chosen large enough to guarantee this condition.

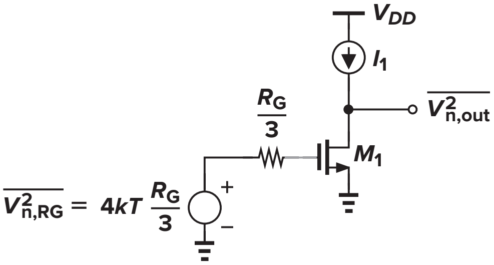


Figure 7.24

7.2.2 Flicker Noise

The interface between the gate oxide and the silicon substrate in a MOSFET entails an interesting phenomenon. Since the silicon crystal reaches an end at this interface, many "dangling" bonds appear, giving rise to extra energy states (Fig. 7.25). As charge carriers move at the interface, some are randomly trapped and later released by such energy states, introducing "flicker" noise in the drain current. In addition to trapping, several other mechanisms are believed to generate flicker noise [4].

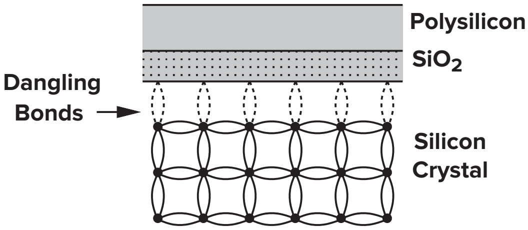


Figure 7.25 Dangling bonds at the oxide-silicon interface.

Unlike thermal noise, the average power of flicker noise cannot be predicted easily. Depending on the "cleanness" of the oxide-silicon interface, flicker noise may assume considerably different values and as such varies from one CMOS technology to another. The flicker noise is more easily modeled as a voltage source in series with the gate and, in the saturation region, roughly given by

$$\overline{V_n^2} = \frac{K}{C_{ox}WL} \cdot \frac{1}{f} \tag{7.34}$$

where K is a process-dependent constant on the order of 10^{-25} $\mathrm{V^2F}$. Note that our notation assumes a bandwidth of 1 Hz. Interestingly, as shown in Fig. 7.26, the noise spectral density is inversely proportional to the frequency. For example, the trap-and-release phenomenon associated with the dangling bonds occurs more often at low frequencies. For this reason, flicker noise is also called $1/f$ noise. Note that (7.34) does not depend on the bias current or the temperature. This is only an approximation; in reality, the flicker noise equation is somewhat more complex [3].

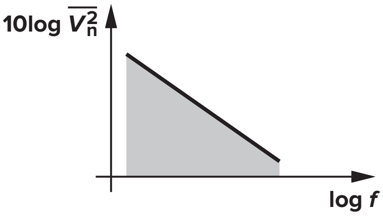


Figure 7.26 Flicker noise spectrum.

The inverse dependence of (7.34) on WL suggests that to reduce $1/f$ noise, the device *area* must be increased. It is therefore not surprising to see devices having areas of several hundred square microns in low-noise applications. (More fundamentally, the noise power trades with the gate capacitance, WLC_{ox}.) Generally, PMOS devices exhibit less $1/f$ noise than NMOS transistors because the former carry the holes in a "buried channel," i.e., at some distance from the oxide-silicon interface, and hence trap and release the carriers to a lesser extent.

▶ **Example 7.8**

For an NMOS current source, calculate the total thermal and $1/f$ noise in the drain current for a band from 1 kHz to 1 MHz.

Solution

The thermal noise current per unit bandwidth is given by $\overline{I_{n,th}^2} = 4kT\gamma g_m$. Thus, the total thermal noise integrated across the band of interest is

$$\overline{I_{n,th,tot}^2} = 4kT\gamma g_m(10^6 - 10^3) \tag{7.35}$$

$$\approx 4kT\gamma g_m \times 10^6 \text{ A}^2 \tag{7.36}$$

For $1/f$ noise, the drain noise current per unit bandwidth is obtained by multiplying the noise voltage at the gate by the device transconductance:

$$\overline{I_{n,1/f}^2} = \frac{K}{C_{ox}WL} \cdot \frac{1}{f} \cdot g_m^2 \tag{7.37}$$

The total $1/f$ noise is then equal to

$$\overline{I_{n,1/f,tot}^2} = \frac{Kg_m^2}{C_{ox}WL} \int_{1\text{ kHz}}^{1\text{ MHz}} \frac{df}{f} \tag{7.38}$$

$$= \frac{Kg_m^2}{C_{ox}WL} \ln 10^3 \tag{7.39}$$

$$= \frac{6.91Kg_m^2}{C_{ox}WL} \tag{7.40}$$

The above example raises an interesting question. What happens to $\overline{I_{n,1/f,tot}^2}$ if the lower end of the band, f_L, is zero rather than 1 kHz? Equation (7.39) then yields an infinite value for the total noise. To overcome the fear of infinite noise, we make two observations. First, extending f_L to zero means that we are interested in *arbitrarily* slow noise components. A noise component at 0.01 Hz varies significantly in roughly 10 s (one-tenth of the period) and one at 10^{-6} Hz in roughly one day. Second, the infinite flicker noise power simply means that if we observe the circuit for a very long time, the very slow noise components can randomly assume a very large power level. At such slow rates, noise becomes indistinguishable from thermal drift or aging of devices.

The foregoing observations lead to the following conclusions. First, since the *signals* encountered in most applications do not contain very-low-frequency components, our observation window need not be very long. For example, voice signals display negligible energy below 20 Hz, and if a noise component varies at a lower rate, it does not corrupt the voice significantly. Second, the logarithmic dependence of the flicker noise power upon f_L allows some margin for error in selecting f_L. For example, if the integral in Eq. (7.38) begins from 100 Hz rather than 1 kHz, the coefficient in (7.40) rises from 6.91 to 9.21.

In order to quantify the significance of $1/f$ noise with respect to thermal noise for a given device, we plot both spectral densities on the same axes (Fig. 7.27). Called the $1/f$ noise "corner frequency," the

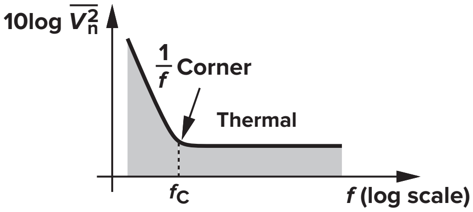


Figure 7.27 Concept of flicker noise corner frequency.

intersection point serves as a measure of what part of the band is mostly corrupted by flicker noise. In the above example, the $1/f$ noise corner, f_C, of the output current is determined as

$$4kT\gamma g_m = \frac{K}{C_{ox}WL} \cdot \frac{1}{f_C} \cdot g_m^2 \tag{7.41}$$

that is,

$$f_C = \frac{K}{\gamma C_{ox}WL} g_m \frac{1}{4kT} \tag{7.42}$$

This result implies that f_C generally depends on the device area and transconductance. Nonetheless, for a given L, the dependence is weak and the $1/f$ noise corner is relatively constant, falling in the vicinity of 10 MHz to 50 MHz for nanometer transistors.

▶ Example 7.9

For a 100-μm/0.5-μm MOS device with $g_m = 1/(100\ \Omega)$, the $1/f$ noise corner frequency is measured to be 500 kHz. If $t_{ox} = 90$ Å, what is the flicker noise coefficient, K, in this technology?

Solution

For $t_{ox} = 90$ Å, we have $C_{ox} = 3.84$ fF/μm^2. Using Eq. (7.42), we write

$$500\ \text{kHz} = \frac{K}{3.84 \times 100 \times 0.5 \times 10^{-15}} \cdot \frac{1}{100} \cdot \frac{3}{8 \times 1.38 \times 10^{-23} \times 300} \tag{7.43}$$

That is, $K = 1.06 \times 10^{-25}$ V^{2}F. ◀

It is important to bear in mind that typical transistor models include thermal and flicker noise but not the gate resistance noise. The latter must therefore be added to each transistor by the designer.

7.3 ■ Representation of Noise in Circuits

Output Noise Consider a general circuit with one input port and one output port (Fig. 7.28). How do we quantify the effect of noise here? The natural approach would be to set the input to zero and calculate the total noise at the output due to various sources of noise in the circuit. This is indeed how the noise is measured in the laboratory or in simulations. Our analysis procedure in Section 7.1.5 methodically leads to the output noise spectrum.

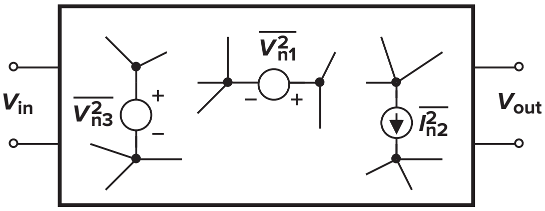


Figure 7.28 Noise sources in a circuit.

▶ Example 7.10

What is the total output noise voltage of the common-source stage shown in Fig. 7.29(a)? Assume that λ = 0.

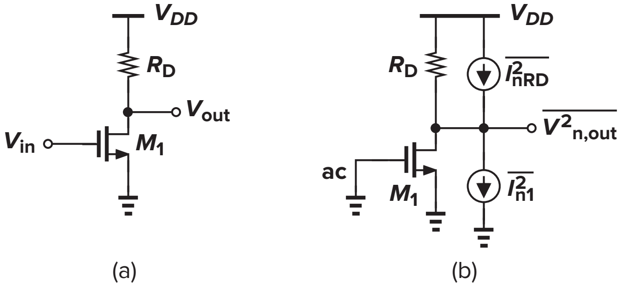


Figure 7.29 (a) CS stage; (b) circuit including noise sources.

Solution

We must identify the sources of noise, find their transfer functions to the output, multiply their spectra by the squared magnitude of the transfer functions and add the results. We model the thermal and flicker noise of M_1 by two current sources: $\overline{I^2_{n,th}} = 4kT\gamma g_m$ and $\overline{I^2_{n,1/f}} = Kg_m^2/(C_{ox}WLf)$. We also represent the thermal noise of R_D by a current source $\overline{I^2_{n,RD}} = 4kT/R_D$. Since these currents flow through R_D, the output noise voltage per unit bandwidth is equal to

$$\overline{V^2_{n,out}} = \left(4kT\gamma g_m + \frac{K}{C_{ox}WL} \cdot \frac{1}{f} \cdot g_m^2 + \frac{4kT}{R_D}\right) R_D^2 \tag{7.44}$$

Note that the noise mechanisms are added as "power" quantities because they are uncorrelated. The value given by (7.44) represents the noise power in 1 Hz at a frequency f. The total output noise is obtained by integration.

Input-Referred Noise While intuitively appealing, the output-referred noise does not allow a fair comparison of the performance of different circuits because it depends on the gain. For example, as depicted in Fig. 7.30, if a common-source stage is followed by a noiseless amplifier having a voltage gain A_1, then the output noise is equal to the expression in (7.44) multiplied by A_1^2. Considering only the output noise, we may conclude that as A_1 increases, the circuit becomes noisier, an incorrect result because a larger A_1 also provides a proportionally higher *signal* level at the output. That is, the output signal-to-noise ratio does not depend on A_1.

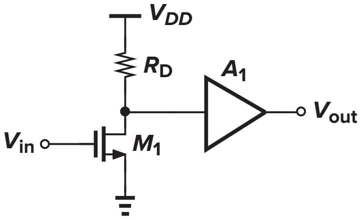


Figure 7.30 Addition of gain stage to a CS stage.

To overcome the above quandary, we usually specify the "input-referred noise" of circuits. Illustrated conceptually in Fig. 7.31, the idea is to represent the effect of all noise sources in the circuit by a single

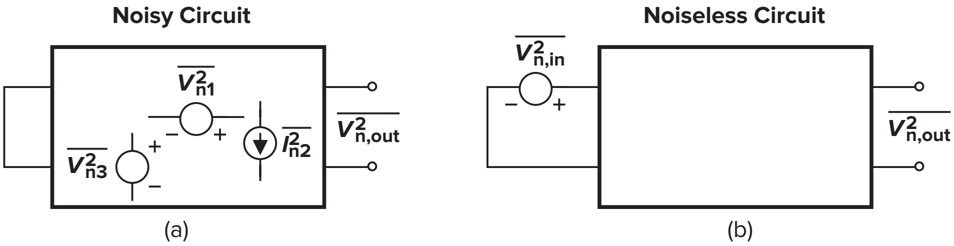


Figure 7.31 Determination of input-referred noise voltage.

source, $\overline{V_{n,in}^2}$, at the input such that the output noise in Fig. 7.31(b) equals that in Fig. 7.31(a). If the voltage gain is A_v, then we must have $\overline{V_{n,out}^2} = A_v^2 \overline{V_{n,in}^2}$, that is, the input-referred noise voltage in this simple case is given by the output noise voltage divided by the gain.

▶ Example 7.11

For the circuit of Fig. 7.29, calculate the input-referred noise voltage.

Solution

We have

$$\overline{V_{n,in}^2} = \frac{\overline{V_{n,out}^2}}{A_v^2} \tag{7.45}$$

$$= \left(4kT\gamma g_m + \frac{K}{C_{ox}WL} \cdot \frac{1}{f} \cdot g_m^2 + \frac{4kT}{R_D}\right) R_D^2 \frac{1}{g_m^2 R_D^2} \tag{7.46}$$

$$= 4kT\frac{\gamma}{g_m} + \frac{K}{C_{ox}WL} \cdot \frac{1}{f} + \frac{4kT}{g_m^2 R_D} \tag{7.47}$$

Note that the first term in (7.47) can be viewed as the thermal noise of a resistor equal to $\gamma/(g_m)$ placed in series with the gate. Similarly, the third term corresponds to the noise of a resistor equal to $(g_m^2 R_D)^{-1}$. We sometimes say the "equivalent thermal noise resistance" of a circuit is equal to R_T, meaning that the total input-referred thermal noise of the circuit in unit bandwidth is equal to $4kTR_T$.

Why does $\overline{V_{n,in}^2}$ decrease as R_D increases? This is because the noise *voltage* due to R_D at the output is proportional to $\sqrt{R_D}$ while the voltage gain of the circuit is proportional to R_D.

◀

At this point of our study, we make two observations. First, the input-referred noise and the input signal are both multiplied by the gain as they are processed by the circuit. Thus, the input-referred noise indicates how much the input signal is corrupted by the circuit's noise, i.e., how small an input the circuit can detect with acceptable SNR. For this reason, input-referred noise allows a fair comparison of different circuits. Second, the input-referred noise is a fictitious quantity in that it cannot be *measured* at the input of the circuit. The two circuits of Figs. 7.31(a) and (b) are *mathematically* equivalent but the physical circuit is still that in Fig. 7.31(a).

In the foregoing discussion, we have assumed that the input-referred noise can be modeled by a single voltage source in series with the input. This is generally an incomplete representation if the circuit has a finite input impedance and is driven by a finite source impedance. To understand why, let us first return to the CS stage of Fig. 7.29 and observe that the output thermal noise due to M_1 is equal to $(4kT\gamma g_m)R_D^2$ regardless of the network driving the gate (i.e., regardless of the preceding stage). Upon dividing this noise by $(g_m R_D)^2$, we obtain an input-referred noise voltage of $4kT\gamma/g_m$—also independent of the preceding stage.

Now, consider the common-source stage of Fig. 7.32(a), where the input capacitance is denoted by C_{in}. The input-referred noise voltage due to M_1 is still given by $4kT\gamma/g_m$. Suppose the preceding stage is modeled by a Thevenin equivalent having an output impedance of R_1 [Fig. 7.32(b)]. Simplifying the circuit for noise calculations as shown in Fig. 7.32(c), we seek the output noise due to M_1, hoping to obtain $4kT\gamma g_m R_D^2$. Owing to the voltage division between R_1 and $1/(C_{in}s)$, the output noise emerges as

$$\overline{V_{n,out}^2} = \overline{V_{n,in}^2} \left| \frac{1}{R_1 C_{in} j\omega + 1} \right|^2 (g_m R_D)^2 \tag{7.48}$$

$$= \frac{4kT\gamma g_m R_D^2}{R_1^2 C_{in}^2 \omega^2 + 1} \tag{7.49}$$

This result is incorrect; after all, the output noise due to M_1 must not diminish as R_1 increases.

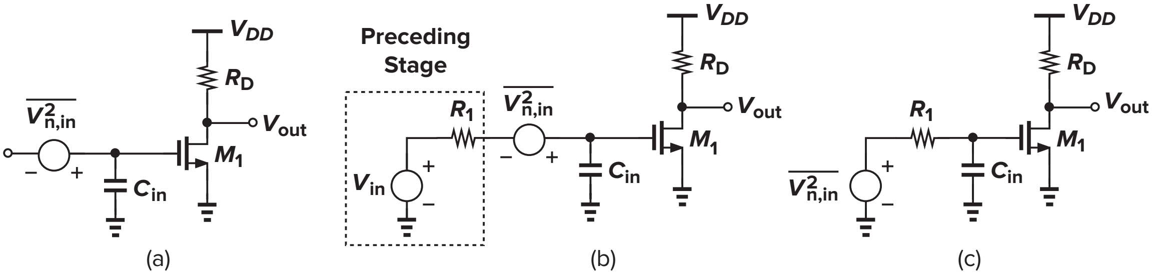


Figure 7.32 CS stage including input capacitance; (b) CS stage stimulated by a finite source impedance; (c) effect of single noise source.

Let us summarize the problem. If the circuit has a finite input impedance, modeling the input-referred noise by merely a voltage source implies that the output noise vanishes as the source impedance becomes large, an incorrect conclusion. To resolve this issue, we model the input-referred noise by both a series voltage source and a parallel current source (Fig. 7.33) so that if the output impedance of the preceding stage assumes large values—thereby reducing the effect of $\overline{V^2_{n,in}}$—the noise current source still flows through a finite impedance, producing noise at the input. It can be proved that $\overline{V^2_{n,in}}$ and $\overline{I^2_{n,in}}$ are necessary and sufficient to represent the noise of any linear two-port circuit [5].

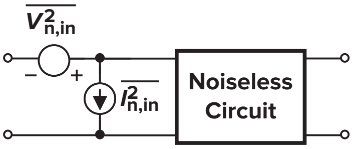


Figure 7.33 Representation of noise by voltage and current sources.

How do we calculate $\overline{V^2_{n,in}}$ and $\overline{I^2_{n,in}}$? Since the model is valid for any source impedance, we consider two extreme cases: zero and infinite source impedances. As shown in Fig. 7.34(a), if the source impedance is zero, $\overline{I^2_{n,in}}$ flows through $\overline{V^2_{n,in}}$ and has no effect on the output. Thus, the output noise measured in this case arises solely from $\overline{V^2_{n,in}}$. Similarly, if the input is open [Fig. 7.34(b)], then $\overline{V^2_{n,in}}$ has no effect and the output noise is due to only $\overline{I^2_{n,in}}$. Let us apply this method to the circuit of Fig. 7.32.

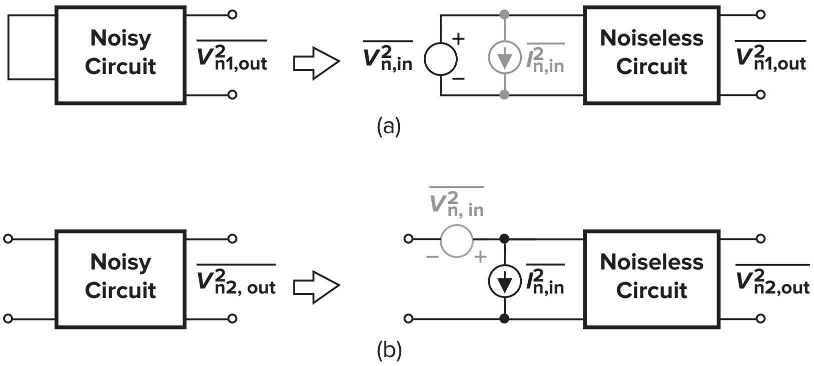


Figure 7.34 Calculation of input-referred noise (a) voltage and (b) current.

▶ **Example 7.12**

Calculate the input-referred noise voltage and current of Fig. 7.32, including only the thermal noise of M_1 and R_D.

Solution

From (7.47), the input-referred noise voltage is simply

$$\overline{V_{n,in}^2} = 4kT\frac{\gamma}{g_m} + \frac{4kT}{g_m^2 R_D} \tag{7.50}$$

As depicted in Fig. 7.35(a), this voltage generates the same output noise as the actual circuit if the input is shorted.

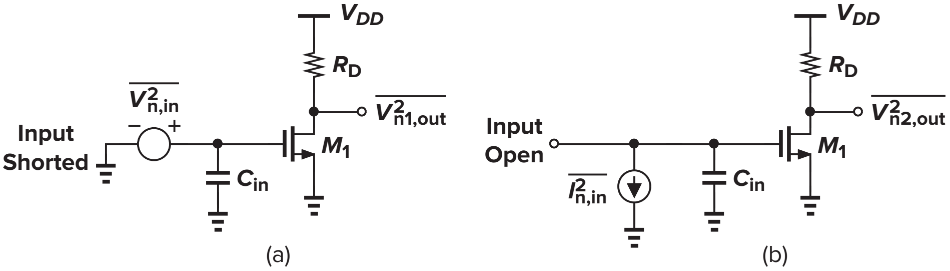


Figure 7.35

To obtain the input-referred noise current, we open the input and find the output noise in terms of $\overline{I_{n,in}^2}$ [Fig. 7.35(b)]. The noise current flows through C_{in}, generating at the output

$$\overline{V_{n2,out}^2} = \overline{I_{n,in}^2}\left(\frac{1}{C_{in}\omega}\right)^2 g_m^2 R_D^2 \tag{7.51}$$

According to Fig. 7.34(b), this value must be equal to the output of the noisy circuit when its input is open:

$$\overline{V_{n2,out}^2} = \left(4kT\gamma g_m + \frac{4kT}{R_D}\right) R_D^2 \tag{7.52}$$

From (7.51) and (7.52), it follows that

$$\overline{I_{n,in}^2} = (C_{in}\omega)^2 \frac{4kT}{g_m^2}\left(\gamma g_m + \frac{1}{R_D}\right) \tag{7.53}$$

As mentioned earlier, the input noise current, $I_{n,in}$, becomes significant if the circuit's input impedance, Z_{in}, is not very high. To see whether $I_{n,in}$ can be neglected or not, we consider the scenario depicted in Fig. 7.36, where Z_S denotes the output impedance of the preceding circuit. The total noise voltage sensed by the second stage at node X is equal to

$$V_{n,X} = \frac{Z_{in}}{Z_{in}+Z_S}V_{n,in} + \frac{Z_{in}Z_S}{Z_{in}+Z_S}I_{n,in} \tag{7.54}$$

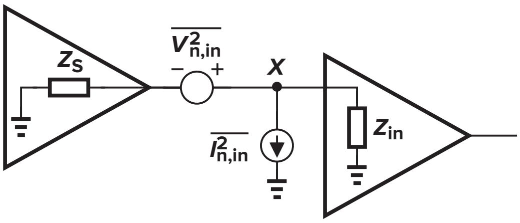


Figure 7.36 Effect of input noise current.

If $\overline{I_{n,in}^2}|Z_S|^2 \ll \overline{V_{n,in}^2}$, then the effect of $I_{n,in}$ is negligible. In other words, ultimately, it is the output impedance of the preceding stage—rather than Z_{in}—that determines the significance of $I_{n,in}$. We conclude that the input-referred noise current can be neglected if

$$|Z_S|^2 \ll \frac{\overline{V_{n,in}^2}}{\overline{I_{n,in}^2}} \tag{7.55}$$

A difficulty in the use of input-referred noise voltages and currents is that they may be *correlated*. After all, $V_{n,in}$ and $I_{n,in}$ may contain effects from the same noise source. For example, in Fig. 7.35, if the noise voltage of R_D is increasing at some point in time, then both $V_{n,in}$ and $I_{n,in}$ also inherit this increase. For this reason, noise calculations must revert to Eq. (7.11) and include the correlation between the two. Methods of avoiding this correlation are described in Appendix A.

The reader may wonder if the use of both a voltage source and a current source to represent the input-referred noise "counts the noise twice." We consider the environment depicted in Fig. 7.37 as an example and prove that the output noise is correct for any source impedance, Z_S. Assuming Z_S is noiseless for simplicity, we first calculate the total noise voltage at the gate of M_1 due to $\overline{V_{n,in}^2}$ and $\overline{I_{n,in}^2}$. This voltage cannot be obtained by superposition of powers because $\overline{V_{n,in}^2}$ and $\overline{I_{n,in}^2}$ are correlated. Nonetheless, superposition still applies to voltages and currents because the circuit is linear and time-invariant. Equations (7.50) and (7.53) must be respectively rewritten as

$$V_{n,in} = V_{n,M1} + \frac{1}{g_m R_D} V_{n,RD} \tag{7.56}$$

$$I_{n,in} = C_{in} s V_{n,M1} + \frac{C_{in} s}{g_m R_D} V_{n,RD} \tag{7.57}$$

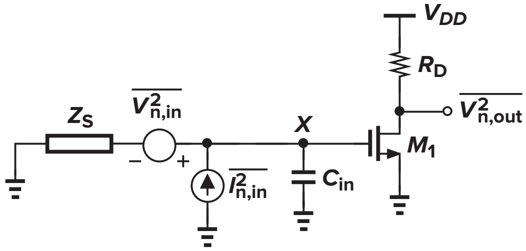


Figure 7.37 CS stage driven by a source impedance.

where $V_{n,M1}$ denotes the gate-referred noise voltage of M_1 and $V_{n,RD}$ the noise voltage of R_D. We recognize that $V_{n,M1}$ and $V_{n,RD}$ appear in both $V_{n,in}$ and $I_{n,in}$, creating a strong correlation between the two. Thus, the calculations must use superposition of voltages—as if $V_{n,in}$ and $I_{n,in}$ were deterministic quantities.

Adding the contributions of $V_{n,in}$ and $I_{n,in}$ at node X in Fig. 7.37, we have

$$V_{n,X} = V_{n,in} \frac{\frac{1}{C_{in}s}}{\frac{1}{C_{in}s} + Z_S} + I_{n,in} \frac{\frac{Z_S}{C_{in}s}}{\frac{1}{C_{in}s} + Z_S} \tag{7.58}$$

$$= \frac{V_{n,in} + I_{n,in} Z_S}{Z_S C_{in} s + 1} \tag{7.59}$$

Substituting for $V_{n,in}$ and $I_{n,in}$ from (7.56) and (7.57), respectively, we obtain

$$\begin{aligned} V_{n,X} &= \frac{1}{Z_S C_{in} s + 1}\left[V_{n,M1} + \frac{1}{g_m R_D} V_{n,RD} + C_{in} s Z_S (V_{n,M1} + \frac{1}{g_m R_D} V_{n,RD})\right] \\ &= V_{n,M1} + \frac{1}{g_m R_D} V_{n,RD} \end{aligned} \tag{7.60}$$

Note that $V_{n,X}$ is independent of Z_S and C_{in}. It follows that

$$\overline{V^2_{n,out}} = g_m^2 R_D^2 \overline{V^2_{n,X}} \tag{7.61}$$

$$= 4kT\left(\gamma g_m + \frac{1}{R_D}\right) R_D^2 \tag{7.62}$$

the same as (7.52). Thus, $V_{n,in}$ and $I_{n,in}$ do not "double count" the noise.

Another Approach In some cases, it is simpler to consider the output short-circuit noise *current*—rather than the output open-circuit noise voltage—for these calculations. This current is then multiplied by the circuit's output resistance to yield the output noise voltage or simply divided by a proper gain to give the input-referred quantities. The following example illustrates this approach.

▶ Example 7.13

Determine the input-referred noise voltage and current for the amplifier shown in Fig. 7.38(a). Assume that I_1 is noiseless and $\lambda = 0$.

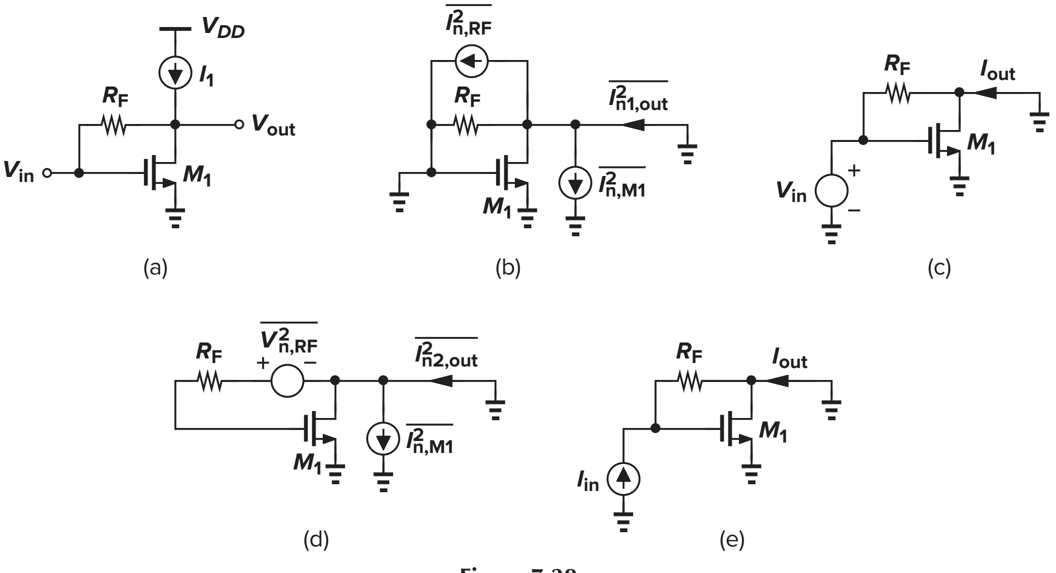


Figure 7.38

Solution

To compute the input-referred noise voltage, we must short the input port. In this case, we can also short the output port as shown in Fig. 7.38(b), and find the output noise current due to R_F and M_1. Since both terminals of R_F are at ac ground, a KVL yields

$$\overline{I^2_{n1,out}} = \frac{4kT}{R_F} + 4kT\gamma g_m \tag{7.63}$$

The output impedance of the circuit with the input shorted is simply equal to R_F, yielding

$$\overline{V_{n1,out}^2} = \left(\frac{4kT}{R_F} + 4kT\gamma g_m\right) R_F^2 \tag{7.64}$$

We can calculate the input-referred noise voltage by dividing (7.64) by the voltage gain or by dividing (7.63) by the *transconductance*, G_m. Let us pursue the latter method. As depicted in Fig. 7.38(c),

$$G_m = \frac{I_{out}}{V_{in}} \tag{7.65}$$

$$= g_m - \frac{1}{R_F} \tag{7.66}$$

Dividing (7.63) by G_m^2 gives

$$\overline{V_{n,in}^2} = \frac{\dfrac{4kT}{R_F} + 4kT\gamma g_m}{(g_m - \dfrac{1}{R_F})^2} \tag{7.67}$$

For the input-referred noise current, we first compute the output noise current with the input left open [Fig. 7.38(d)]. Since $V_{n,RF}$ directly modulates the gate-source voltage of M_1, producing a drain current of $4kTR_Fg_m^2$, we have

$$\overline{I_{n2,out}^2} = 4kTR_Fg_m^2 + 4kT\gamma g_m \tag{7.68}$$

Next, we must determine the current gain of the circuit according to the arrangement shown in Fig. 7.38(c). Noting that $V_{GS} = I_{in}R_F$, and hence $I_D = g_mI_{in}R_F$, we obtain

$$I_{out} = g_mR_FI_{in} - I_{in} \tag{7.69}$$

$$= (g_mR_F - 1)I_{in} \tag{7.70}$$

Dividing (7.68) by the square of the current gain yields

$$\overline{I_{n,in}^2} = \frac{4kTR_Fg_m^2 + 4kT\gamma g_m}{(g_mR_F - 1)^2} \tag{7.71}$$

The reader is encouraged to repeat this analysis using the output noise voltage rather than the output noise current.

The above circuit exemplifies cases where the output noise voltage is not the same for short-circuit and open-circuit input ports. The reader can prove that, if the input is left open, then

$$\overline{V_{n2,out}^2} = \frac{4kT\gamma}{g_m} + 4kTR_F \tag{7.72}$$

◀

7.4 ■ Noise in Single-Stage Amplifiers

Having developed basic mathematical tools and models for noise analysis, we now study the noise performance of single-stage amplifiers at low frequencies. Before considering specific topologies, we describe a lemma that simplifies noise calculations.

Lemma The circuits shown in Fig. 7.39(a) and (b) are equivalent at low frequencies if $\overline{V_n^2} = \overline{I_n^2}/g_m^2$ and the circuits are driven by a finite impedance.

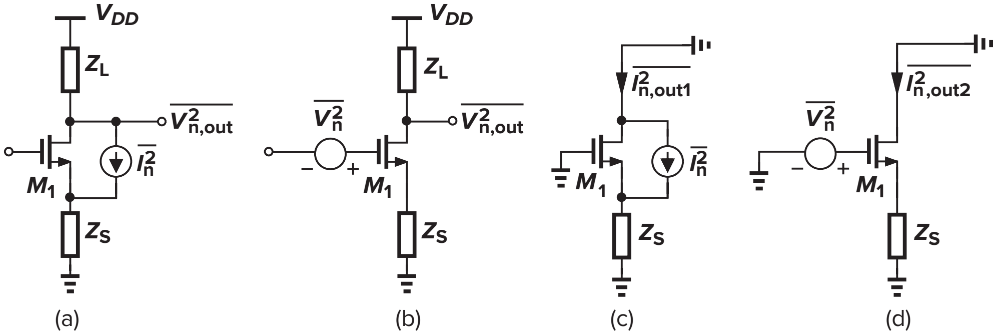


Figure 7.39 Equivalent CS stages.

Proof Since the circuits have equal output impedances, we simply examine the output short-circuit currents [Figs. 7.39(c) and (d)]. It can be proved (Problem 7.4) that the output noise current of the circuit in Fig. 7.39(c) is given by

$$I_{n,out1} = \frac{I_n}{Z_S(g_m + g_{mb} + 1/r_O) + 1} \tag{7.73}$$

and that of Fig. 7.39(d) is

$$I_{n,out2} = \frac{g_m V_n}{Z_S(g_m + g_{mb} + 1/r_O) + 1} \tag{7.74}$$

Equating (7.73) and (7.74), we have $V_n = I_n/g_m$. We call V_n the "gate-referred" noise of M_1.

This lemma suggests that the noise source can be transformed from a drain-source current to a gate series voltage for arbitrary Z_S. We repeat this analysis in the presence of the gate-source capacitance in Problem 7.29.

▶ **Example 7.14**

Prove the above lemma using Thevenin equivalents.

Solution

We construct a Thevenin model for the circuits in Figs. 7.39(a) and (b) but exclude Z_L, as depicted in Figs. 7.40(a) and (b). With $I_n = 0$ and $V_n = 0$, the two topologies are identical, and hence $Z_{Thev1} = Z_{Thev2}$. We thus need only find the condition under which $V_{Thev1} = V_{Thev2}$.

To obtain the Thevenin voltages, we must replace Z_L with an open circuit [Fig. 7.40(c)].[10] Since the current flowing through Z_S is zero in both circuits, we have $V_{Thev1} = I_n r_O$ and $V_{Thev2} = g_m V_n r_O$. It follows that $V_n = I_n/g_m$. ◀

7.4.1 Common-Source Stage

From Example 7.11, the input-referred noise voltage per unit bandwidth of a simple CS stage is equal to

$$\overline{V^2_{n,in}} = 4kT\left(\frac{\gamma}{g_m} + \frac{1}{g_m^2 R_D}\right) + \frac{K}{C_{ox}WL}\frac{1}{f} \tag{7.75}$$

[10] The Thevenin voltage is calculated by disconnecting the port of interest from external loads.

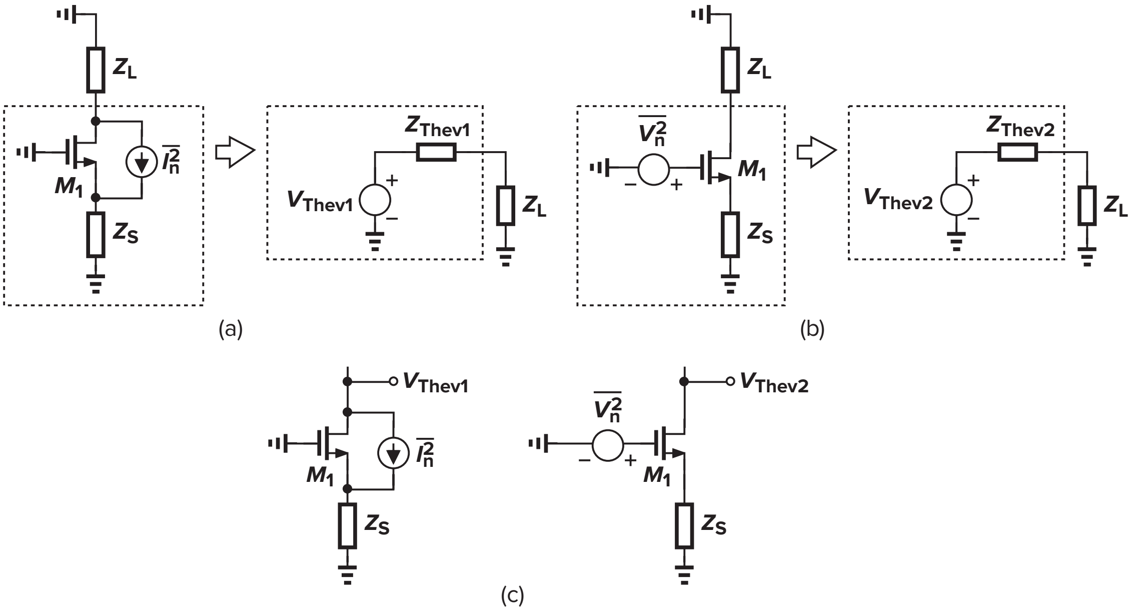


Figure 7.40

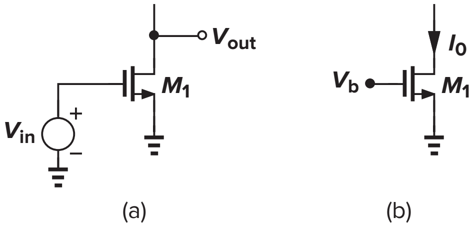


Figure 7.41 Voltage amplification versus current generation.

From the above lemma, we recognize that the term $4kT\gamma/g_m$ is in fact the thermal noise current of M_1 expressed as a voltage in series with the gate.

How can we reduce the input-referred noise voltage? Equation (7.75) implies that the transconductance of M_1 must be maximized. Thus, the transconductance must be maximized if the transistor is to amplify a voltage signal applied to its gate [Fig. 7.41(a)] whereas it must be minimized if the transistor operates as a constant current source [Fig. 7.41(b)], as illustrated by the following example.

▶ Example 7.15

Calculate the input-referred thermal noise voltage of the amplifier shown in Fig. 7.42(a), assuming both transistors are in saturation. Also, determine the total output thermal noise if the circuit drives a load capacitance C_L. What is the output signal-to-noise ratio if a low-frequency sinusoid of peak amplitude V_m is applied to the input?

Solution

Representing the thermal noise of M_1 and M_2 by current sources [Fig. 7.42(b)] and noting that they are uncorrelated, we write

$$\overline{V_{n,out}^2} = 4kT(\gamma g_{m1} + \gamma g_{m2})(r_{O1}\|r_{O2})^2 \tag{7.76}$$

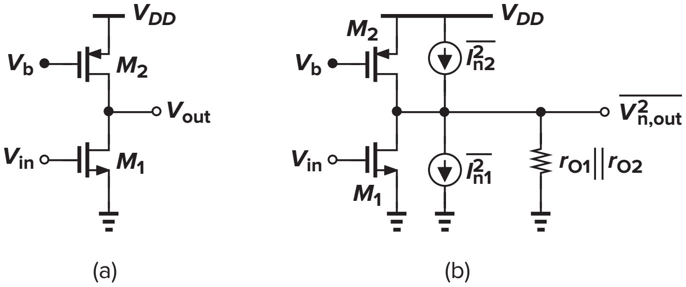


Figure 7.42

(In reality, γ may not be the same for NMOS and PMOS devices.) Since the voltage gain is equal to $g_{m1}(r_{O1}\|r_{O2})$, the total noise voltage referred to the gate of M_1 is

$$\overline{V_{n,in}^2} = 4kT(\gamma g_{m1} + \gamma g_{m2})\frac{1}{g_{m1}^2} \tag{7.77}$$

$$= 4kT\gamma\left(\frac{1}{g_{m1}} + \frac{g_{m2}}{g_{m1}^2}\right) \tag{7.78}$$

Equation (7.78) reveals the dependence of $\overline{V_{n,in}^2}$ upon g_{m1} and g_{m2}, confirming that g_{m2} must be minimized because M_2 serves as a current source rather than a transconductor.[11]

The reader may wonder why M_1 and M_2 in Fig. 7.42 exhibit different noise effects. After all, if the noise currents of both transistors flow through $r_{O1}\|r_{O2}$, why should g_{m1} be maximized and g_{m2} minimized? This is simply because, as g_{m1} increases, the output noise *voltage* rises in proportion to $\sqrt{g_{m1}}$ whereas the *voltage gain* of the stage increases in proportion to g_{m1}. As a result, the input-referred noise voltage decreases. Such a trend does not apply to M_2.

To compute the total output noise, we integrate (7.76) across the band:

$$\overline{V_{n,out,tot}^2} = \int_0^\infty 4kT\gamma(g_{m1} + g_{m2})(r_{O1}\|r_{O2})^2 \frac{df}{1 + (r_{O1}\|r_{O2})^2 C_L^2 (2\pi f)^2} \tag{7.79}$$

Using the results of Example 7.3, we have

$$\overline{V_{n,out,tot}^2} = \gamma(g_{m1} + g_{m2})(r_{O1}\|r_{O2})\frac{kT}{C_L} \tag{7.80}$$

A low-frequency input sinusoid of amplitude V_m yields an output amplitude equal to $g_{m1}(r_{O1}\|r_{O2})V_m$. The output SNR is equal to the ratio of the signal power and the noise power:

$$\text{SNR}_{out} = \left[\frac{g_{m1}(r_{O1}\|r_{O2})V_m}{\sqrt{2}}\right]^2 \cdot \frac{1}{\gamma(g_{m1} + g_{m2})(r_{O1}\|r_{O2})(kT/C_L)} \tag{7.81}$$

$$= \frac{C_L}{2\gamma kT} \cdot \frac{g_{m1}^2(r_{O1}\|r_{O2})}{g_{m1} + g_{m2}} V_m^2 \tag{7.82}$$

We note that to maximize the output SNR, C_L must be maximized, i.e., the bandwidth must be minimized. Of course, the bandwidth is also dictated by the input signal spectrum. This example indicates that it becomes exceedingly difficult to design broadband circuits while maintaining low noise.

[11] A device or a circuit that converts a voltage to a current is called a transconductor or a V/I converter.

▶ **Example 7.16**

Determine the input-referred thermal noise voltage of the complementary common-source stage shown in Fig. 7.43.

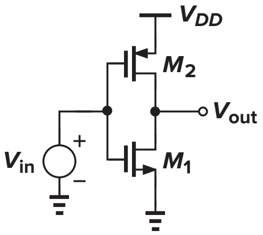


Figure 7.43

Solution

With the input signal set to zero, this circuit produces the same output noise voltage as the circuit in Fig. 7.42(a) does. But the complementary stage provides a higher voltage gain, $(g_{m1} + g_{m2})(r_{O1}||r_{O2})$. The input-referred noise voltage is thus given by

$$\overline{V_{n,in}^2} = \frac{4kT\gamma}{g_{m1} + g_{m2}} \tag{7.83}$$

an expected result because M_1 and M_2 operate in "parallel," and hence their transconductances add. Why does this topology exhibit a lower input noise than the circuit of Fig. 7.42(a)? In both cases, M_2 injects noise to the output node, but in the complementary stage, this device operates as a transconductor and *amplifies* the input. ◀

For a simple CS stage with resistive load, Eq. (7.75) suggests that the thermal noise can be reduced by increasing the bias current. But, for a given headroom, this requires that we decrease R_D and hence increase its noise contribution. In order to quantify this trade-off, we express g_m as $2I_D/(V_{GS} - V_{TH})$ and write the input-referred thermal noise as

$$\overline{V_{n,in}^2} = 4kT\left[\frac{\gamma(V_{GS} - V_{TH})}{2I_D} + \frac{(V_{GS} - V_{TH})^2}{4I_D \cdot I_D R_D}\right] \tag{7.84}$$

This equation suggests that $V_{n,in}$ falls if I_D is increased and $I_D R_D$ kept constant provided that $V_{GS} - V_{TH}$ also remains constant, i.e., if the transistor width increases in proportion to I_D.

▶ **Example 7.17**

Calculate the input-referred $1/f$ and thermal noise voltage of the CS stage depicted in Fig. 7.44(a), assuming M_1 and M_2 are in saturation.

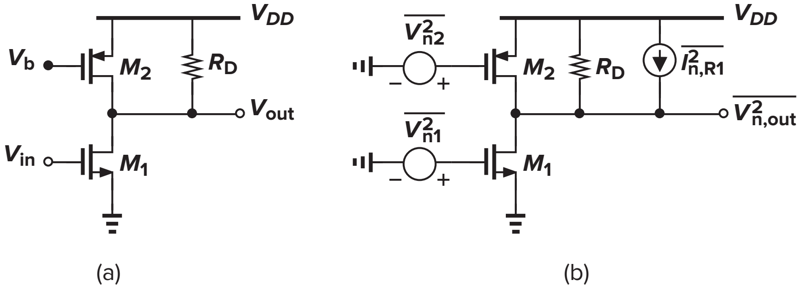


Figure 7.44

Solution

We model the $1/f$ and thermal noise of the transistors as voltage sources in series with their gates [Fig. 7.44(b)]. The noise voltage at the gate of M_2 experiences a gain of $g_{m2}(R_D \| r_{O1} \| r_{O2})$ as it appears at the output. The result must then be divided by $g_{m1}(R_D \| r_{O1} \| r_{O2})$ to be referred to the main input. The noise current of R_D is multiplied by $R_D \| r_{O1} \| r_{O2}$ and divided by $g_{m1}(R_D \| r_{O1} \| r_{O2})$. Thus, the overall input-referred noise voltage is given by

$$\overline{V_{n,in}^2} = 4kT\gamma\left(\frac{g_{m2}}{g_{m1}^2} + \frac{1}{g_{m1}}\right) + \frac{1}{C_{ox}}\left[\frac{K_P g_{m2}^2}{(WL)_2 g_{m1}^2} + \frac{K_N}{(WL)_1}\right]\frac{1}{f} + \frac{4kT}{g_{m1}^2 R_D} \tag{7.85}$$

where K_P and K_N denote the flicker noise coefficients of PMOS and NMOS devices, respectively. Note that the circuit reduces to that in Fig. 7.42(a) or 7.29(a) if $R_D = \infty$ or $g_{m2} = 0$, respectively. How should the bias current of M_2 be chosen to minimize $V_{n,in}$ if the dc voltage drop across R_D is fixed? This is left as an exercise for the reader.

How do we design a common-source stage for low-noise operation? For thermal noise in the simple topology of Fig. 7.41, we must maximize g_{m1} by increasing the drain current or the device width. A higher I_D translates to greater power dissipation and limited output voltage swings while a wider device leads to larger input and output capacitance. We can also increase R_D, but at the cost of limiting the voltage headroom and lowering the speed.

For $1/f$ noise, the primary approach is to increase the area of the transistor. If WL is increased while W/L remains constant, then the device transconductance, and hence its thermal noise, do not change, but the device capacitances increase. These observations point to the trade-offs between noise, power dissipation, voltage headroom, and speed.

▶ Example 7.18

A student writes the drain flicker noise current of a MOS device as $[K/(WLC_{ox}f)]g_m^2 = [K/(WLC_{ox}f)]$ $(\sqrt{2\mu_n C_{ox}(W/L)I_D})^2 = 2K\mu_n I_D/(L^2 f)$, concluding that the flicker noise current is independent of W. Explain the flaw in this argument.

Solution

A fair comparison must keep both the overdrive and I_D constant as W changes. (If we allow $V_{GS} - V_{TH}$ to change, then the drain voltage headroom also changes.). Thus, we can express the drain flicker noise current as $[K/(WLC_{ox}f)](4I_D^2)/(V_{GS} - V_{TH})^2$, which reveals that the noise current decreases as WL increases.

▶ Example 7.19

Design a resistively-loaded common-source stage with a total input-referred noise voltage of 100 μV$_{\text{rms}}$, a power budget of 1 mW, a bandwidth of 1 GHz, and a supply voltage of 1 V. Neglect channel-length modulation and flicker noise and assume that the bandwidth is limited by the load capacitance.

Solution

Illustrated in Fig. 7.45(a), the circuit produces noise at the output in a bandwidth given by R_D and C_L. From the noise model shown in Fig. 7.45(b), the reader can derive a Thevenin equivalent for the circuit in the dashed box, obtaining the output noise spectrum as

$$\overline{V_{n,out}^2} = (\overline{V_{n,RD}^2} + R_D^2\overline{I_{n,M1}^2})\frac{1}{R_D^2 C_L^2\omega^2 + 1} \tag{7.86}$$

$$= (4kTR_D + 4kT\gamma g_m R_D^2)\frac{1}{R_D^2 C_L^2\omega^2 + 1} \tag{7.87}$$

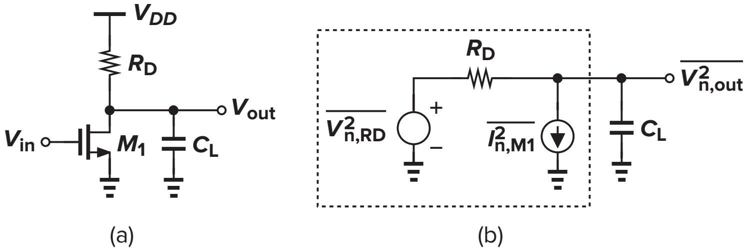


Figure 7.45

Since we know that the integral of $4kTR_D/(R_D^2C_L^2\omega^2+1)$ from 0 to ∞ yields a value of kT/C_L, we manipulate the transistor noise contribution as follows:

$$\overline{V^2_{n,out}} = \frac{4kTR_D}{R_D^2C_L^2\omega^2+1} + \gamma g_m R_D \frac{4kTR_D}{R_D^2C_L^2\omega^2+1} \tag{7.88}$$

Integration from 0 to ∞ thus gives

$$\overline{V^2_{n,out,tot}} = \frac{kT}{C_L} + \gamma g_m R_D \frac{kT}{C_L} \tag{7.89}$$

$$= (1+\gamma g_m R_D)\frac{kT}{C_L} \tag{7.90}$$

This noise must be divided by $g_m^2R_D^2$ and equated to $(100\ \mu\mathrm{V_{rms}})^2$. We also note that $1/(2\pi R_D C_L) = 1$ GHz and $kT = 4.14\times10^{-21}$ J at the room temperature, arriving at

$$\frac{1+\gamma g_m R_D}{g_m^2 R_D}\cdot\frac{2\pi kT}{2\pi R_D C_L} = (100\ \mu\mathrm{V_{rms}})^2 \tag{7.91}$$

and hence

$$\frac{1}{g_m}\left(\frac{1}{g_m R_D}+\gamma\right) = 384\ \Omega \tag{7.92}$$

We have some flexibility in the choice of g_m and R_D here. For example, if $g_m R_D = 3$ and $\gamma = 1$, then $1/g_m = 288\ \Omega$ and $R_D = 864\ \Omega$. With a drain-current budget of 1 mW/$V_{DD} = 1$ mA, we can choose W/L so as to obtain this amount of transconductance.

The above choice of the voltage gain and the resulting values of R_D and g_m must be checked against the bias conditions. Since $R_D I_D = 864$ mV, $V_{DS,min} = 136$ mV, leaving little headroom for voltage swings. The reader is encouraged to try $g_m R_D = 2$ or 4 to see how the voltage headroom depends on the choice of the gain.

7.4.2 Common-Gate Stage

Thermal Noise Consider the common-gate configuration shown in Fig. 7.46(a). Neglecting channel-length modulation, we represent the thermal noise of M_1 and R_D by two current sources [Fig. 7.46(b)]. Note that, owing to the low input impedance of the circuit, the input-referred noise current is not negligible even at low frequencies. To calculate the input-referred noise voltage, we short the input to ground and equate the output noises of the circuits in Figs. 7.47(a) and (b):

$$\left(4kT\gamma g_m + \frac{4kT}{R_D}\right)R_D^2 = \overline{V^2_{n,in}}(g_m+g_{mb})^2R_D^2 \tag{7.93}$$

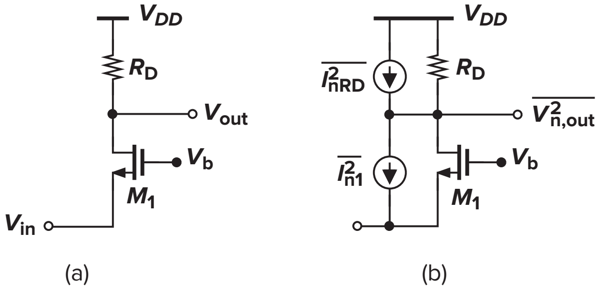


Figure 7.46 (a) CG stage; (b) circuit including noise sources.

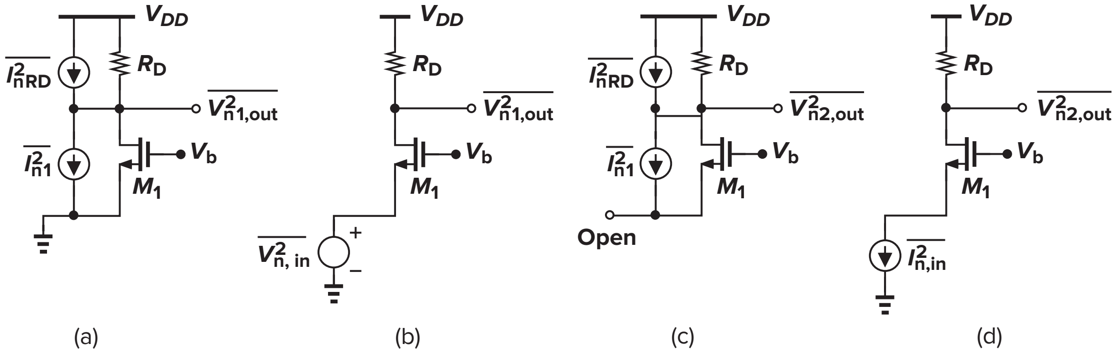


Figure 7.47 Calculation of input-referred noise of a CG stage.

That is

$$\overline{V_{n,in}^2} = \frac{4kT(\gamma g_m + 1/R_D)}{(g_m + g_{mb})^2} \tag{7.94}$$

Similarly, equating the output noises of the circuits in Figs. 7.47(c) and (d) yields the input-referred noise current. What is the effect of $\overline{I_{n1}^2}$ at the output in Fig. 7.47(c)? Since the sum of the currents at the source of M_1 is zero, $I_{n1} + I_{D1} = 0$. Consequently, I_{n1} creates an equal and opposite current in M_1, producing *no* noise at the output. The output noise voltage of Fig. 7.46(a) is therefore equal to $4kTR_D$, and hence $\overline{I_{n,in}^2} R_D^2 = 4kTR_D$. That is

$$\overline{I_{n,in}^2} = \frac{4kT}{R_D} \tag{7.95}$$

An important drawback of the common-gate topology is that it directly refers the noise current produced by the load to the input. Exemplified by (7.95), this effect arises because such a circuit provides no *current* gain, a point of contrast to common-source amplifiers.

We have thus far neglected the noise contributed by the bias-current source of a common-gate stage. Shown in Fig. 7.48 is a simple mirror arrangement establishing the bias current of M_1 as a multiple of I_1. Capacitor C_0 shunts the noise generated by M_0 to ground. We note that if the input of the circuit is shorted to ground, then the drain noise current of M_2 does not flow through R_D, contributing no input-referred noise voltage. On the other hand, if the input is open, all of $\overline{I_{n2}^2}$ flows from M_1 and R_D (at low frequencies), producing an output noise equal to $\overline{I_{n2}^2} R_D^2$ and hence an input-referred noise current of $\overline{I_{n2}^2}$. As a result, the noise current of M_2 directly adds to the input-referred noise current, making it desirable to *minimize*

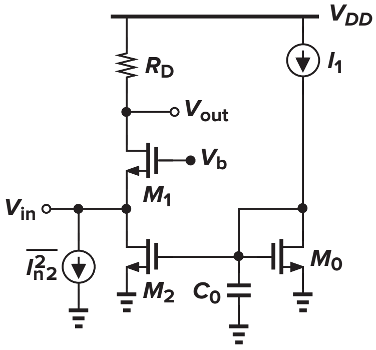


Figure 7.48 Noise contributed by bias-current source.

the transconductance of M_2. For a given bias current, however, this translates to a higher drain-source voltage for M_2 because $g_{m2} = 2I_{D2}/(V_{GS2} - V_{TH2})$, requiring a high value for V_b and limiting the voltage swing at the output node.

▶ **Example 7.20**

Calculate the input-referred thermal noise voltage and current of the circuit shown in Fig. 7.49 assuming that all of the transistors are in saturation.

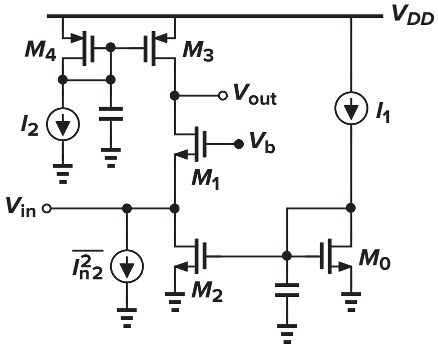


Figure 7.49

Solution

To compute the input-referred noise voltage, we short the input to ground, obtaining

$$\overline{V_{n1,out}^2} = 4kT\gamma(g_{m1} + g_{m3})(r_{O1}\|r_{O3})^2 \tag{7.96}$$

Thus, the input-referred noise voltage, $V_{n,in}$, must satisfy this relationship:

$$\overline{V_{n,in}^2}(g_{m1} + g_{mb1})^2(r_{O1}\|r_{O3})^2 = 4kT\gamma(g_{m1} + g_{m3})(r_{O1}\|r_{O3})^2 \tag{7.97}$$

where the voltage gain from V_{in} to V_{out} is approximated by $(g_{m1} + g_{mb1})(r_{O1}||r_{O3})$. It follows that

$$\overline{V_{n,in}^2} = 4kT\gamma\frac{(g_{m1} + g_{m3})}{(g_{m1} + g_{mb1})^2} \tag{7.98}$$

As expected, the noise is proportional to g_{m3}.

To calculate the input-referred noise current, we open the input and note that the output noise voltage due to M_3 is simply given by $\overline{I_{n3}^2}R_{out}^2$, where $R_{out} = r_{O3}||[r_{O2} + (g_{m1} + g_{mb1})r_{O1}r_{O2} + r_{O1}]$ denotes the output impedance when the input is open. The reader can prove that, in response to an input current I_{in}, the circuit generates an output voltage given by

$$V_{out} = \frac{(g_{m1} + g_{mb1})r_{O1} + 1}{r_{O1} + (g_{m1} + g_{mb1})r_{O1}r_{O2} + r_{O2} + r_{O3}} r_{O3} r_{O2} I_{in} \tag{7.99}$$

Dividing $I_{n3}R_{out}$ by this gain to refer the noise of M_3 to the input, we have

$$I_{n,in}|_{M3} = \frac{r_{O2} + (g_{m1} + g_{mb1})r_{O1}r_{O2} + r_{O1}}{r_{O2}[(g_{m1} + g_{mb1})r_{O1} + 1]} I_{n3} \tag{7.100}$$

which reduces to

$$I_{n,in}|_{M3} \approx I_{n3} \tag{7.101}$$

$$\approx 4kT\gamma g_{m3} \tag{7.102}$$

if any $g_m r_O$ product is much greater than unity. Since the noise current of M_2 directly adds to the input, we have

$$\overline{I^2_{n,in}} = 4kT\gamma(g_{m2} + g_{m3}) \tag{7.103}$$

Again, the noise is proportional to the transconductance of the two current sources. In the above calculations, we have neglected the effect of I_{n1} when the input is left open even though the source of M_1 sees a finite degeneration (r_{O2}). In Problem 7.31, we refer this noise to the input and prove that it is still negligible.

Flicker Noise The effect of $1/f$ noise in a common-gate topology is also of interest. As a typical case, we compute the input-referred $1/f$ noise voltage and current of the circuit shown in Fig. 7.49. Illustrated in Fig. 7.50, each $1/f$ noise generator is modeled by a voltage source in series with the gate of the corresponding transistor. Note that the $1/f$ noise of M_0 and M_4 is neglected. A more realistic case is studied in Problem 7.10.

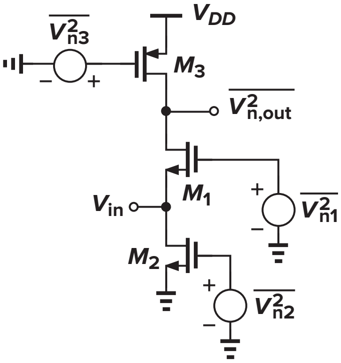


Figure 7.50 Flicker noise in a CG stage.

With the input shorted to ground, we have

$$\overline{V^2_{n1,out}} = \frac{1}{C_{ox}f}\left[\frac{g^2_{m1}K_N}{(WL)_1} + \frac{g^2_{m3}K_P}{(WL)_3}\right](r_{O1}\|r_{O3})^2 \tag{7.104}$$

where K_N and K_P denote the flicker noise coefficients of NMOS and PMOS devices, respectively. Approximating the voltage gain as $(g_{m1} + g_{mb1})(r_{O1}||r_{O3})$, we obtain

$$\overline{V^2_{n,in}} = \frac{1}{C_{ox}f}\left[\frac{g^2_{m1}K_N}{(WL)_1} + \frac{g^2_{m3}K_P}{(WL)_3}\right]\frac{1}{(g_{m1} + g_{mb1})^2} \tag{7.105}$$

With the input open, the output noise voltage is approximately given by

$$\overline{V_{n2,out}^2} = \frac{1}{C_{ox}f}\left[\frac{g_{m2}^2 K_N}{(WL)_2} + \frac{g_{m3}^2 K_P}{(WL)_3}\right] R_{out}^2 \tag{7.106}$$

where it is assumed that the transconductance from the gate of M_2 to the output is equal to g_{m2}. It follows that

$$\overline{I_{n,in}^2} = \frac{1}{C_{ox}f}\left[\frac{g_{m2}^2 K_N}{(WL)_2} + \frac{g_{m3}^2 K_P}{(WL)_3}\right] \tag{7.107}$$

Equations (7.105) and (7.107) describe the $1/f$ noise behavior of the circuit and must be added to (7.98) and (7.103), respectively, to obtain the overall noise per unit bandwidth.

7.4.3 Source Followers

Consider the source follower depicted in Fig. 7.51(a), where M_2 serves as the bias-current source. Since the input impedance of the circuit is quite high, even at relatively high frequencies, the input-referred noise current can usually be neglected for moderate driving source impedances. To compute the input-referred thermal noise voltage, we employ the representation in Fig. 7.51(b), expressing the output noise due to M_2 as

$$\overline{V_{n,out}^2}|_{M2} = \overline{I_{n2}^2}\left(\frac{1}{g_{m1}}\bigg\|\frac{1}{g_{mb1}}\bigg\| r_{O1}\| r_{O2}\right)^2 \tag{7.108}$$

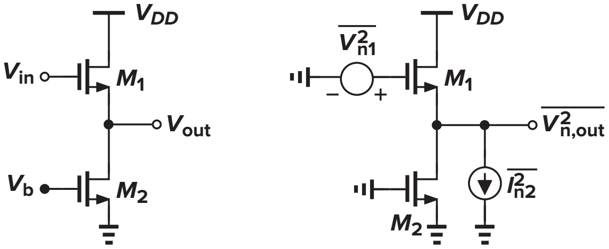


Figure 7.51 (a) Source follower; (b) circuit including noise sources.

From Chapter 3,

$$A_v = \frac{\dfrac{1}{g_{mb1}}\bigg\| r_{O1}\| r_{O2}}{\dfrac{1}{g_{mb1}}\bigg\| r_{O1}\| r_{O2} + \dfrac{1}{g_{m1}}} \tag{7.109}$$

Thus, the total input-referred noise voltage is

$$\overline{V_{n,in}^2} = \overline{V_{n1}^2} + \frac{\overline{V_{n,out}^2}|_{M2}}{A_v^2} \tag{7.110}$$

$$= 4kT\gamma\left(\frac{1}{g_{m1}} + \frac{g_{m2}}{g_{m1}^2}\right) \tag{7.111}$$

Note the similarity between (7.78) and (7.111).

Since source followers add noise to the input signal while providing a voltage gain of less than unity, they are usually avoided in low-noise amplification. The $1/f$ noise performance of source followers is studied in Problem 7.11.

7.4.4 Cascode Stage

Consider the cascode stage of Fig. 7.52(a). Since at low frequencies the noise currents of M_1 and R_D mostly flow through R_D, the noise contributed by these two devices is quantified as in a common-source stage:

$$\overline{V_{n,in}^2}|_{M1,RD} = 4kT\left(\frac{\gamma}{g_{m1}} + \frac{1}{g_{m1}^2 R_D}\right) \tag{7.112}$$

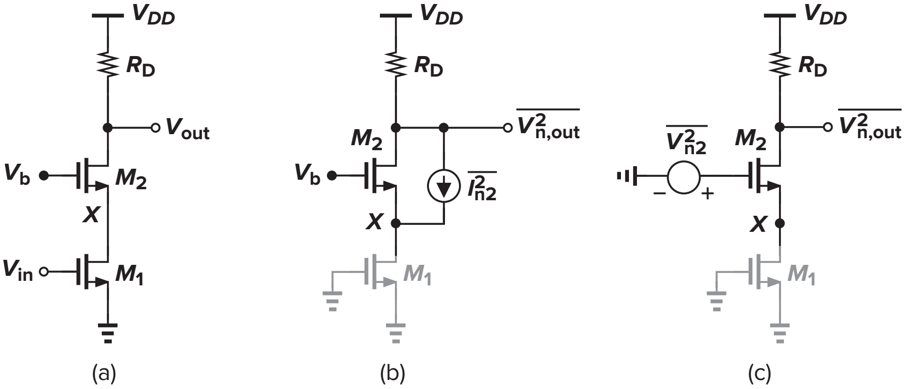


Figure 7.52 (a) Cascode stage; (b) noise of M_2 modeled by a current source; (c) noise of M_2 modeled by a voltage source.

where $1/f$ noise of M_1 is ignored. What is the effect of noise of M_2? Modeled as in Fig. 7.52(b), this noise contributes negligibly to the output, especially at low frequencies. This is because, if channel-length modulation in M_1 is neglected, then $I_{n2} + I_{D2} = 0$, and hence M_2 does not affect $V_{n,out}$. From another point of view, using the lemma of Fig. 7.39 to construct the equivalent in Fig. 7.52(c), we note that the voltage gain from V_{n2} to the output is quite small if the impedance at node X is large. At high frequencies, on the other hand, the total capacitance at node X, C_X, gives rise to a gain:

$$\frac{V_{n,out}}{V_{n2}} \approx \frac{-R_D}{1/g_{m2} + 1/(C_X s)} \tag{7.113}$$

increasing the output noise. This capacitance also reduces the gain from the main input to the output by shunting the signal current produced by M_1 to ground. As a result, the input-referred noise of a cascode stage may rise considerably at high frequencies.

If R_D in Fig. 7.52(c) is large, e.g., if it represents the output resistance of a PMOS cascode load, then the gain from V_{n2} to V_{out} may not be small. The reader can show that, if $R_D \approx g_m r_O^2$ (for a cascode), then V_{out}/V_n is still much greater, making the contribution of V_n negligible.

7.5 ■ Noise in Current Mirrors

The noise produced by the devices in current mirrors may propagate to the output of interest. In Figs. 7.48 and 7.49, for example, the diode-connected device may contribute substantial *flicker* noise unless an extremely large bypass capacitor is used. This effect is exacerbated by the bias-current multiplication factor in the current mirror.

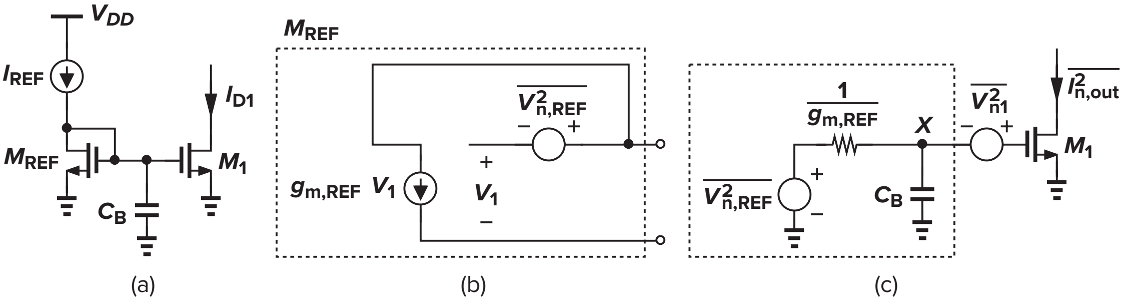


Figure 7.53 (a) Current mirror using a capacitor to suppress diode-connected device's noise, (b) small-signal model, and (c) overall equivalent circuit.

To appreciate the difficulty with current-mirror flicker noise, let us study the simple topology shown in Fig. 7.53(a), where $(W/L)_1 = N(W/L)_{REF}$. The multiplication factor, N, is in the range of 5 to 10 so as to minimize the power consumed by the reference branch. We wish to determine the flicker noise in I_{D1}. We assume that $\lambda = 0$ and I_{REF} is noiseless but caution the reader that, as described in Chapter 12, the noise of the reference (bandgap) current may not be negligible. We first construct a Thevenin equivalent for M_{REF} and its flicker noise, $V_{n,REF}$: as depicted in Fig. 7.53(b), the open-circuit voltage is equal to $V_{n,REF}$ because V_1 must be zero (why?). Noting that the Thevenin resistance is equal to $1/g_{m,REF}$, we arrive at the arrangement in Fig. 7.53(c), where the noise voltage at node X and V_{n1} add (without correlation) and drive the gate of M_1, producing

$$\overline{I^2_{n,out}} = \left(\frac{g^2_{m,REF}}{C_B^2\omega^2 + g^2_{m,REF}}\overline{V^2_{n,REF}} + \overline{V^2_{n1}}\right)g^2_{m1} \tag{7.114}$$

Since $(W/L)_1 = N(W/L)_{REF}$ and, typically, $L_1 = L_{REF}$, we observe that $\overline{V^2_{n,REF}} = N\overline{V^2_{n1}}$ because the flicker noise power spectral density is inversely proportional to the channel area, WL. It follows that

$$\overline{I^2_{n,out}} = \left(\frac{Ng^2_{m,REF}}{C_B^2\omega^2 + g^2_{m,REF}} + 1\right)g^2_{m1}\overline{V^2_{n1}} \tag{7.115}$$

For the noise of the diode-connected device to be negligible, we must ensure that the first term inside the parentheses is small:

$$(N-1)g^2_{m,REF} \ll C_B^2\omega^2 \tag{7.116}$$

and hence

$$C_B^2 \gg \frac{(N-1)g^2_{m,REF}}{\omega^2} \tag{7.117}$$

For example, if $N = 5$, $g_{m,REF} \approx 1/(200\ \Omega)$, and the minimum frequency of interest is 1 MHz, we have $C_B^2 \gg 2.533 \times 10^{-18}$ F. For a tenfold suppression of the M_{REF} noise, this translates to 5.03 nF!

In order to reduce the noise contributed by M_{REF} while avoiding such a large capacitor, we can insert a resistance between its gate and C_B [Fig. 7.54(a)] and revise Eq. (7.114) as

$$\overline{I^2_{n,out}} = \left[\frac{g^2_{m,REF}}{(1+g_{m,REF}R_B)^2C_B^2\omega^2 + g^2_{m,REF}}(\overline{V^2_{n,REF}} + \overline{V^2_{n,RB}}) + \overline{V^2_{n1}}\right]g^2_{m1} \tag{7.118}$$

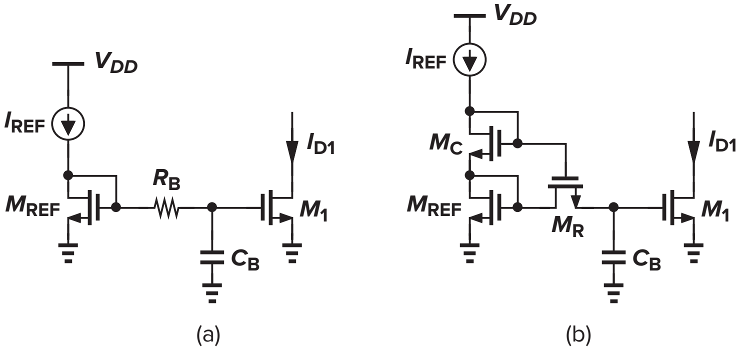


Figure 7.54 (a) Use of a resistor to filter a diode-connected device's noise, and (b) realization of the resistor by a MOSFET.

The series resistance lowers the filter cutoff frequency to $[(1/g_{m1,REF} + R_B)C_B]^{-1}$ but also contributes its own noise. We can thus increase R_B before $\overline{V_{n,RB}^2}$ becomes an appreciable fraction of $\overline{V_{n,REF}^2}$.

In practice, R_B can be quite large before its thermal noise becomes comparable with the flicker noise of M_{REF}. The upper bound on R_B is therefore dictated by the area trade-off between R_B and C_B.[12] We thus seek a circuit arrangement that provides a high resistance and occupies a moderate area. Fortunately, we have developed such a topology in Chapter 5: as shown in Fig. 7.54(b), a MOS device, M_R, with a small, but controlled overdrive serves our purpose. As explained in Chapter 5, M_R is chosen narrow and long, and M_C wide and short.

7.6 ■ Noise in Differential Pairs

With our understanding of noise in basic amplifiers, we can now study the noise behavior of differential pairs. Shown in Fig. 7.55(a), a differential pair can be viewed as a two-port circuit. It is therefore possible to model the overall noise as depicted in Fig. 7.55(b). For low-frequency operation, $\overline{I_{n,in}^2}$ is negligible.

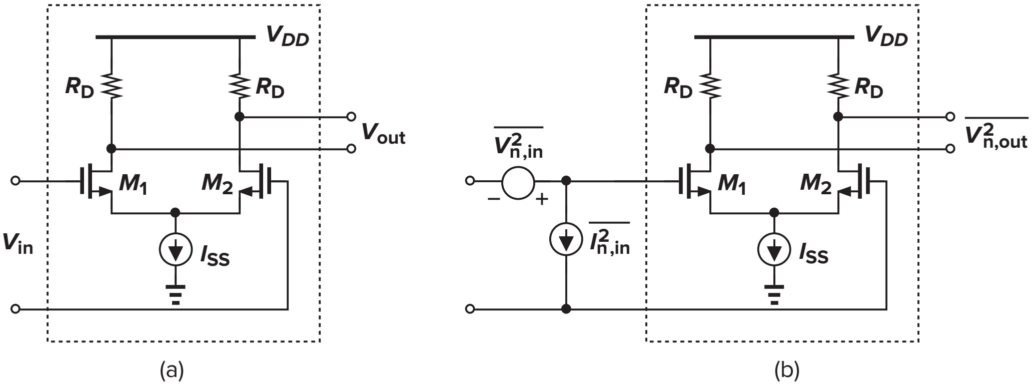


Figure 7.55 (a) Differential pair; (b) circuit including input-referred noise sources.

To calculate the thermal component of $\overline{V_{n,in}^2}$, we first obtain the total output noise with the inputs shorted together [Fig. 7.56(a)], noting that superposition of power quantities is possible because the noise

[12] Also, the transistors' gate leakage currents flow through R_B, introducing a significant dc error if this resistor is very large.

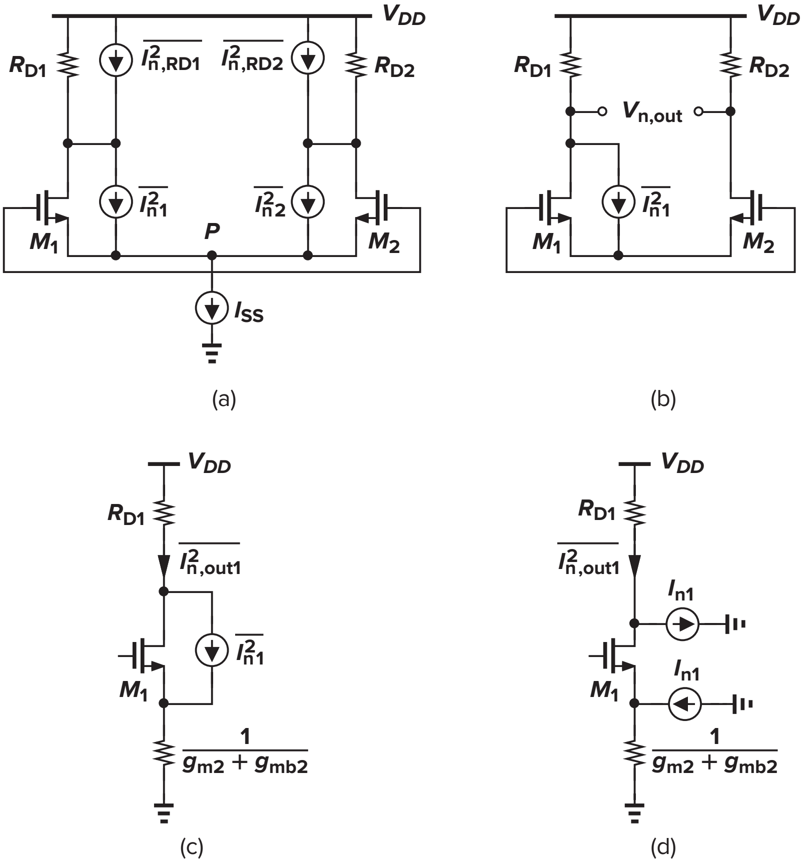


Figure 7.56 Calculation of input-referred noise of a differential pair.

sources in the circuit are uncorrelated. Since I_{n1} and I_{n2} are uncorrelated, node P cannot be considered a virtual ground, making it difficult to use the half-circuit concept. Thus, we simply derive the effect of each source individually. Depicted in Fig. 7.56(b), the contribution of I_{n1} is obtained by first reducing the circuit to that in Fig. 7.56(c). With the aid of this figure and neglecting channel-length modulation, the reader can prove that half of I_{n1} flows through R_{D1} and the other half through M_2 and R_{D2}. [As shown in Fig. 7.56(d), this can also be proved by decomposing I_{n1} into two (correlated) current sources and calculating their effect at the output.] Thus, the differential output noise due to M_1 is equal to

$$V_{n,out}|_{M1} = \frac{I_{n1}}{2} R_{D1} + \frac{I_{n1}}{2} R_{D2} \tag{7.119}$$

Note that the two noise voltages are directly added because they both arise from I_{n1} and are therefore correlated. It follows that, if $R_{D1} = R_{D2} = R_D$,

$$\overline{V^2_{n,out}}|_{M1} = \overline{I^2_{n1}} R^2_D \tag{7.120}$$

Similarly,

$$\overline{V^2_{n,out}}|_{M2} = \overline{I^2_{n2}} R^2_D \tag{7.121}$$

yielding

$$\overline{V_{n,out}^2}\Big|_{M1,M2} = \left(\overline{I_{n1}^2} + \overline{I_{n2}^2}\right) R_D^2 \tag{7.122}$$

Taking into account the noise of R_{D1} and R_{D2}, we have for the total output noise

$$\overline{V_{n,out}^2} = \left(\overline{I_{n1}^2} + \overline{I_{n2}^2}\right) R_D^2 + 2(4kTR_D) \tag{7.123}$$

$$= 8kT\left(\gamma g_m R_D^2 + R_D\right) \tag{7.124}$$

Dividing the result by the square of the differential gain, $g_m^2 R_D^2$, we obtain

$$\overline{V_{n,in}^2} = 8kT\left(\frac{\gamma}{g_m} + \frac{1}{g_m^2 R_D}\right) \tag{7.125}$$

This is simply twice the input noise voltage squared of a common-source stage.

The input-referred noise voltage can also be calculated by exploiting the lemma illustrated in Fig. 7.39. As shown in Fig. 7.57, the noise of M_1 and M_2 is modeled as a voltage source in series with their gates, and the noise of R_{D1} and R_{D2} is divided by $g_m^2 R_D^2$, thereby resulting in (7.125). The reader is encouraged to repeat these calculations if the tail current source is replaced with a short circuit.

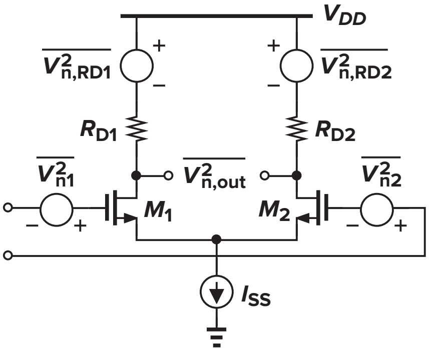


Figure 7.57 Alternative method of calculating the input-referred noise.

It is instructive to compare the noise performance of a differential pair and a common-source stage, as expressed by (7.75) and (7.125). We conclude that, if each transistor has a transconductance g_m, then the input-referred noise *voltage* of a differential pair is $\sqrt{2}$ times that of a common-source stage. This is simply because the former includes twice as many devices in the signal path, as exemplified by the two series voltage sources in Fig. 7.57. (Since the noise sources are uncorrelated, their powers add.) It is also important to recognize that, with the assumption of equal device transconductances, a differential pair consumes twice as much power as a common-source stage if the transistors have the same dimensions.

The noise modeling of Fig. 7.57 can readily account for $1/f$ noise of the transistors as well. Placing the voltage sources given by $K/(C_{ox}WL)$ in series with each gate, we can rewrite (7.125) as

$$\overline{V_{n,in,tot}^2} = 8kT\left(\frac{\gamma}{g_m} + \frac{1}{g_m^2 R_D}\right) + \frac{2K}{C_{ox}WL}\frac{1}{f} \tag{7.126}$$

These derivations suggest that the input-referred noise voltage squared of a fully-differential circuit is equal to twice that of its half-circuit equivalent (because the latter employs half as many devices in the signal path). The following example reinforces this point.

▶ Example 7.21

A differential pair with current-source loads can be configured to act as a large "floating" resistor [7]. Illustrated in Fig. 7.58(a), the idea is to bias M_1 and M_2 at a very small current so as to obtain a high incremental resistance between A and B, approximately equal to $1/g_{m1} + 1/g_{m2}$. Determine the noise associated with this resistor. Neglect channel-length modulation.

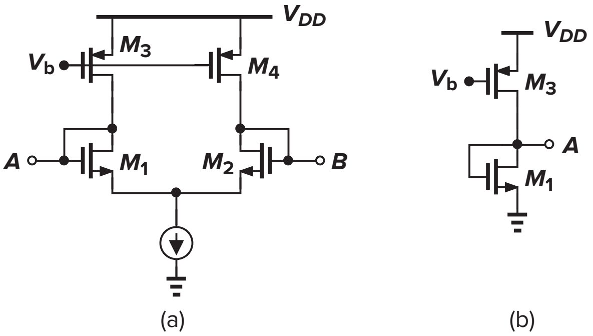


Figure 7.58

Solution

Viewing A and B as the outputs and modeling the circuit by its Thevenin equivalent, we must determine the noise voltage that appears between these nodes. To this end, we construct the half circuit shown in Fig. 7.58(b) and write the noise voltage at A as

$$\overline{V_{n,A}^2} = (4kT\gamma g_{m1} + 4kT\gamma g_{m3})\frac{1}{g_{m1}^2} + \frac{K}{(WL)_1C_{ox}}\frac{1}{f} + \frac{K}{(WL)_3C_{ox}}\frac{1}{f}(\frac{g_{m3}}{g_{m1}})^2 \tag{7.127}$$

The noise measured between A and B is thus equal to

$$\overline{V_{n,AB}^2} = 8kT\gamma(g_{m1} + g_{m3})\frac{1}{g_{m1}^2} + \frac{2K}{(WL)_1C_{ox}}\frac{1}{f} + \frac{2K}{(WL)_3C_{ox}}\frac{1}{f}(\frac{g_{m3}}{g_{m1}})^2 \tag{7.128}$$

We recognize that this resistor is noisier than a simple ohmic resistor of the same value ($\approx 2/g_{m1}$). It is also much less linear (why?).

◀

Does the tail current source in Fig. 7.55 contribute noise? If the differential input signal is zero and the circuit is symmetric, then the noise in I_{SS} divides equally between M_1 and M_2, producing only a common-mode noise voltage at the output. On the other hand, for a small differential input, ΔV_{in}, we have

$$\Delta I_{D1} - \Delta I_{D2} = g_m \Delta V_{in} \tag{7.129}$$

$$= \sqrt{2\mu_n C_{ox}\frac{W}{L}\left(\frac{I_{SS} + I_n}{2}\right)}\Delta V_{in} \tag{7.130}$$

where I_n denotes the noise in I_{SS} and $I_n \ll I_{SS}$. In essence, the noise modulates the transconductance of each device. Equation (7.130) can be written as

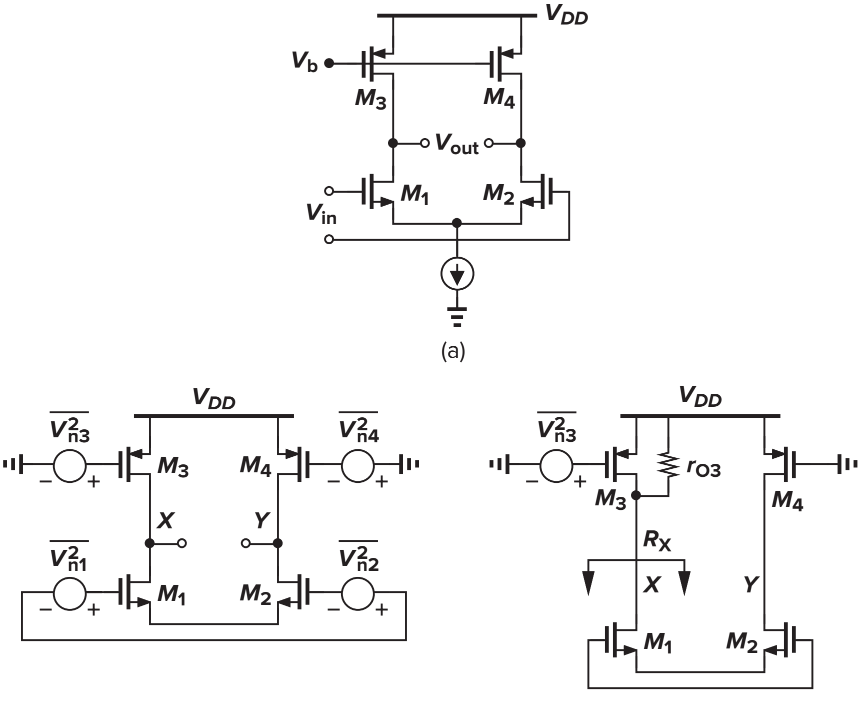
$$\Delta I_{D1} - \Delta I_{D2} \approx \sqrt{2\mu_n C_{ox}\frac{W}{L}\cdot\frac{I_{SS}}{2}}\left(1 + \frac{I_n}{2I_{SS}}\right)\Delta V_{in} \tag{7.131}$$

$$= g_{m0}\left(1 + \frac{I_n}{2I_{SS}}\right)\Delta V_{in} \tag{7.132}$$

where g_{m0} is the transconductance of the noiseless circuit. Equation (7.132) suggests that as the circuit departs from equilibrium, I_n is more unevenly divided between M_1 and M_2, thereby generating differential noise at the output. This effect is nonetheless usually negligible.

▶ Example 7.22

Assuming that the devices in Fig. 7.59(a) operate in saturation and the circuit is symmetric, calculate the input-referred noise voltage.



(a)

(b)

(c)

Figure 7.59

Solution

Since the thermal and $1/f$ noise of M_1 and M_2 can be modeled as voltage sources in series with the input, we need only refer the noise of M_3 and M_4 to the input. Let us calculate the output noise contributed by M_3. The drain noise current of M_3 is divided between r_{O3} and the resistance seen looking into the drain of M_1 [Fig. 7.59(c)]. From Chapter 5, this resistance equals $R_X = r_{O4} + 2r_{O1}$. Denoting the resulting noise currents flowing through r_{O3} and R_X by I_{nA} and I_{nB}, respectively, we have

$$I_{nA} = g_{m3}V_{n3}\frac{r_{O4} + 2r_{O1}}{2r_{O4} + 2r_{O1}} \tag{7.133}$$

and

$$I_{nB} = g_{m3}V_{n3}\frac{r_{O3}}{2r_{O4} + 2r_{O1}} \tag{7.134}$$

The former produces a noise voltage of $g_{m3}V_{n3}r_{O3}(r_{O4} + 2r_{O1})/(2r_{O4} + 2r_{O1})$ at node X with respect to ground whereas the latter flows through M_1, M_2, and r_{O4}, generating $g_{m3}V_{n3}r_{O3}r_{O4}/(2r_{O4} + 2r_{O1})$ at node Y with respect to ground. Thus, the total differential output noise due to M_3 is equal to

$$V_{nXY} = V_{nX} - V_{nY} \tag{7.135}$$

$$= g_{m3} V_{n3} \frac{r_{O3} r_{O1}}{r_{O3} + r_{O1}} \tag{7.136}$$

(The reader can verify that V_{nY} must be *subtracted* from V_{nX}.)

Equation (7.136) implies that the noise current of M_3 is simply multiplied by the parallel combination of r_{O1} and r_{O3} to produce the differential output voltage. This is of course not surprising because, as depicted in Fig. 7.60, the effect of V_{n3} at the output can also be derived by decomposing V_{n3} into two differential components applied to the gates of M_3 and M_4 and subsequently using the half-circuit concept. Since this calculation relates to a *single* noise source, we can temporarily ignore the random nature of noise and treat V_{n3} and the circuit as familiar deterministic, linear components.

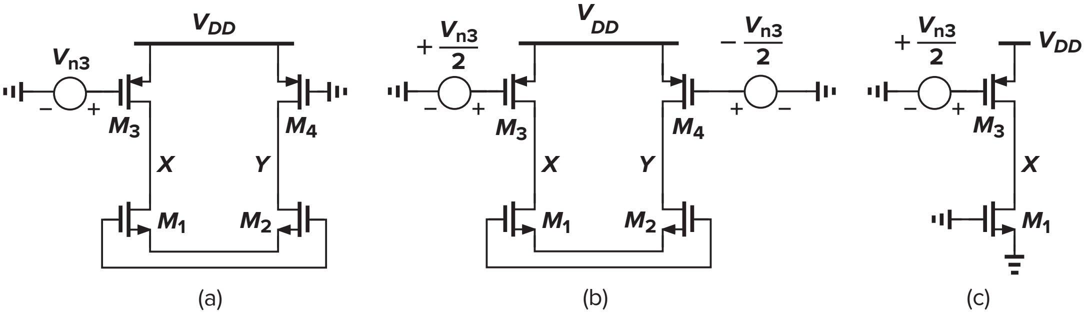


Figure 7.60 Calculation of input-referred noise in a differential pair with current-source loads.

Applying (7.136) to M_4 as well and adding the resulting powers, we have

$$\overline{V_{n,out}^2}|_{M3,M4} = g_{m3}^2 (r_{O1} \| r_{O3})^2 \overline{V_{n3}^2} + g_{m4}^2 (r_{O2} \| r_{O4})^2 \overline{V_{n4}^2} \tag{7.137}$$

$$= 2 g_{m3}^2 (r_{O1} \| r_{O3})^2 \overline{V_{n3}^2} \tag{7.138}$$

To refer the noise to the input, we divide (7.138) by $g_{m1}^2 (r_{O1} \| r_{O3})^2$, obtaining the *total* input-referred noise voltage per unit bandwidth as

$$\overline{V_{n,in}^2} = 2\overline{V_{n1}^2} + 2\frac{g_{m3}^2}{g_{m1}^2} \overline{V_{n3}^2} \tag{7.139}$$

which, upon substitution for $\overline{V_{n1}^2}$ and $\overline{V_{n3}^2}$, reduces to

$$\overline{V_{n,in}^2} = 8kT\gamma \left(\frac{1}{g_{m1}} + \frac{g_{m3}}{g_{m1}^2} \right) + \frac{2K_N}{C_{ox}(WL)_1 f} + \frac{2K_P}{C_{ox}(WL)_3 f} \frac{g_{m3}^2}{g_{m1}^2} \tag{7.140}$$

◀

It is instructive to compare the above input-referred noise to that of a differential pair with active load (the five-transistor OTA). We analyze the thermal noise of the latter and leave the flicker noise as an exercise for the reader. Due to lack of perfect symmetry, we seek the Norton noise equivalent of the circuit by first computing the output short-circuit noise current (Fig. 7.61). This result can be multiplied by the output resistance and divided by the gain to obtain the input-referred noise voltage.

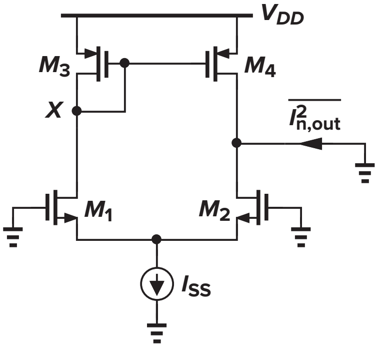


Figure 7.61 OTA output short-circuit noise current.

We recall from Chapter 5 that the transconductance of the five-transistor OTA is approximately equal to $g_{m1,2}$. Thus, the output noise current due to M_1 and M_2 is given by this transconductance multiplied by the gate-referred noises of M_1 and M_2, i.e., $g^2_{m1,2}(4kT\gamma/g_{m1} + 4kT\gamma/g_{m2})$.

Let us consider the noise current of M_3, $4kT\gamma g_{m3}$. This current primarily circulates through the diode-connected impedance, $1/g_{m3}$, producing a voltage at the gate of M_4 with a spectral density of $4kT\gamma/g_{m3}$. This noise is multiplied by g^2_{m4} as it emerges from the drain of M_4. The noise current of M_4 itself also directly flows through the output short circuit. We therefore have

$$\overline{I^2_{n,out}} = 4kT\gamma(2g_{m1,2} + 2g_{m3,4}) \tag{7.141}$$

Multiplying this noise by $R^2_{out} \approx (r_{O1,2}||r_{O3,4})^2$ and dividing the result by $A^2_v = G^2_m R^2_{out}$, we obtain the total input-referred noise voltage as

$$\overline{V^2_{n,in}} = 8kT\gamma\left(\frac{1}{g_{m1,2}} + \frac{g_{m3,4}}{g^2_{m1,2}}\right) \tag{7.142}$$

which is the same as that of the fully-differential circuit.

An interesting difference between the OTA and the fully-differential topology relates to the noise contributed by the tail current when $V_{in1} = V_{in2}$. Recall from Chapter 5 that the output voltage of the OTA in Fig. 7.62 is equal to V_X. If I_{SS} fluctuates, so do V_X and V_{out}. Since the tail noise current, I_n, splits equally between M_1 and M_2, the noise voltage at X is given by $\overline{I^2_n}/(4g^2_{m3})$, and so is the noise voltage at the output. (Why does I_n split equally, even though the impedance seen looking into the source of M_2 appears to be higher than that seen looking into the source of M_1?)

The effect of noise must be studied for many other analog circuits as well. For example, feedback systems, op amps, and bandgap references exhibit interesting and important noise characteristics. We return to these topics in other chapters.

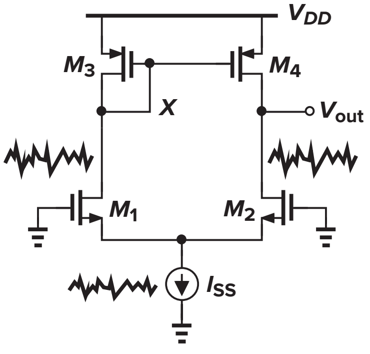


Figure 7.62 Effect of tail noise current in OTA.

7.7 ■ Noise-Power Trade-Off

In our analysis of the input-referred thermal noise, we have seen that the noise contributed by the transistors "in the signal path" is inversely proportional to their transconductance. This dependence suggests a trade-off between noise and power consumption.

The noise-power trade-off can in fact be generalized to *any* circuit (so long as the input noise current is negligible). To understand this point, let us begin with a simple CS stage as shown in Fig. 7.63(a): we double W/L and the bias current of M_1 and halve the load resistor. This transformation maintains the voltage gain and the output swing regardless of the transistor characteristics. But we note that the input-referred thermal and flicker noise power is exactly halved (because both the g_m and the gate area of the transistor are doubled). This 3-dB reduction in noise accrues at a cost of doubling the power consumption (and the input capacitance).

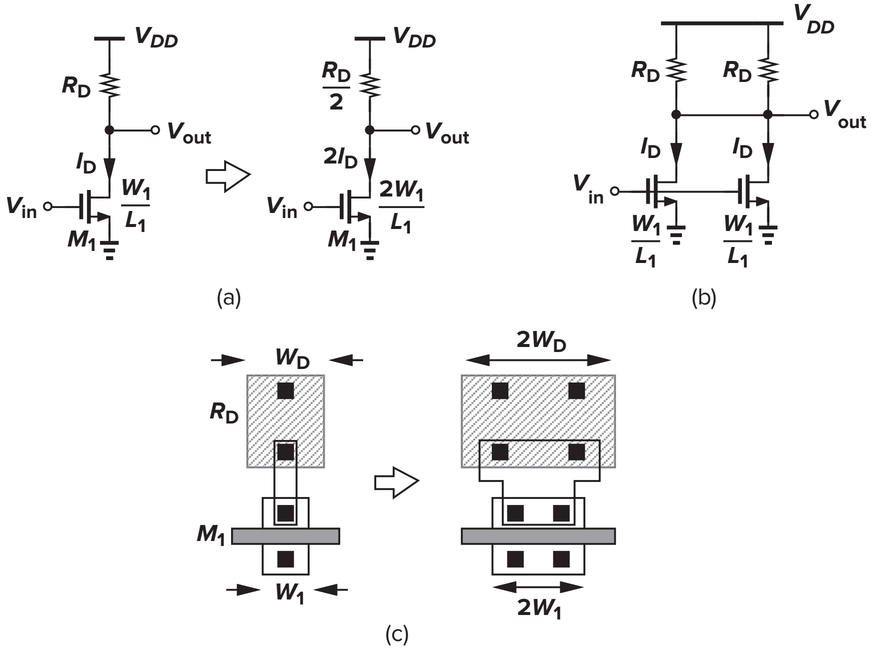


Figure 7.63 (a) Output noise reduction by scaling, (b) equivalent operation, and (c) scaling viewed at layout level.

Called "linear scaling," the transformation depicted in Fig. 7.63(a) can also be viewed as placing two instances of the original circuit *in parallel*, as illustrated in Fig. 7.63(b). Alternatively, we can say that the widths of the transistor and the resistor are doubled [Fig. 7.63(c)].

In general, if two instances of a circuit are placed in parallel, the output noise power is halved [Fig. 7.64(a)]. This can be proved by setting the input to zero and constructing a Thevenin noise equivalent for each [Fig. 7.64(b)]. Since $V_{n1,out}$ and $V_{n2,out}$ are uncorrelated, we can use superposition of powers to write

$$\overline{V^2_{n,out}} = \frac{\overline{V^2_{n1,out}}}{4} + \frac{\overline{V^2_{n2,out}}}{4} \tag{7.143}$$

$$= \frac{\overline{V^2_{n1,out}}}{2} \tag{7.144}$$

Thus, the output noise is traded for power consumption while retaining the voltage gain and output swings. Note that this can also be proved if the input is left open, revealing that the input-referred noise current, $\overline{I^2_{n,in}}$, is *doubled* (why?).

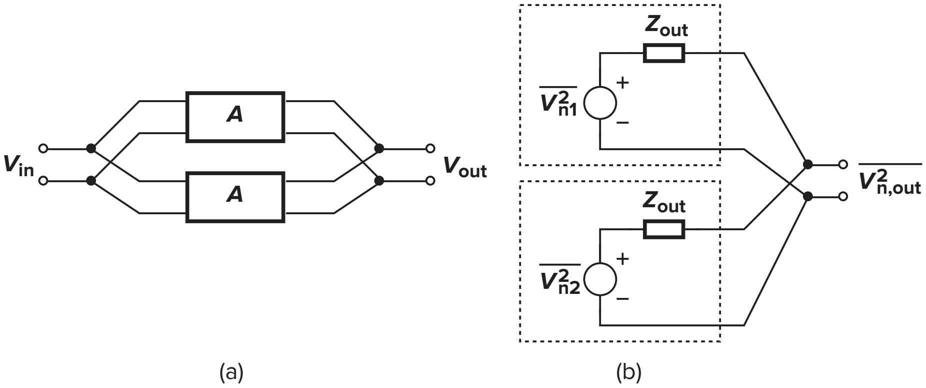


Figure 7.64 (a) General scaling for noise reduction, and (b) equivalent circuit.

We should also remark that noise spectrum must eventually be integrated across the circuit's bandwidth. The foregoing linear scaling assumes that the bandwidth is dictated by the application and hence constant.

7.8 ■ Noise Bandwidth

The total noise corrupting a signal in a circuit results from all of the frequency components that fall in the bandwidth of the circuit. Consider a multipole circuit having the output noise spectrum shown in Fig. 7.65(a). Since the noise components above ω_{p1} are not negligible, the total output noise must be evaluated by calculating the total area under the spectral density:

$$\overline{V_{n,out,tot}^2} = \int_0^\infty \overline{V_{n,out}^2} df \tag{7.145}$$

However, as depicted in Fig. 7.65(b), it is sometimes helpful to represent the total noise simply as $V_0^2 \cdot B_n$, where the bandwidth B_n is chosen such that

$$V_0^2 \cdot B_n = \int_0^\infty \overline{V_{n,out}^2} df \tag{7.146}$$

Called the "noise bandwidth," B_n allows a fair comparison of circuits that exhibit the same low-frequency noise, V_0^2, but different high-frequency transfer functions. As an exercise, the reader can prove that the noise bandwidth of a one-pole system is equal to $\pi/2$ times the pole frequency.

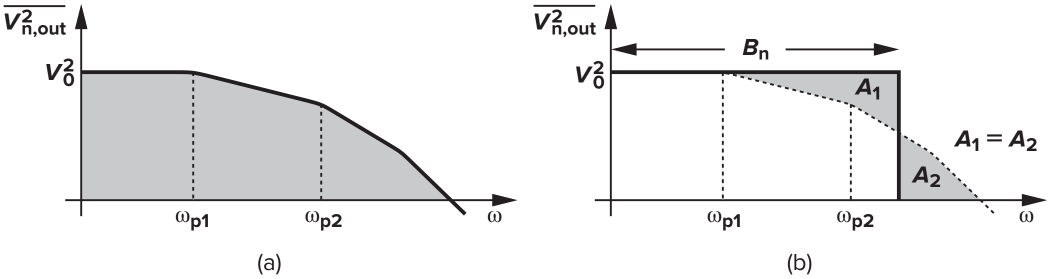


Figure 7.65 (a) Output noise spectrum of a circuit; (b) concept of noise bandwidth.

7.9 ■ Problem of Input Noise Integration

In our noise studies thus far, we have computed the output noise spectrum and, by integration, the total output noise voltage. Is it possible to perform the integration on the input-referred noise instead?

Consider the CS stage shown in Fig. 7.66, where we assume that $\lambda = 0$ and M_1 exhibits only thermal noise. For simplicity, let us neglect the noise of R_D. We note that the output noise spectrum is equal to the amplified and low-pass filtered noise of M_1; this spectrum readily lends itself to integration (Example 7.19). The input-referred noise voltage, on the other hand, is simply equal to $\overline{V_{n,M1}^2}$, carrying an *infinite* power and prohibiting integration at the input.

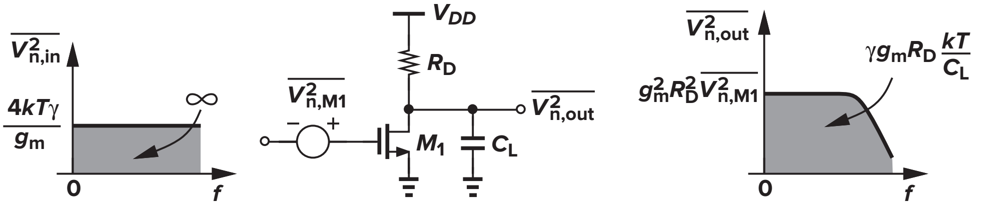


Figure 7.66 Difficulty with referring output noise to input.

The above quandary arises in most circuits, encouraging only *output* noise integration. After all, the physical and observable noise appears only at the output, and the input-referred noise remains a fictitious quantity. However, for a fair comparison of different designs, we can divide the *integrated* output noise by the low-frequency (or mid-band) gain of the circuit. For example, the CS stage of Fig. 7.66 can be characterized by a total input-referred noise equal to

$$\overline{V_{n,in,tot}^2} = \gamma g_m R_D \frac{kT}{C_L} \cdot \frac{1}{g_m^2 R_D^2} \tag{7.147}$$

$$= \frac{\gamma}{g_m R_D} \frac{kT}{C_L} \tag{7.148}$$

if the noise of R_D is neglected. The reader is encouraged to repeat these calculations with channel-length modulation and the noise of R_D included.

7.10 ■ Appendix A: Problem of Noise Correlation

As explained in Section 7.1.3, the input-referred noise voltage and current are generally correlated, complicating noise calculations. In this appendix, we consider alternative methods that avoid the correlation. Recall from (7.55) that the input-referred noise current manifests itself only if the magnitude squared of the impedance driving the circuit is comparable to $\overline{V_{n,in}^2}/\overline{I_{n,in}^2}$.

In many circuits, the output noise voltage remains approximately the same as the driving impedance, Z_S, goes from zero to infinity, i.e., the input port termination goes from a short circuit to an open circuit.[13] For example, a common-source stage with negligible C_{GD} exhibits this behavior [Fig. 7.67(a)]:

$$\overline{V_{n1,out}^2} = \overline{V_{n2,out}^2} = 4kT\gamma g_m R_D^2 + 4kTR_D \tag{7.149}$$

[13] The noise of Z_S is excluded here.

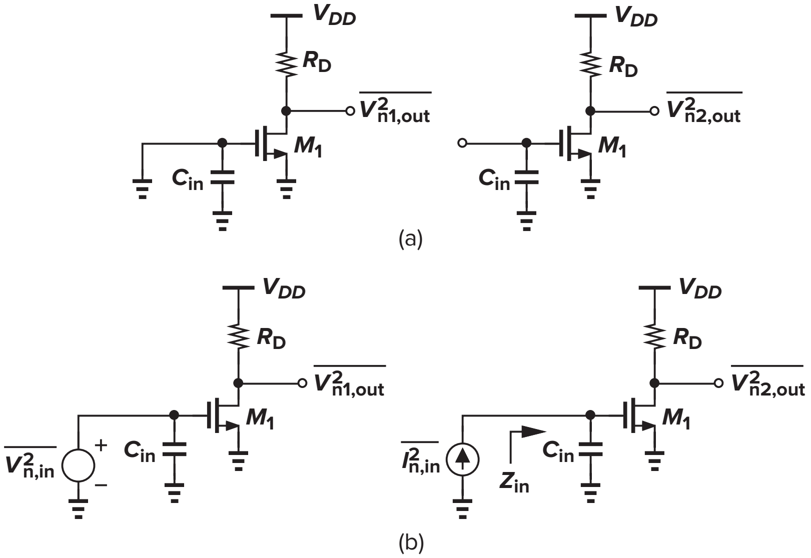


Figure 7.67 (a) Output noise of CS stage with input shorted or open; (b) calculation of input-referred sources.

We now note from Fig. 7.67(b) that

$$\overline{V^2_{n1,out}} = \overline{V^2_{n,in}}|H(f)|^2 \tag{7.150}$$

where $H(s) = V_{out}/V_{in}$, and also

$$\overline{V^2_{n2,out}} = \overline{I^2_{n,in}}|Z_{in}(f)|^2|H(f)|^2 \tag{7.151}$$

It follows that $\overline{I^2_{n,in}} = \overline{V^2_{n,in}}/|Z_{in}(f)|^2$. Since $Z_{in}(s)$ is a deterministic quantity, we have $I_{n,in} = V_{n,in}/Z_{in}(s)$, and hence 100% correlation between the two sources. In order to account for both $V_{n,in}$ and $I_{n,in}$, we must carry out lengthy calculations similar to those for Fig. 7.37.

Now, consider the arrangement shown in Fig. 7.68(a), where Z_S denotes the output impedance of the preceding stage. We assume that the output noise of the circuit negligibly changes as Z_S varies. The noise voltage at node X is equal to

$$V_{n,X} = \frac{Z_{in}}{Z_{in}+Z_S}V_{n,in} + \frac{Z_{in}Z_s}{Z_{in}+Z_S}I_{n,in} \tag{7.152}$$

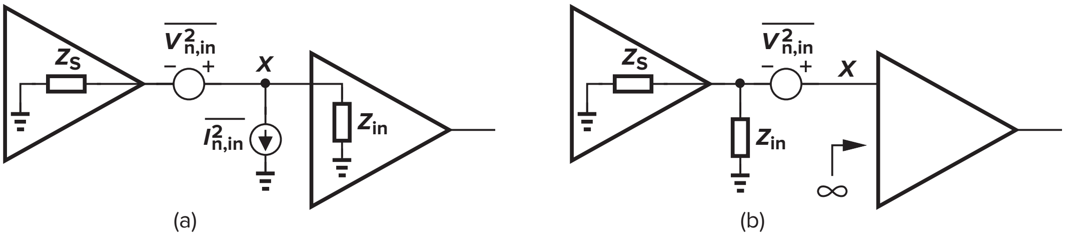


Figure 7.68 (a) Cascade of two stages, and (b) transformation to omit $I_{n,in}$.

Replacing for $I_{n,in}$ from above, we obtain

$$V_{n,X} = V_{n,in} \tag{7.153}$$

That is, $I_{n,in}$ simply serves to keep $V_{n,X}$ (with respect to ground) equal to $V_{n,in}$ for different values of Z_S. This interesting result helps simplify the analysis.

Based on this observation, we modify the arrangement to that in Fig. 7.68(b), where Z_{in} simply loads the preceding stage but $I_{n,in}$ is absent. Here, too, we have $V_{n,X} = V_{n,in}$. Thus, in circuits whose output noise voltage is a weak function of the input termination, $I_{n,in}$ can be omitted if an impedance equal to Z_{in} is used to load the preceding stage.

If the above condition for $V_{n,out}$ does not hold, we may simply consider the preceding stage as part of the circuit and view the two stages as one entity. For example, the amplifier shown in Fig. 7.69 can be modeled as one stage with input-referred noise sources $V_{n,in}$ and $I_{n,in}$, thereby avoiding the complications associated with the second stage's noise voltage and current.

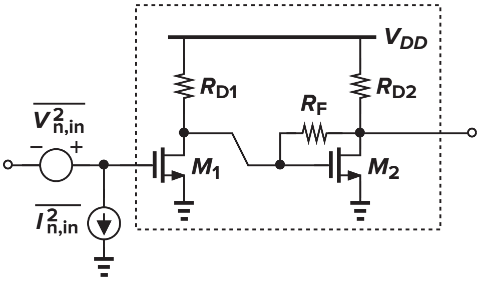


Figure 7.69 Viewing a cascade as a single circuit.

References

[1] L. W. Couch, *Digital and Analog Communication Systems,* 4th ed. (New York: Macmillan Co., 1993).

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

[3] B. Razavi, Y. Ran, and K. F. Lee, "Impact of Distributed Gate Resistance on the Performance of MOS Devices," *IEEE Trans. Circuits and Systems, Part I,* pp. 750–754, Nov. 1994.

[4] Y. Tsividis, *Operation and Modeling of the MOS Transistor,* 2nd ed. (Boston: McGraw-Hill, 1999).

[5] A. A. Abidi, "High-Frequency Noise Measurements on FETs with Small Dimensions," *IEEE Trans. Electron Devices,* vol. 33, pp. 1801–1805, Nov. 1986.

[6] H. A. Haus et al., "Representation of Noise in Linear Twoports," *Proc. IRE*, vol. 48, pp. 69–74, Jan. 1960.

[7] S. Asai et al, "High-Resistance Resistor Consisting of a Subthreshold CMOS Differential Pair," *IEICE Trans. Electronics,* vol. E93, pp. 741–746, June 2010.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

7.1. A common-source stage incorporates a 50-μm/0.5-μm NMOS device biased at $I_D = 1$ mA along with a load resistor of 2 kΩ. What is the total input-referred thermal noise voltage in a 100-MHz bandwidth?

7.2. Consider the common-source stage of Fig. 7.42. Assume that $(W/L)_1 = 50/0.5$, $I_{D1} = I_{D2} = 0.1$ mA, and $V_{DD} = 3$ V. If the contribution of M_2 to the input-referred noise voltage (not voltage squared) must be one-fifth of that of M_1, what is the maximum output voltage swing of the amplifier?

7.3. Using the distributed model of Fig. 7.21(c) and ignoring the channel thermal noise, prove that, for gate noise calculations, a distributed gate resistance of R_G can be replaced by a lumped resistance equal to $R_G/3$. (Hint: model the noise of R_{Gj} by a series voltage source and calculate the total drain noise current. Watch for correlated sources of noise.)

7.4. Prove that the output noise current of Fig. 7.39(c) is given by Eq. (7.73).

7.5. Calculate the input-referred flicker noise voltage of the circuit shown in Fig. 7.70.

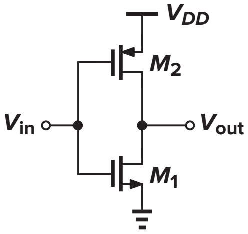


Figure 7.70

7.6. Calculate the input-referred thermal noise voltage of each circuit in Fig. 7.71. Assume that $\lambda = \gamma = 0$.

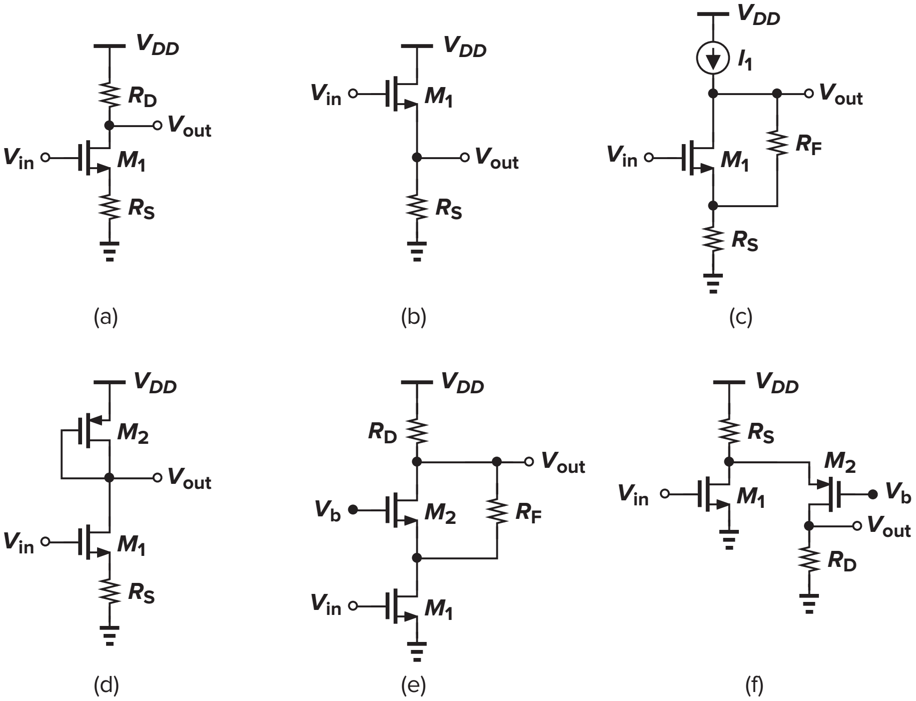


Figure 7.71

7.7. Calculate the input-referred thermal noise voltage of each circuit in Fig. 7.72. Assume that $\lambda = \gamma = 0$.

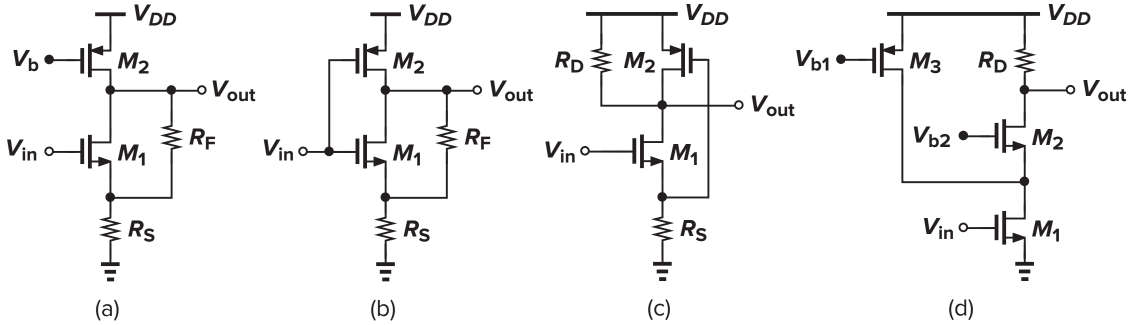


Figure 7.72

7.8. Calculate the input-referred thermal noise voltage and current of each circuit in Fig. 7.73. Assume that $\lambda = \gamma = 0$.

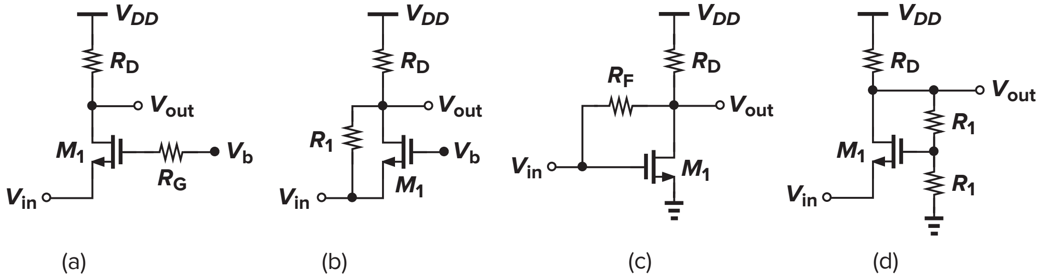


Figure 7.73

7.9. Calculate the input-referred thermal noise voltage and current of each circuit in Fig. 7.74. Assume that $\lambda = \gamma = 0$.

7.10. Calculate the input-referred $1/f$ noise voltage and current of Fig. 7.49 if the two capacitors are removed.

7.11. Calculate the input-referred $1/f$ noise voltage of the source follower shown in Fig. 7.51.

7.12. Assuming that $\lambda = \gamma = 0$, calculate the input-referred thermal noise voltage of each circuit in Fig. 7.75. For part (a), assume that $g_{m3,4} = 0.5g_{m5,6}$.

7.13. Consider the degenerated common-source stage shown in Fig. 7.76.

(a) Calculate the input-referred thermal noise voltage if $\lambda = \gamma = 0$.

(b) Suppose linearity requirements necessitate that the dc voltage drop across R_S be equal to the overdrive voltage of M_1. How does the thermal noise contributed by R_S compare with that contributed by M_1?

7.14. Explain why Miller's theorem cannot be applied to calculate the effect of the thermal noise of a floating resistor.

7.15. The circuit of Fig. 7.20 is designed with $(W/L)_1 = 50/0.5$ and $I_{D1} = 0.05$ mA. Calculate the total rms thermal noise voltage at the output in a 50-MHz bandwidth.

7.16. For the circuit shown in Fig. 7.77, calculate the total output thermal and $1/f$ noise in a bandwidth $[f_L, f_H]$. Assume that $\lambda \neq 0$, but neglect other capacitances.

7.17. Suppose in the circuit of Fig. 7.42, $(W/L)_{1,2} = 50/0.5$ and $I_{D1} = |I_{D2}| = 0.5$ mA. What is the input-referred thermal noise voltage?

7.18. The circuit of Fig. 7.42 is modified as depicted in Fig. 7.78.

(a) Calculate the input-referred thermal noise voltage.

(b) For a given bias current and output voltage swing, what value of R_S minimizes the input-referred thermal noise?

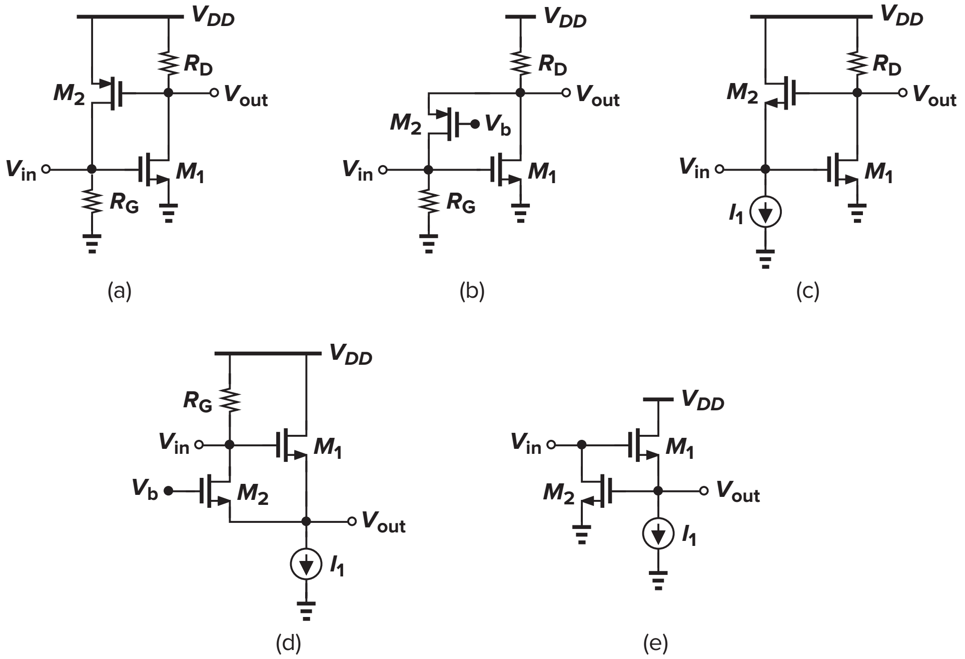


Figure 7.74

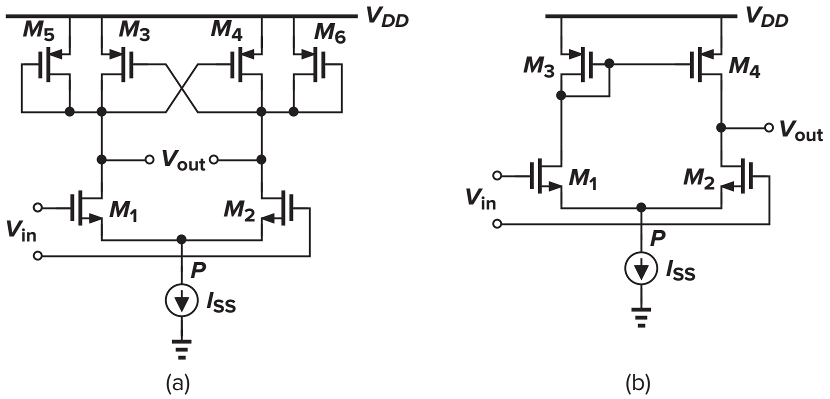


Figure 7.75

7.19. A common-gate stage incorporates an NMOS device with $W/L = 50/0.5$ biased at $I_D = 1$ mA and a load resistor of 1 kΩ. Calculate the input-referred thermal noise voltage and current.

7.20. The circuit of Fig. 7.48 is designed with $(W/L)_1 = 50/0.5$, $I_{D1} = I_{D2} = 0.5$ mA, and $R_D = 1$ kΩ.

(a) Determine $(W/L)_2$ such that the contribution of M_2 to the input-referred thermal noise current (not current squared) is one-fifth of that due to R_D.

(b) Now calculate the minimum value of V_b to place M_2 at the edge of the triode region. What is the maximum allowable output voltage swing?

7.21. Design the circuit of Fig. 7.48 for an input-referred thermal noise voltage of 3 nV/$\sqrt{\text{Hz}}$ and maximum output swing. Assume that $I_{D1} = I_{D2} = 0.5$ mA.

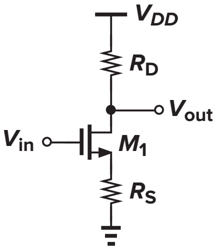


Figure 7.76

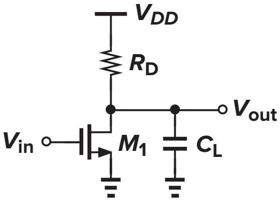


Figure 7.77

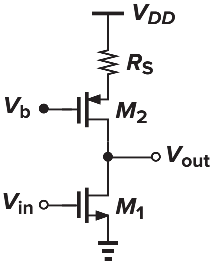


Figure 7.78

7.22. Consider the circuit of Fig. 7.49. If $(W/L)_{1-3} = 50/0.5$ and $I_{D1-3} = 0.5$ mA, determine the input-referred thermal noise voltage and current.

7.23. The circuit of Fig. 7.49 is designed with $(W/L)_1 = 50/0.5$ and $I_{D1-3} = 0.5$ mA. If an output swing of 2 V is required, estimate by iteration the dimensions of M_2 and M_3 such that the input-referred thermal noise current is minimum.

7.24. The source follower of Fig. 7.51 is to provide an output resistance of 100 Ω with a bias current of 0.1 mA.

(a) Calculate $(W/L)_1$.

(b) Determine $(W/L)_2$ such that the input-referred thermal noise voltage (not voltage squared) contributed by M_2 is one-fifth of that due to M_1. What is the maximum output swing?

7.25. The cascode stage of Fig. 7.52(a) exhibits a capacitance C_X from node X to ground. Neglecting other capacitances, determine the input-referred thermal noise voltage.

7.26. Determine the input-referred thermal and $1/f$ noise voltages of the circuits shown in Fig. 7.79 and compare the results. Assume that the circuits draw equal supply currents.

7.27. Repeat the analysis in Example 7.13 but assume that $\lambda > 0$.

7.28. Suppose the circuit of Fig. 7.38(a) is driven by a finite source impedance, as shown in Fig. 7.80. Assume that $\lambda = 0$, and neglect the noise of R_S.

(a) Determine the output noise voltage of the circuit.

(b) In a manner similar to the analysis of Fig. 7.37, compute in terms of $V_{n,RF}$ and $V_{n,M1}$ the input-referred noise voltage and current, paying close attention to their correlation.

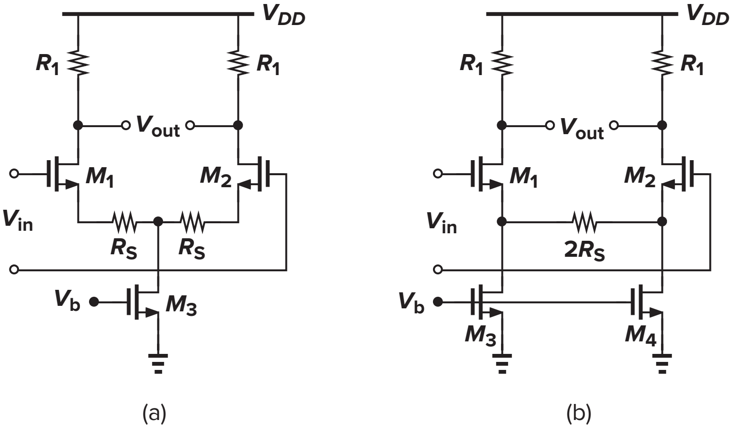


Figure 7.79

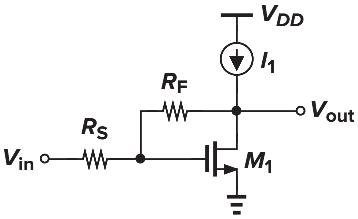


Figure 7.80

(c) Using superposition of voltages and currents (not powers), calculate the output noise voltage in terms of $V_{n,in}$ and $I_{n,in}$, as obtained in (b). Now make the substitutions $\overline{V_{n,RF}^2} = 4kTR_F$ and $\overline{I_{n,M1}^2} = 4kT\gamma g_m$. Is this result the same as that derived in (a)?

7.29. Consider the circuits in Figs. 7.39(c) and (d), but include C_{GS} and a noiseless impedance Z_1 in series with the gate. Derive expressions for $I_{n,out1}$ and $I_{n,out2}$. Does the lemma hold in this case?

7.30. Repeat Example 7.14 while including C_{GS} and an impedance Z_1 in series with the gate. Does the lemma hold in this case?

7.31. Model the thermal noise of M_1 in Fig. 7.49 by a voltage source in series with its gate and assuming the input is open,

(a) Determine the resulting output voltage. (The voltage gain for a degenerated CS stage was derived in Chapter 3.)

(b) Now refer this voltage to the input as a current and compare the result with the contributions of M_2 and M_3.

7.32. Figure 7.81 shows a noiseless amplifier driven by a source resistance of R_S. If the amplifier can be modeled by a low-frequency gain of A_0 and a single pole at ω_0, determine the total integrated noise at the output due to R_S.

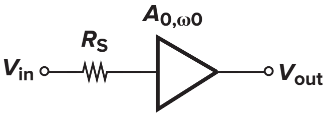


Figure 7.81

7.33. Considering only thermal noise in Fig. 7.82, determine the output noise spectrum and the total integrated noise. Assume that $\lambda > 0$.

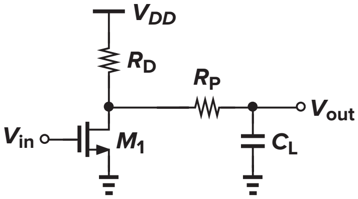


Figure 7.82

7.34. Calculate the input-referred thermal and flicker noise of the circuit shown in Fig. 7.83, where the output of interest is $I_{D3} - I_{D4}$. Consider two cases: (a) the current sources are ideal, and (b) the current sources are realized by MOSFETs. Neglect channel-length modulation and body effect.

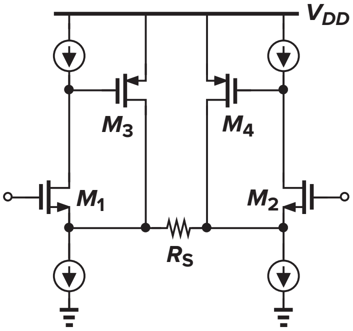


Figure 7.83

CHAPTER

8

Feedback

On a mild August morning in 1927, Harold Black was riding the ferry from New York to New Jersey, where he worked at Bell Laboratories. Black and many other researchers had been investigating the problem of nonlinearity in amplifiers used in long-distance telephone networks, seeking a practical solution. While reading the newspaper on the ferry, Black was suddenly struck by an idea and began to draw a diagram on the newspaper, which would later be used as the evidence in his patent application. The idea is known to us as the negative-feedback amplifier.

Feedback is a powerful technique that finds wide application in analog circuits. For example, negative feedback allows high-precision signal processing, and positive feedback makes it possible to build oscillators. In this chapter, we consider only negative feedback and use the term feedback to mean that.

We begin with a general view of feedback circuits, describing important benefits that result from feedback. Next, we study four feedback topologies and their properties. We then deal with difficulties in feedback circuit analysis and introduce the two-port technique, Bode's technique, and Blackman's theorem as possible solutions.

8.1 ■ General Considerations

Figure 8.1 shows a negative-feedback system, where $H(s)$ and $G(s)$ are called the feedforward and the feedback networks, respectively. Since the output of $G(s)$ is equal to $G(s)Y(s)$, the input to $H(s)$, called the feedback error, is given by $X(s) - G(s)Y(s)$. That is

$$Y(s) = H(s)[X(s) - G(s)Y(s)] \tag{8.1}$$

Thus,

$$\frac{Y(s)}{X(s)} = \frac{H(s)}{1 + G(s)H(s)} \tag{8.2}$$

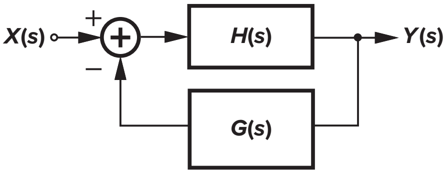


Figure 8.1 General feedback system.

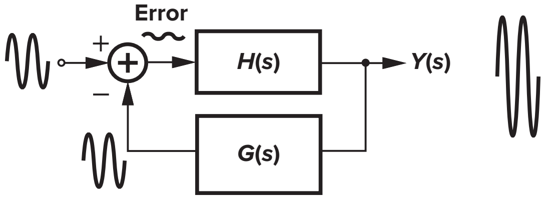


Figure 8.2 Similarity between output of feedback network and input signal.

We call $H(s)$ the "open-loop" transfer function and $Y(s)/X(s)$ the "closed-loop" transfer function. In most cases of interest in this book, $H(s)$ represents an amplifier and $G(s)$ is a frequency-independent quantity. In other words, a fraction of the output signal is sensed and compared with the input, generating an error term. In a well-designed negative-feedback system, the error term is minimized, thereby making the output of $G(s)$ an accurate "copy" of the input and hence the output of the system a faithful (scaled) replica of the input (Fig. 8.2). We also say that the input of $H(s)$ is a "virtual ground" because the signal amplitude at this point is small. In subsequent developments, we replace $G(s)$ by a frequency-independent quantity β and call it the "feedback factor."

It is instructive to identify four elements in the feedback system of Fig. 8.1: (1) the feedforward amplifier, (2) a means of sensing the output, (3) the feedback network, and (4) a means of generating the feedback error, i.e., a subtractor (or an adder). These elements exist in every feedback system, even though they may not be obvious in cases such as a simple common-source stage with resistive degeneration.

8.1.1 Properties of Feedback Circuits

Before proceeding to the analysis of feedback circuits, we study some simple examples to describe the benefits of negative feedback.

Gain Desensitization Consider the common-source stage shown in Fig. 8.3(a), where the voltage gain is equal to $g_{m1}r_{O1}$. A critical drawback of this circuit is the poor definition of the gain: both g_{m1} and r_{O1} vary with process and temperature. Now suppose the circuit is configured as in Fig. 8.3(b), where the gate bias of M_1 is set by means not shown here (Chapter 13). Let us calculate the overall voltage gain of the circuit at relatively low frequencies such that C_2 draws a negligible (small-signal) current from the output node, i.e., $V_{out}/V_X = -g_{m1}r_{O1}$ because the entire drain current flows through r_{O1}. Since $(V_{out} - V_X)C_2s = (V_X - V_{in})C_1s$, we have

$$\frac{V_{out}}{V_{in}} = -\frac{1}{\left(1 + \frac{1}{g_{m1}r_{O1}}\right)\frac{C_2}{C_1} + \frac{1}{g_{m1}r_{O1}}} \tag{8.3}$$

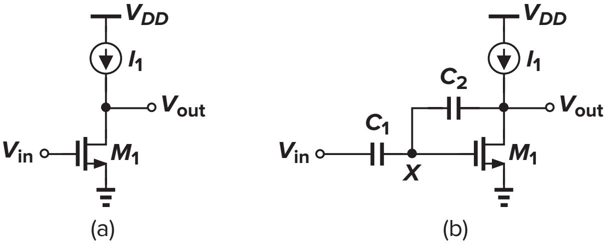


Figure 8.3 (a) Simple common-source stage; (b) circuit of (a) with feedback.

If $g_{m1}r_{O1}$ is sufficiently large, the $1/(g_{m1}r_{O1})$ terms in the denominator can be neglected, yielding

$$\frac{V_{out}}{V_{in}} = -\frac{C_1}{C_2} \tag{8.4}$$

Compared to $g_{m1}r_{O1}$, this gain can be controlled with much higher accuracy because it is given by the *ratio* of two capacitors. If C_1 and C_2 are made of the same material, then process and temperature variations do not change C_1/C_2.

The above example reveals that negative feedback provides gain "desensitization," i.e., the closed-loop gain is less sensitive to device parameters than the open-loop gain is. One may also say that negative feedback "stabilizes" the gain and hence "improves the stability." But this nomenclature may be confused with frequency stability (Chapter 10), which typically *worsens* as a result of negative feedback. Illustrated for a more general case in Fig. 8.4, gain desensitization can be quantified by writing

$$\frac{Y}{X} = \frac{A}{1+\beta A} \tag{8.5}$$

$$\approx \frac{1}{\beta}\left(1-\frac{1}{\beta A}\right) \tag{8.6}$$

where we have assumed that $\beta A \gg 1$. We note that the closed-loop gain is determined, to the first order by the feedback factor, β. More important, even if the open-loop gain, A, varies by a factor of, say, 2, Y/X varies by a small percentage because $1/(\beta A) \ll 1$.

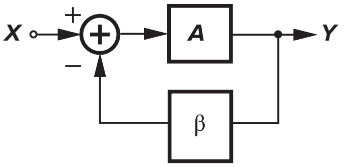


Figure 8.4 Simple feedback system.

Called the "loop gain," the quantity βA plays an important role in feedback systems.[1] We see from (8.6) that the higher βA is, the less sensitive Y/X will be to variations in A. From another perspective, the accuracy of the closed-loop gain improves by maximizing βA. Note that as β increases, the closed-loop gain, $Y/X \approx 1/\beta$, decreases, suggesting a trade-off between precision and the closed-loop gain. In other words, we begin with a high-gain amplifier and apply feedback to obtain a low, but less sensitive, closed-loop gain. Another conclusion here is that the output of the feedback network is equal to $\beta Y = X \cdot \beta A/(1+\beta A)$, approaching X as βA becomes much greater than unity. This result agrees with the illustration in Fig. 8.2.

The calculation of the loop gain can proceed as follows. As illustrated in Fig. 8.5, we set the main input to (ac) zero, break the loop at some point, inject a test signal in the "right direction," follow the signal around the loop, and obtain the value that returns to the break point. The negative of the transfer function thus derived is the loop gain. Note that the loop gain is a dimensionless quantity. In Fig. 8.5, we have $V_t\beta(-1)A = V_F$ and hence $V_F/V_t = -\beta A$. Similarly, as depicted in Fig. 8.6, for the simple feedback circuit, we can write $V_X = V_tC_2/(C_1+C_2)$ and[2]

$$V_t\frac{C_2}{C_1+C_2}(-g_{m1}r_{O1}) = V_F \tag{8.7}$$

[1] The loop gain, βA, and the open-loop gain, A, must not be confused with each other.

[2] A common mistake here is to say that C_2 does not pass signals at very low frequencies, and hence $V_X = 0$. This is not true because C_1 also has a high impedance at very low frequencies.

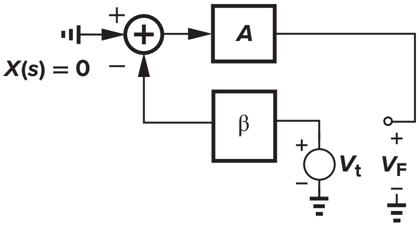


Figure 8.5 Computation of loop gain.

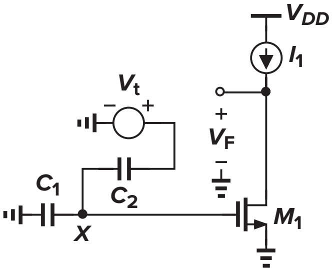


Figure 8.6 Computation of loop gain in a simple feedback circuit.

That is

$$\frac{V_F}{V_t} = -\frac{C_2}{C_1 + C_2} g_{m1} r_{O1} \tag{8.8}$$

Note that the current drawn by C_2 from the output is neglected here. This issue will be addressed in Sec. 8.5.

▶ Example 8.1

Determine the loop gain for the feedback common-gate stage shown in Fig. 8.7(a).

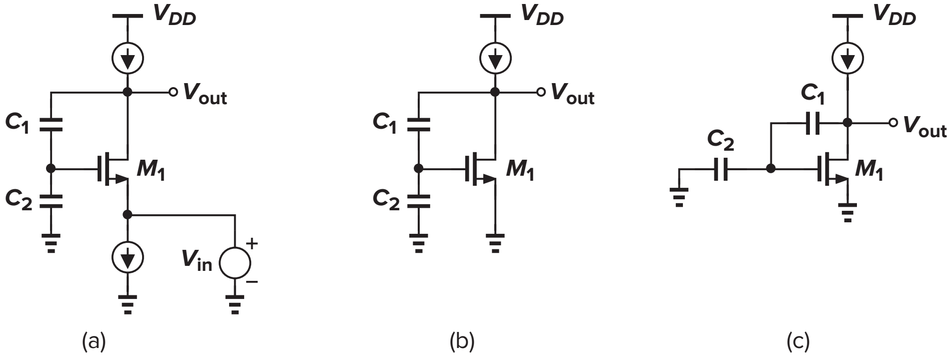


Figure 8.7

Solution

In order to compute the loop gain, we must first set the main input to (ac) zero, arriving at the arrangement shown in Fig. 8.7(b). Redrawing the circuit as in Fig. 8.7(c), we recognize that this topology is identical to the CS stage of Fig. 8.3(b) with $V_{in} = 0$. The loop gain is therefore given by Eq. (8.8).

The important point here is that, when computing the loop gain, we no longer know where the main input and output terminals are. Thus, seemingly different circuit topologies may have the same loop gain. ◀

We should emphasize that the desensitization of gain by feedback leads to many other properties of feedback systems. Our examination of Eq. (8.6) indicates that large variations in A affect Y/X negligibly if βA is large. Such variations can arise from different sources: process, temperature, frequency, and loading. For example, if A drops at high frequencies, Y/X varies to a lesser extent, and the bandwidth is increased. Similarly, if A decreases because the amplifier drives a heavy load, Y/X is not affected much. These concepts become clearer below.

Terminal Impedance Modification As a second example, let us study the circuit shown in Fig. 8.8(a), where a capacitive voltage divider senses the output voltage of a common-gate stage, applying the result to the gate of current source M_2 and hence returning a signal to the input.[3] Our objective is to compute the input resistance at relatively low frequencies with and without feedback. Neglecting channel-length modulation and the current drawn by C_1, we break the feedback loop as shown in Fig. 8.8(b) and write

$$R_{in,open} = \frac{1}{g_{m1} + g_{mb1}} \tag{8.9}$$

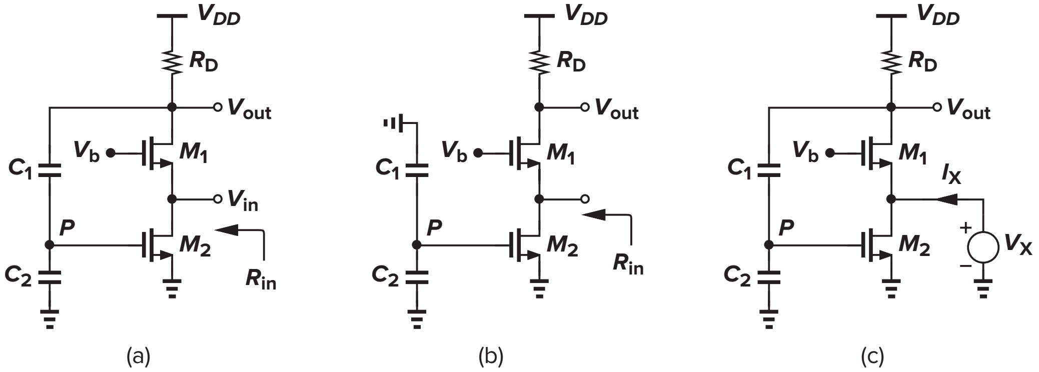


Figure 8.8 (a) Common-gate circuit with feedback; (b) open-loop circuit; (c) calculation of input resistance.

For the closed-loop circuit, as depicted in Fig. 8.8(c), we write $V_{out} = (g_{m1} + g_{mb1})V_X R_D$ and

$$V_P = V_{out}\frac{C_1}{C_1 + C_2} \tag{8.10}$$

$$= (g_{m1} + g_{mb1})V_X R_D \frac{C_1}{C_1 + C_2} \tag{8.11}$$

Thus, the small-signal drain current of M_2 equals $g_{m2}(g_{m1} + g_{mb1})V_X R_D C_1/(C_1 + C_2)$. Adding this current to the drain current of M_1 with proper polarity yields I_X:

$$I_X = (g_{m1} + g_{mb1})V_X + g_{m2}(g_{m1} + g_{mb1})\frac{C_1}{C_1 + C_2} R_D V_X \tag{8.12}$$

$$= (g_{m1} + g_{mb1})\left(1 + g_{m2}R_D\frac{C_1}{C_1 + C_2}\right)V_X \tag{8.13}$$

[3] The bias network for M_2 is not shown.

It follows that

$$R_{in,closed} = V_X/I_X \tag{8.14}$$

$$= \frac{1}{g_{m1}+g_{mb1}}\frac{1}{1+g_{m2}R_D\dfrac{C_1}{C_1+C_2}} \tag{8.15}$$

We therefore conclude that this type of feedback reduces the input resistance by a factor of $1 + g_{m2}R_DC_1/(C_1 + C_2)$. The reader can prove that the quantity $g_{m2}R_DC_1/(C_1 + C_2)$ is the loop gain.

Let us now consider the circuit of Fig. 8.9(a) as an example of output impedance modification by feedback. Here M_1, R_S, and R_D constitute a common-source stage and C_1, C_2, and M_2 sense the output voltage,[4] returning a current equal to $[C_1/(C_1 + C_2)]V_{out}g_{m2}$ to the source of M_1. The reader can prove that the feedback is indeed negative. To compute the output resistance at relatively low frequencies, we set the input to zero [Fig. 8.9(b)] and write

$$I_{D1} = V_X\frac{C_1}{C_1+C_2}g_{m2}\frac{R_S}{R_S+\dfrac{1}{g_{m1}+g_{mb1}}} \tag{8.16}$$

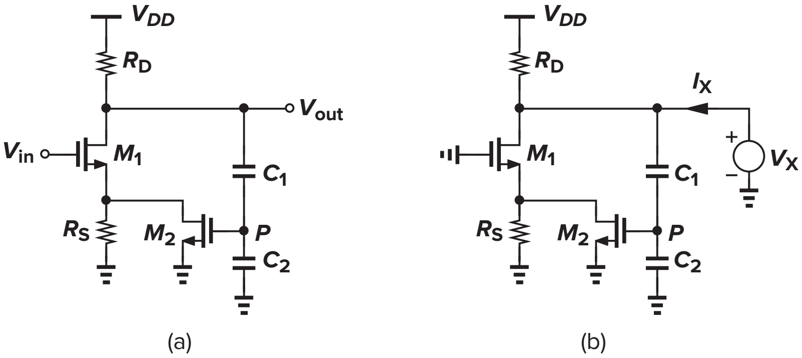


Figure 8.9 (a) CS stage with feedback; (b) calculation of output resistance.

Since $I_X = V_X/R_D + I_{D1}$, we have

$$\frac{V_X}{I_X} = \frac{R_D}{1+\dfrac{g_{m2}R_S(g_{m1}+g_{mb1})R_D}{(g_{m1}+g_{mb1})R_S+1}\dfrac{C_1}{C_1+C_2}} \tag{8.17}$$

Equation (8.17) implies that this type of feedback decreases the output resistance. The denominator of (8.17) is indeed equal to one plus the loop gain.

Bandwidth Modification. The next example illustrates the effect of negative feedback on the bandwidth. Suppose the feedforward amplifier has a one-pole transfer function:

$$A(s) = \frac{A_0}{1+\dfrac{s}{\omega_0}} \tag{8.18}$$

[4] Biasing of M_2 is not shown.

where A_0 denotes the low-frequency gain and ω_0 is the 3-dB bandwidth. What is the transfer function of the closed-loop system? From (8.5), we have

$$\frac{Y}{X}(s) = \frac{\dfrac{A_0}{1+\dfrac{s}{\omega_0}}}{1+\beta\dfrac{A_0}{1+\dfrac{s}{\omega_0}}} \tag{8.19}$$

$$= \frac{A_0}{1+\beta A_0+\dfrac{s}{\omega_0}} \tag{8.20}$$

$$= \frac{\dfrac{A_0}{1+\beta A_0}}{1+\dfrac{s}{(1+\beta A_0)\omega_0}} \tag{8.21}$$

The numerator of (8.21) is simply the closed-loop gain at low frequencies—as predicted by (8.5)—and the denominator reveals a pole at $(1+\beta A_0)\omega_0$. Thus, the 3-dB bandwidth has increased by a factor of $1+\beta A_0$, albeit at the cost of a proportional reduction in the gain (Fig. 8.10).

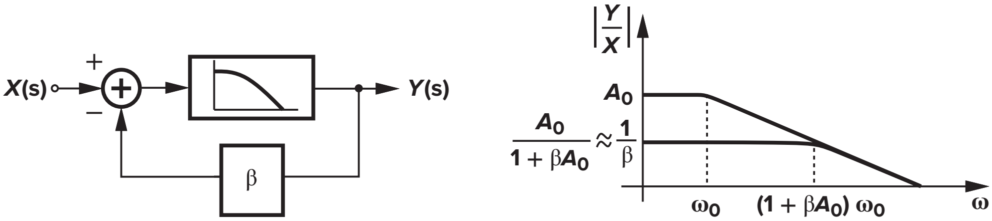


Figure 8.10 Bandwidth modification as a result of feedback.

The increase in the bandwidth fundamentally originates from the gain desensitization property of feedback. Recall from (8.6) that, if A is large enough, the closed-loop gain remains approximately equal to $1/\beta$ even if A experiences substantial variations. In the example of Fig. 8.10, A varies with frequency rather than process or temperature, but negative feedback still suppresses the effect of this variation. Of course, at high frequencies, A drops to such low levels that βA becomes comparable with unity and the closed-loop gain falls below $1/\beta$.

Equation (8.21) suggests that the "gain-bandwidth product" of a one-pole system is equal to $A_0\omega_0$ and does not change much with feedback, making the reader wonder how feedback improves the speed if a high gain is required. Suppose we need to amplify a 20-MHz square wave by a factor of 100 and maximum bandwidth, but we have only a single-pole amplifier with an open-loop gain of 100 and 3-dB bandwidth of 10 MHz. If the input is applied to the open-loop amplifier, the response appears as shown in Fig. 8.11(a), exhibiting a long risetime and falltime because the time constant is equal to $1/(2\pi f_{3\text{-dB}}) \approx 16$ ns.

Now suppose we apply feedback to the amplifier such that the gain and bandwidth are modified to 10 and 100 MHz, respectively. Placing two of these amplifiers in a cascade [Fig. 8.11(b)], we obtain a much faster response with an overall gain of 100. Of course, the cascade consumes twice as much power, but it would be quite difficult to achieve this performance with the original amplifier even if its power dissipation were doubled.

Nonlinearity Reduction An important property of negative feedback is the reduction of nonlinearity in analog circuits. A nonlinear characteristic is one that departs from a straight line, i.e., one whose *slope* varies (Fig. 8.12). A familiar example is the input-output characteristic of differential pairs. Note that

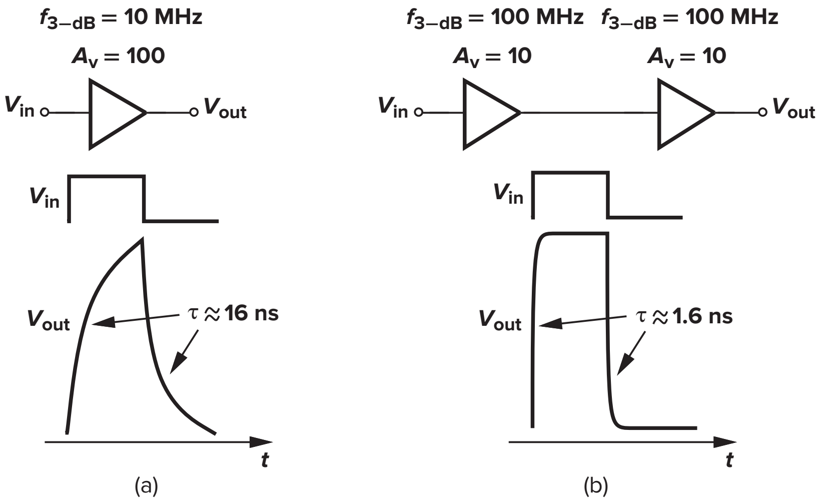


Figure 8.11 Amplification of a 20-MHz square wave by (a) a 10-MHz amplifier and (b) a cascade of two 100-MHz feedback amplifiers.

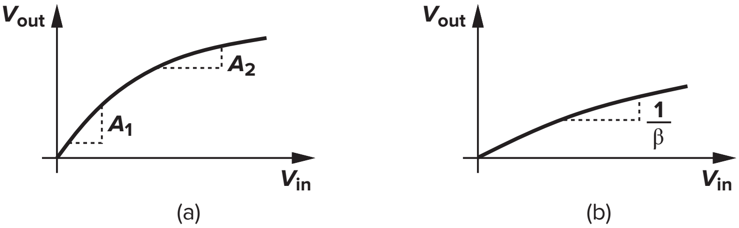


Figure 8.12 Input-output characteristic of a nonlinear amplifier (a) before and (b) after applying feedback.

the slope can be viewed as the small-signal gain. We predict that, even though the gain of an open-loop amplifier varies from A_1 to A_2 in Fig. 8.12, a closed-loop feedback system incorporating such an amplifier exhibits less gain variation and hence a higher linearity. To quantify this effect, we note that the open-loop gain ratio between regions 1 and 2 in Fig. 8.12 is equal to

$$r_{open} = \frac{A_2}{A_1} \tag{8.22}$$

For example, $r_{open} = 0.9$ means that the gain falls by 10% from region 1 to region 2. Assuming $A_2 = A_1 - \Delta A$, we can write

$$r_{open} = 1 - \frac{\Delta A}{A_1} \tag{8.23}$$

Let us place this amplifier in a negative-feedback loop. For the closed-loop gain ratio, we have

$$r_{closed} = \frac{\dfrac{A_2}{1+\beta A_2}}{\dfrac{A_1}{1+\beta A_1}} \tag{8.24}$$

$$= \frac{1+\dfrac{1}{\beta A_1}}{1+\dfrac{1}{\beta A_2}} \tag{8.25}$$

It follows that

$$r_{closed} \approx 1 - \frac{\dfrac{1}{\beta A_2} - \dfrac{1}{\beta A_1}}{1 + \dfrac{1}{\beta A_2}} \tag{8.26}$$

$$\approx 1 - \frac{A_1 - A_2}{1 + \beta A_2}\frac{1}{A_1} \tag{8.27}$$

$$\approx 1 - \frac{\Delta A}{1 + \beta A_2}\frac{1}{A_1} \tag{8.28}$$

Comparison of (8.23) and (8.28) suggests that the gain ratio is much closer to 1 in the latter if the loop gain, $1 + \beta A_2$, is large.

We study nonlinearity and its behavior in feedback systems more extensively in Chapter 14.

8.1.2 Types of Amplifiers

Most of the circuits studied thus far can be considered "voltage amplifiers" because they sense a voltage at the input and produce a voltage at the output. However, three other types of amplifiers can also be constructed such that they sense or produce currents. Shown in Fig. 8.13, the four configurations have quite different properties: (1) circuits sensing a voltage must exhibit a high input impedance (a voltmeter measures a voltage with minimal loading) whereas those sensing a current must provide a low input impedance (a current meter inserted in a wire must negligibly disturb the current); (2) circuits generating a voltage must exhibit a low output impedance (as a voltage source) while those generating a current must provide a high output impedance (as a current source). Note that the gains of transimpedance and transconductance[5] amplifiers have a dimension of resistance and conductance, respectively. For example, a transimpedance amplifier may have a gain of 2 kΩ, which means that it produces a 2-V output in response to a 1-mA input. Also, we use the sign conventions depicted in Fig. 8.13; for example, the transimpedance $R_0 = V_{out}/I_{in}$ if I_{in} flows *into* the amplifier.

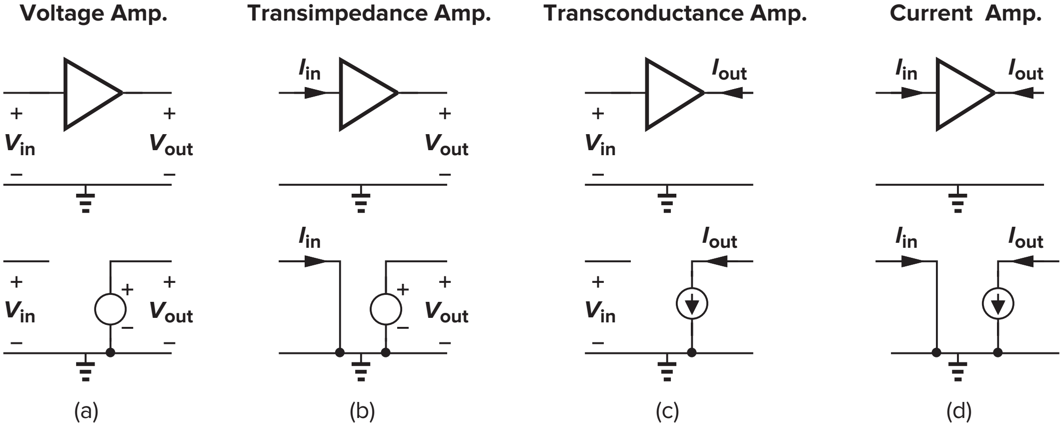


Figure 8.13 Types of amplifiers along with their idealized models.

[5] This terminology is standard but not consistent. One should use either transimpedance and transadmittance or transresistance and transconductance.

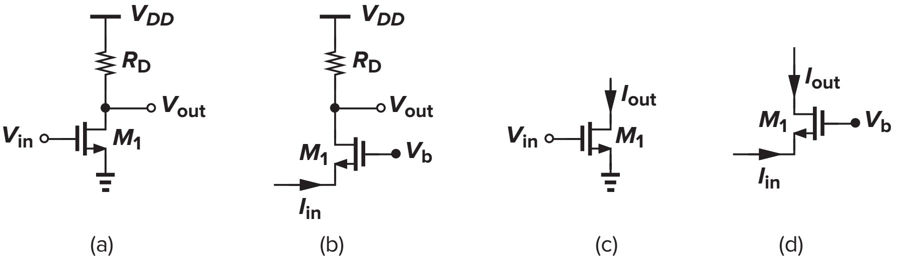


Figure 8.14 Simple implementations of four types of amplifiers.

Figure 8.14 illustrates simple implementations of each amplifier. In Fig. 8.14(a), a common-source stage senses and produces Voltages, and in Fig. 8.14(b), a common-gate circuit serves as a transimpedance amplifier, converting the source current to a voltage at the drain. In Fig. 8.14(c), a common-source transistor operates as a transconductance amplifier (also called a V/I converter), generating an output current in response to an input voltage, and in Fig. 8.14(d), a common-gate device senses and produces currents.

The circuits of Fig. 8.14 may not provide adequate performance in many applications. For example, the circuits of Figs. 8.14(a) and (b) suffer from a relatively high output impedance. Figure 8.15 depicts modifications that alter the output impedance or increase the gain.

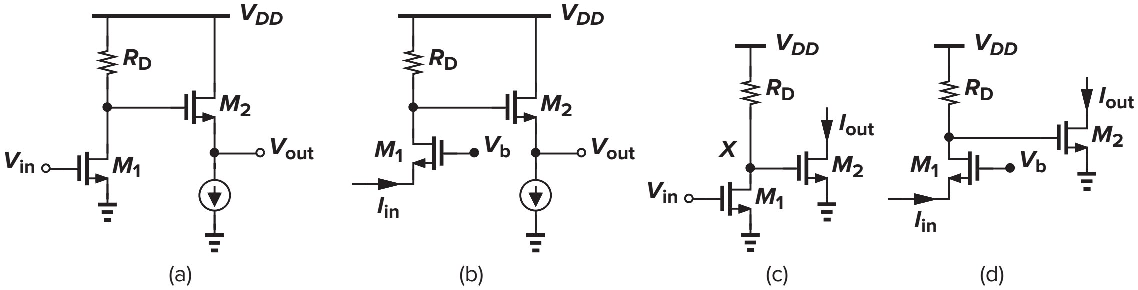


Figure 8.15 Four types of amplifiers with improved performance.

▶ **Example 8.2**

Calculate the gain of the transconductance amplifier shown in Fig. 8.15(c).

Solution

The gain in this case is defined as $G_m = I_{out}/V_{in}$. That is

$$G_m = \frac{V_X}{V_{in}} \cdot \frac{I_{out}}{V_X} \tag{8.29}$$

$$= -g_{m1}(r_{O1} \| R_D) \cdot g_{m2} \tag{8.30}$$

◀

While most familiar amplifiers are of the voltage-voltage type, the other three configurations do find usage. For example, transimpedance amplifiers are an integral part of optical fiber receivers because they must sense the current produced by a photodiode, eventually generating a voltage that can be processed by subsequent circuits.

▶ Example 8.3

Reconstruct the models of Fig. 8.13 for nonideal amplifiers.

Solution

A nonideal voltage amplifier draws current from its input and exhibits a finite output impedance, as depicted in Fig. 8.16(a).

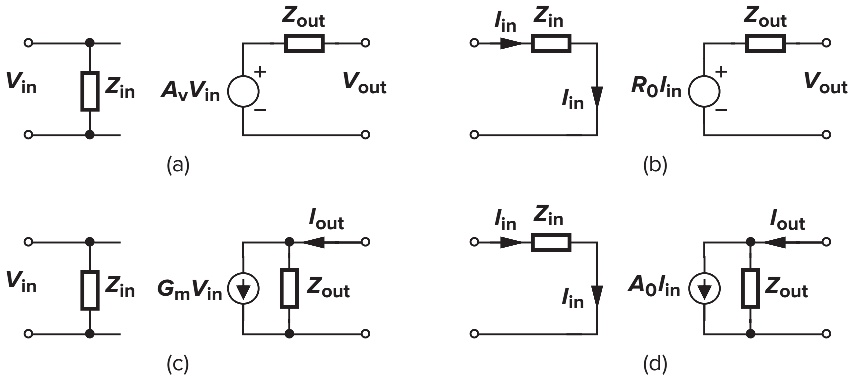


Figure 8.16

A nonideal transimpedance amplifier may have finite input and output impedances [Fig. 8.16(b)]. Note that Z_{in} is in *parallel* with the input port in Fig. 8.16(a) and in *series* with the input port in Fig. 8.16(b). This is to ensure a meaningful result in the ideal case: if Z_{in} goes to infinity in the former or to zero in the latter, the models reduce to those of Fig. 8.13.

The reader is encouraged to justify the models shown in Figs. 8.16(c) and (d) for the other two amplifier types. We should mention that these amplifiers may also have internal feedback from their output to their input, e.g., due to C_{GD}, but we neglect that for now. ◀

8.1.3 Sense and Return Mechanisms

Placing a circuit in a feedback loop requires sensing the output signal and returning (a fraction) of the result to the summing node at the input. With voltage or current quantities as input and output signals, we can identify four types of feedback: voltage-voltage, voltage-current, current-current, and current-voltage, where the first entry in each case denotes the quantity sensed at the *output* and the second the type of signal returned to the input.[6]

It is instructive to review methods of sensing and summing voltages or currents. To sense a voltage, we place a voltmeter *in parallel* with the corresponding port [Fig. 8.17(a)], ideally introducing no loading.

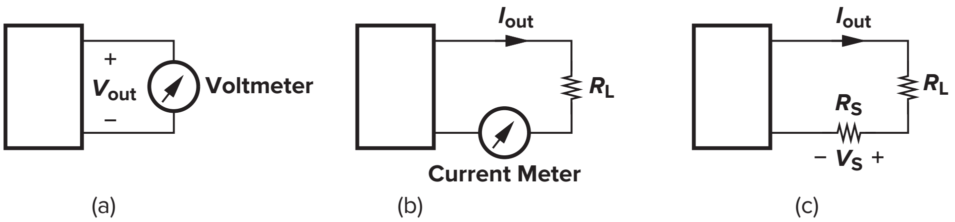


Figure 8.17 Sensing (a) a voltage by a voltmeter; (b) a current by a current meter; (c) a current by a small resistor.

[6] Different authors use different orders or terminologies for the four types of feedback.

When used in a feedback system, this type of sensing is also called "shunt feedback" (regardless of the quantity returned to the input).

To sense a current, a current meter is inserted *in series* with the signal [Fig. 8.17(b)], ideally exhibiting zero series resistance. Thus, this type of sensing is also called "series feedback." In practice, a small resistor replaces the current meter [Fig. 8.17(c)], with the voltage drop across the resistor serving as a measure of the output current.

The addition of the feedback signal and the input signal can be performed in the voltage domain or current domain. To add two quantities, we place them in series if they are voltages and in parallel if they are currents (Fig. 8.18).

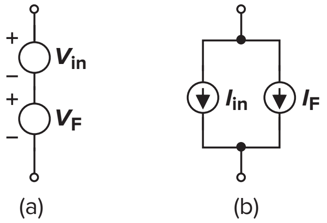


Figure 8.18 Addition of (a) voltages and (b) currents.

To visualize the methods of Figs. 8.17 and 8.18, we consider a number of practical implementations. A voltage can be sensed by a resistive (or capacitive) divider in parallel with the port [Fig. 8.19(a)] and a

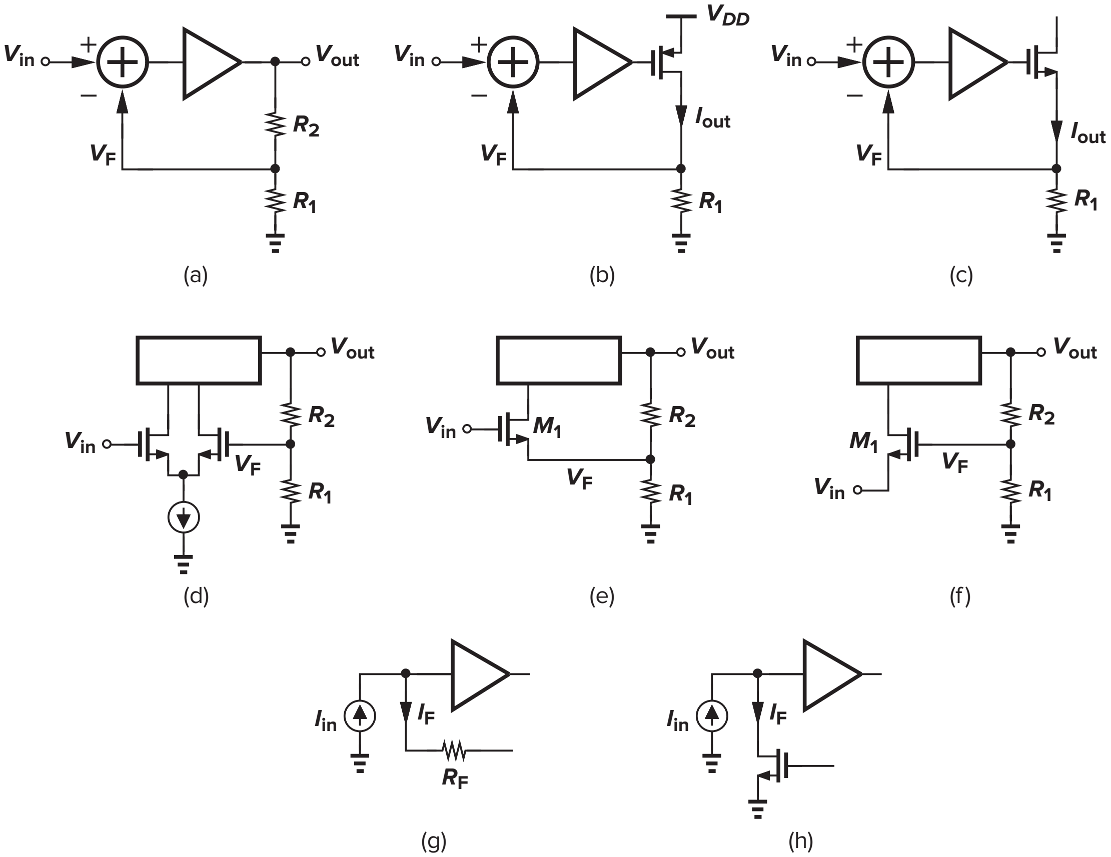


Figure 8.19 Practical means of sensing and adding voltages and currents.

current by placing a small resistor in series with the wire and sensing the voltage across it [Figs. 8.19(b) and (c)]. To subtract two voltages, a differential pair can be used [Fig. 8.19(d)]. Alternatively, a single transistor can perform voltage subtraction as shown in Figs. 8.19(e) and (f) because I_{D1} is a function of $V_{in} - V_F$. Subtraction of currents can be accomplished as depicted in Fig. 8.19(g) or (h). Note that for voltage subtraction, the input and feedback signals are applied to *two* distinct nodes, whereas for current subtraction, they are applied to a single node. This observation proves helpful in identifying the type of feedback.

While ideally having no influence on the operation of the open-loop amplifier itself, the feedback network in reality introduces loading effects that must be taken into account. This issue is discussed in Sec. 8.5.

8.2 ■ Feedback Topologies

In this section, we study four "canonical" topologies resulting from placing each of the four amplifier types in a negative-feedback loop. As depicted in Fig. 8.20, X and Y can be a current or a voltage quantity. The main amplifier is called the "feedforward" or simply the "forward" amplifier, around which we apply feedback to improve the performance.

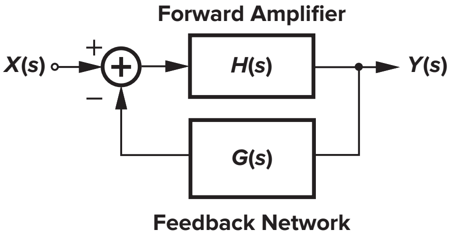


Figure 8.20 Canonical feedback system.

We should remark that some feedback circuits do not conform to the four canonical topologies. We return to this point later in the chapter, but the intuition gained from the analysis of these topologies proves essential to analog design. For example, we greatly benefit from the knowledge that one type of feedback lowers the output impedance while another raises it.

8.2.1 Voltage-Voltage Feedback

This topology senses the output voltage and returns the feedback signal as a voltage.[7] Following the conceptual illustrations of Figs. 8.17 and 8.18, we note that the feedback network is connected in *parallel* with the output and in *series* with the input port (Fig. 8.21). An ideal feedback network in this case exhibits infinite input impedance and zero output impedance because it senses a voltage and generates a voltage. We can therefore write $V_F = \beta V_{out}$, $V_e = V_{in} - V_F$, $V_{out} = A_0(V_{in} - \beta V_{out})$, and hence

$$\frac{V_{out}}{V_{in}} = \frac{A_0}{1 + \beta A_0} \tag{8.31}$$

We recognize that βA_0 is the loop gain and that the overall gain has dropped by $1 + \beta A_0$. Note that here both A_0 and β are dimensionless quantities.

As a simple example of voltage-voltage feedback, suppose we employ a differential voltage amplifier with single-ended output as the feedforward amplifier and a resistive divider as the feedback network

[7] This configuration is also called "series-shunt" feedback, where the first term refers to the *input* connection and the second to the *output* connection.

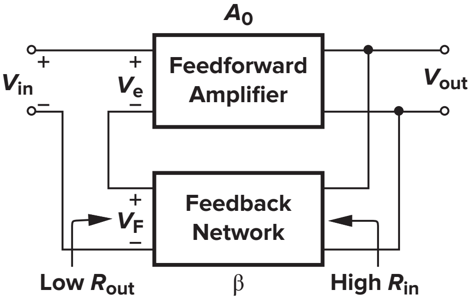


Figure 8.21 Voltage-voltage feedback.

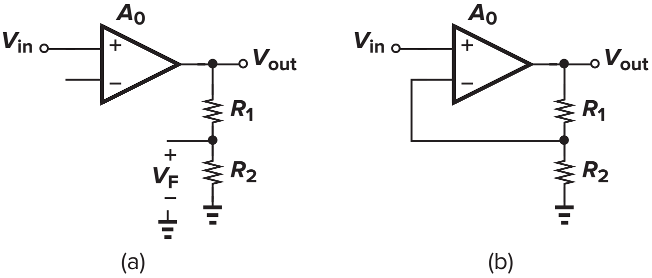


Figure 8.22 (a) Amplifier with output sensed by a resistive divider; (b) voltage-voltage feedback amplifier.

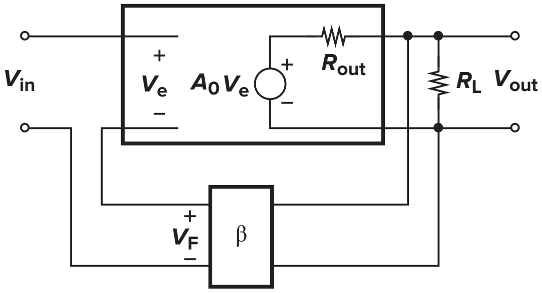


Figure 8.23 Effect of voltage-voltage feedback on output resistance.

[Fig. 8.22(a)]. The divider senses the output voltage, producing a fraction thereof as the feedback signal V_F. Following the block diagram of Fig. 8.21, we place V_F in series with the input of the amplifier to perform subtraction of voltages [Fig. 8.22(b)].

How does voltage-voltage feedback modify the input and output impedances? Let us first consider the output impedance. Recall that a negative-feedback system attempts to make the output an accurate (scaled) replica of the input. Now suppose, as shown in Fig. 8.23, we load the output by a resistor R_L, gradually decreasing its value. While in the open-loop configuration, the output would simply drop in proportion to $R_L/(R_L + R_{out})$, in the feedback system, V_{out} is maintained as a reasonable replica of V_{in} even though R_L decreases. That is, so long as the loop gain remains much greater than unity, $V_{out}/V_{in} \approx 1/\beta$, regardless of the value of R_L. From another point of view, since the circuit stabilizes ("regulates") the output voltage amplitude despite load variations, it behaves as a *voltage* source, thus exhibiting a low output impedance. This property fundamentally originates from the gain desensitization provided by feedback.

In order to formally prove that voltage feedback lowers the output impedance, we consider the simple model in Fig. 8.24, where R_{out} represents the output impedance of the feedforward amplifier. Setting the input to zero and applying a voltage at the output, we write $V_F = \beta V_X$, $V_e = -\beta V_X$, $V_M = -\beta A_0 V_X$,

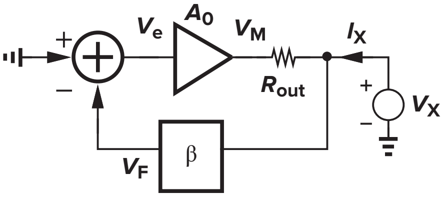


Figure 8.24 Calculation of output resistance of a voltage-voltage feedback circuit.

and hence $I_X = [V_X - (-\beta A_0 V_X)]/R_{out}$ (if the current drawn by the feedback network is neglected). It follows that

$$\frac{V_X}{I_X} = \frac{R_{out}}{1+\beta A_0} \tag{8.32}$$

Thus, the output impedance and the gain are lowered by the same factor. In the circuit of Fig. 8.22(b), for example, the output impedance is lowered by $1 + A_0 R_2/(R_1 + R_2)$.

▶ Example 8.4

The circuit shown in Fig. 8.25(a) is an implementation of the feedback configuration depicted in Fig. 8.22(b), but with the resistors replaced by capacitors. (The bias network of M_2 is not shown.) Calculate the closed-loop gain and output resistance of the amplifier at relatively low frequencies.

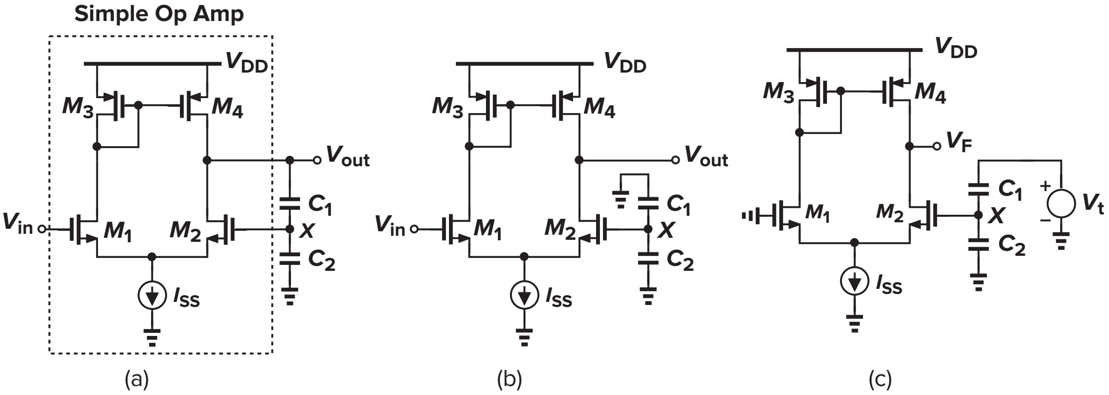


Figure 8.25

Solution

At low frequencies, C_1 and C_2 draw a negligible current from the output node. To find the open-loop voltage gain, we break the feedback loop as shown in Fig. 8.25(b), grounding the top plate of C_1 to ensure zero voltage feedback. The open-loop gain is thus equal to $g_{m1}(r_{O2} \| r_{O4})$.

We must also compute the loop gain, βA_0. With the aid of Fig. 8.25(c), we have

$$V_F = -V_t \frac{C_1}{C_1+C_2} g_{m1}(r_{O2} \| r_{O4}) \tag{8.33}$$

That is

$$\beta A_0 = \frac{C_1}{C_1+C_2} g_{m1}(r_{O2} \| r_{O4}) \tag{8.34}$$

and hence

$$A_{closed} = \frac{g_{m1}(r_{O2}\|r_{O4})}{1 + \dfrac{C_1}{C_1 + C_2} g_{m1}(r_{O2}\|r_{O4})} \tag{8.35}$$

As expected, if $\beta A_0 \gg 1$, then $A_{closed} \approx 1 + C_2/C_1$.

The open-loop output resistance of the circuit is equal to $r_{O2}\|r_{O4}$ (Chapter 5). It follows that

$$R_{out,closed} = \frac{r_{O2}\|r_{O4}}{1 + \dfrac{C_1}{C_1 + C_2} g_{m1}(r_{O2}\|r_{O4})} \tag{8.36}$$

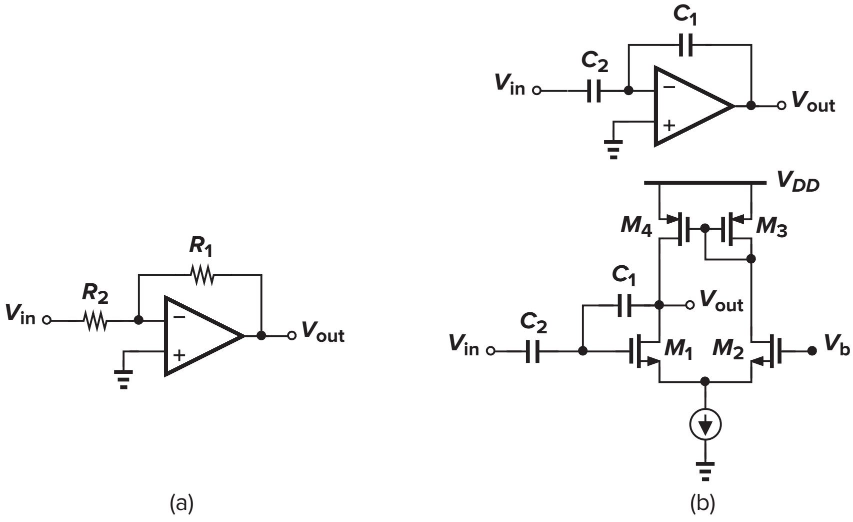
It is interesting to note that if $\beta A_0 \gg 1$, then

$$R_{out,closed} \approx \left(1 + \frac{C_2}{C_1}\right) \frac{1}{g_{m1}} \tag{8.37}$$

In other words, even if the open-loop amplifier suffers from a *high* output resistance, the closed-loop output resistance is independent of $r_{O2}\|r_{O4}$, simply because the open-loop *gain* scales with $r_{O2}\|r_{O4}$ as well.

▶ **Example 8.5**

Figure 8.26(a) shows an inverting amplifier using an op amp, and Fig. 8.26(b) illustrates a circuit implementation incorporating capacitors rather than resistors for the feedback network. Determine the loop gain and output impedance of the latter at low frequencies.



(a) (b)

Figure 8.26

Solution

With V_{in} set to zero, this circuit becomes indistinguishable from that in Fig. 8.25(a). Thus, the loop gain is given by (8.34) and the output impedance by (8.36).

The circuits in Figs. 8.25(a) and 8.26(b) appear similar, but provide different closed-loop gains, approximately $1 + C_2/C_1$ and $-C_2/C_1$, respectively. Thus, for a gain of, say, 4, $C_2/C_1 \approx 3$ in the former and $C_2/C_1 \approx 4$ in the latter. Which topology exhibits a higher loop gain in this case?

Voltage-voltage feedback also modifies the input impedance. Comparing the configurations in Fig. 8.27, we note that the input impedance of the feedforward amplifier sustains the entire input voltage in Fig. 8.27(a), but only a fraction of V_{in} in Fig. 8.27(b). As a result, the current drawn by R_{in} in the feedback topology is *less* than that in the open-loop system, suggesting that returning a voltage quantity to the input *increases* the input impedance.

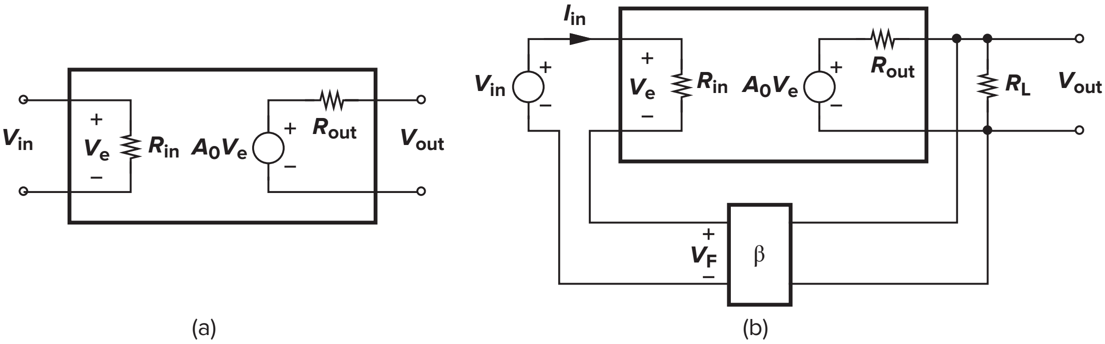


Figure 8.27 Effect of voltage-voltage feedback on input resistance.

The foregoing observation can be confirmed analytically with the aid of Fig. 8.28. Since $V_e = I_X R_{in}$ and $V_F = \beta A_0 I_X R_{in}$, we have $V_e = V_X - V_F = V_X - \beta A_0 I_X R_{in}$. Thus, $I_X R_{in} = V_X - \beta A_0 I_X R_{in}$, and

$$\frac{V_X}{I_X} = R_{in}(1 + \beta A_0) \tag{8.38}$$

The input impedance therefore increases by the ubiquitous factor $1 + \beta A_0$, bringing the circuit closer to an ideal voltage amplifier.

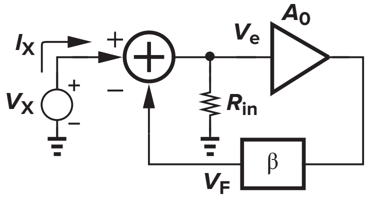


Figure 8.28 Calculation of input impedance of a voltage-voltage feedback circuit.

▶ Example 8.6

Figure 8.29(a) shows a common-gate topology placed in a voltage-voltage feedback configuration. Note that the summation of the feedback voltage and the input voltage is accomplished by applying the former to the gate and the latter to the source.[8] Calculate the input resistance at low frequencies if channel-length modulation is negligible.

Solution

Breaking the loop as depicted in Fig. 8.29(b), we recognize that the open-loop input resistance is equal to $(g_{m1} + g_{mb1})^{-1}$. To find the loop gain, we set the input to zero and inject a test signal in to the loop [Fig. 8.29(c)], obtaining $V_F/V_t = -g_{m1}R_DC_1/(C_1 + C_2)$. The closed-loop input impedance is then equal to

$$R_{in,closed} = \frac{1}{g_{m1} + g_{mb1}}\left(1 + \frac{C_1}{C_1 + C_2}g_{m1}R_D\right) \tag{8.39}$$

[8] This circuit is similar to the right half of the topology shown in Fig. 8.25(a).

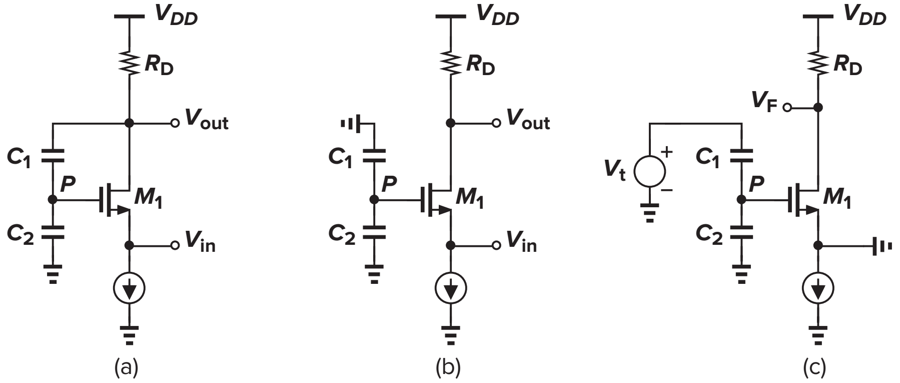


Figure 8.29

The increase in the input impedance can be explained as follows. Suppose the input voltage decreases by ΔV, causing the output voltage to fall. As a result, the gate voltage of M_1 *decreases*, thereby lowering the gate-source voltage of M_1 and yielding a change in V_{GS1} that is *less* than ΔV. This means that the drain current changes by an amount less than $(g_m + g_{mb})\Delta V$. By contrast, if the gate of M_1 were connected to a constant potential, the gate-source voltage would change by ΔV, resulting in a larger current change.

In summary, voltage-voltage feedback decreases the output impedance and increases the input impedance, thereby proving useful as a “buffer” stage that can be interposed between a high-impedance source and a low-impedance load.

8.2.2 Current-Voltage Feedback

In some circuits, it is desirable or simpler to sense the output current to perform feedback. The current is actually sensed by placing a (preferably small) resistor in series with the output and using the voltage drop across the resistor as the feedback information. This voltage may even serve as the return signal that is directly subtracted from the input.

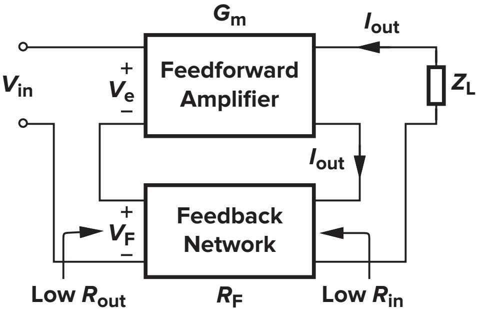


Figure 8.30 Current-voltage feedback.

Let us consider the general current-voltage feedback system illustrated in Fig. 8.30.[9] Since the feedback network senses the output current and returns a voltage, its feedback factor, β, has the dimension of resistance and is denoted by R_F. It is important to note that a G_m stage must be loaded (“terminated”) by a finite impedance, Z_L, to ensure that it can deliver its output current. If $Z_L = \infty$, then an ideal G_m

[9] This topology is also called “series-series” feedback.

stage would sustain an infinite output voltage. We write $V_F = R_F I_{out}$, $V_e = V_{in} - R_F I_{out}$, and hence $I_{out} = G_m(V_{in} - R_F I_{out})$. It follows that

$$\frac{I_{out}}{V_{in}} = \frac{G_m}{1 + G_m R_F} \tag{8.40}$$

An ideal feedback network in this case exhibits zero input and output impedances.

It is instructive to confirm that $G_m R_F$ is indeed the loop gain. As shown in Fig. 8.31, we set the input voltage to zero and break the loop by disconnecting the feedback network from the output and replacing it with a *short* at the output (if the feedback network is ideal). We then inject the test signal I_t, producing $V_F = R_F I_t$, and hence $I_{out} = -G_m R_F I_t$. Thus, the loop gain is equal to $G_m R_F$ and the transconductance of the amplifier is reduced by $1 + G_m R_F$ when feedback is applied.

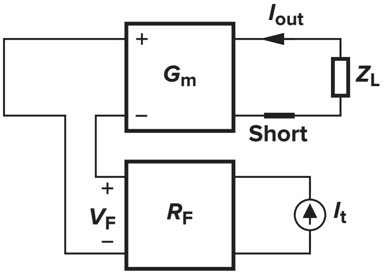


Figure 8.31 Calculation of loop gain for current-voltage feedback.

Is it realistic to assume that the input impedance of the feedback network is zero? Why do we use a test current rather than a test voltage? Does the type of test source affect the loop gain calculations? These questions are addressed later in this chapter.

Sensing the current at the output of a feedback system *increases* the output impedance. This is because the system attempts to make the output *current* a faithful replica of the input signal (with a proportionality factor if the input is a voltage quantity). Consequently, the system delivers the same current waveform as the load varies, in essence approaching an ideal current source and hence exhibiting a high output impedance.

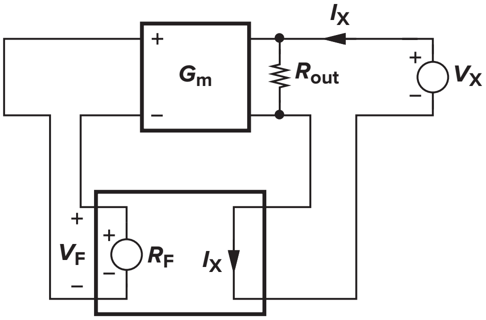


Figure 8.32 Calculation of output resistance of a current-voltage feedback amplifier.

To prove the above result, we consider the current-voltage feedback topology shown in Fig. 8.32, where R_{out} represents the finite output impedance of the feedforward amplifier.[10] The feedback network produces a voltage V_F proportional to I_X:$V_F = R_F I_X$, and the current generated by G_m equals $-R_F I_X G_m$. As a

[10] Note that R_{out} is placed in *parallel* with the output because the ideal transimpedance amplifier is modeled by a voltage-dependent current source.

result, $-R_F I_X G_m = I_X - V_X/R_{out}$, yielding

$$\frac{V_X}{I_X} = R_{out}(1 + G_m R_F) \tag{8.41}$$

The output impedance therefore increases by a factor of $1 + G_m R_F$.

▶ **Example 8.7**

Rechargeable batteries must be charged by a constant current (rather than a constant voltage) to avoid damage. The battery charger must therefore generate a constant current from a golden reference, V_{REF}. As shown in Fig. 8.33(a), we can insert a small resistor r in the output current path, apply the voltage across r to an amplifier A_1, and subtract the output of A_1 from V_{REF}. Calculate the output current and impedance of this circuit, assuming $|Z_L| \ll r_O$ (the output resistance of M_1).

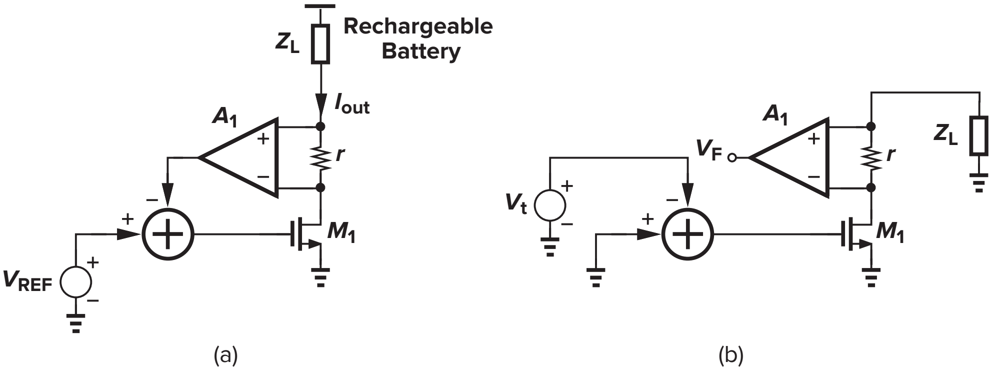


Figure 8.33

Solution

With a high loop gain, the output voltage of A_1 is approximately equal to V_{REF}, and hence $I_{out} = (V_{REF}/A_1)/r$. Using the circuit of Fig. 8.33(b) to determine the loop gain, we have

$$\frac{V_F}{V_t} \approx -g_m r A_1 \tag{8.42}$$

Thus, the open-loop output impedance seen by Z_L is multiplied by $1 + g_m r A_1$, yielding

$$R_{out,closed} = (1 + g_m r A_1)(r_O + r) \tag{8.43}$$

We observe that Z_L is now driven by a better current source. ◀

As with voltage-voltage feedback, current-voltage feedback increases the input impedance by a factor equal to one plus the loop gain. As illustrated in Fig. 8.34, we have $I_X R_{in} G_m = I_{out}$. Thus, $V_e = V_X - G_m R_F I_X R_{in}$ and

$$\frac{V_X}{I_X} = R_{in}(1 + G_m R_F) \tag{8.44}$$

The reader can show that the loop gain is indeed equal to $G_m R_F$.

In summary, current-voltage feedback increases both the input and the output impedances while decreasing the feedforward transconductance. As explained in Chapter 9, the high output impedance proves useful in high-gain op amps.

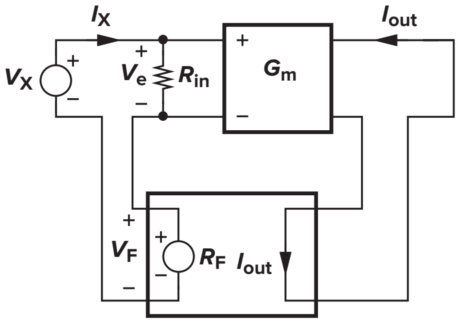


Figure 8.34 Calculation of input resistance of a current-voltage feedback amplifier.

8.2.3 Voltage-Current Feedback

In this type of feedback, the output voltage is sensed and a proportional current is returned to the summing point at the input.[11] Note that the feedforward path incorporates a transimpedance amplifier with gain R_0 and the feedback factor has a dimension of conductance.

A voltage-current feedback topology is shown in Fig. 8.35. Sensing a voltage and producing a current, the feedback network is characterized by a transconductance g_{mF}, ideally exhibiting infinite input and output impedances. Since $I_F = g_{mF} V_{out}$ and $I_e = I_{in} - I_F$, we have $V_{out} = R_0 I_e = R_0(I_{in} - g_{mF} V_{out})$. It follows that

$$\frac{V_{out}}{I_{in}} = \frac{R_0}{1 + g_{mF} R_0} \tag{8.45}$$

The reader can prove that $g_{mF} R_0$ is indeed the loop gain, concluding that this type of feedback lowers the transimpedance by a factor equal to one plus the loop gain.

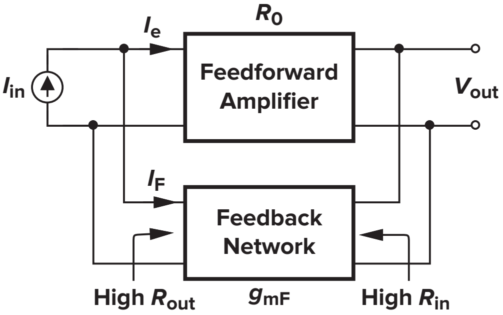


Figure 8.35 Voltage-current feedback.

▶ **Example 8.8**

Calculate the transimpedance, V_{out}/I_{in}, of the circuit shown in Fig. 8.36(a) at relatively low frequencies. Assume that $\lambda = 0$. (The bias network of M_1 is not shown.)

Solution

In this circuit, the capacitive divider C_1-C_2 senses the output voltage, applying the result to the gate of M_1 and producing a current that is subtracted from I_{in}. The open-loop transimpedance equals that of the core common-gate stage, R_D. The loop gain is obtained by setting I_{in} to zero and breaking the loop at the output [Fig. 8.36(b)]:

$$-V_t \frac{C_1}{C_1 + C_2} g_{m1} R_D = V_F \tag{8.46}$$

[11] This topology is also called "shunt-shunt" feedback.

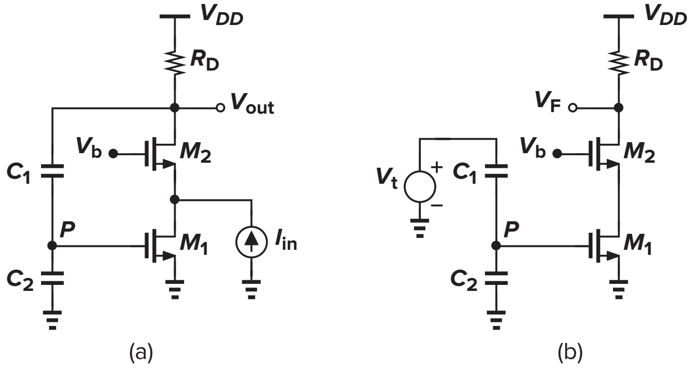


Figure 8.36

Thus, the overall transimpedance is equal to

$$R_{tot} = \frac{R_D}{1 + \dfrac{C_1}{C_1 + C_2} g_{m1} R_D} \tag{8.47}$$

▶ Example 8.9

We know from the previous example that

$$R_{in} = \frac{1}{g_{m2}} \frac{1}{1 + \dfrac{C_1}{C_1 + C_2} g_{m1} R_D} \tag{8.48}$$

A student repeats the analysis, but with the input driven by a voltage source, concluding that the loop gain is *zero* and the input impedance is not affected by the feedback loop. Explain the flaw in the student's argument.

Solution

Consider the arrangement shown in Fig. 8.37(a). We know that R_{in} *is* affected by the feedback because M_1 generates a current in response to V_{in}. On the other hand, it appears from Fig. 8.37(b) that the loop gain is zero in this case. How do we reconcile these two views?

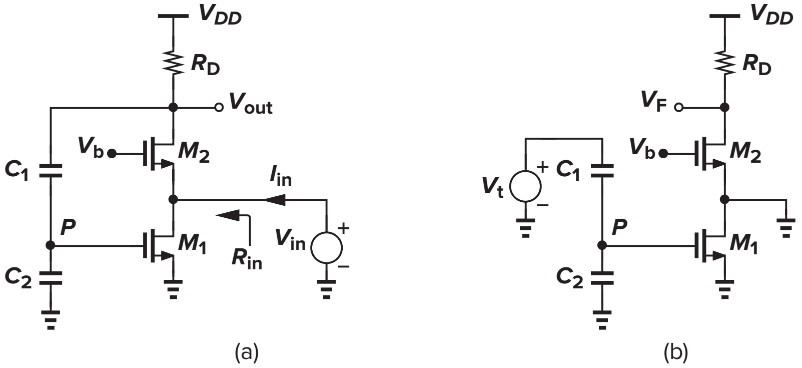


Figure 8.37

We must recall that returning current to the input assumes that the circuit is driven by a current source; i.e., our generic negative-feedback system requires that the returned quantity and the input have the same dimension. In other words, the circuit of Fig. 8.37(a) does *not* map to our canonical feedback system because it returns a current but is driven by a voltage. We therefore cannot compute the loop gain by setting the input voltage to zero and breaking the loop. Of course, the input impedance is still given by Eq. (8.48). We will return to this circuit in Sec. 8.6.4 and apply Blackman's theorem to it.

Following our reasoning for the other two types of feedback studied above, we surmise that voltage-current feedback decreases both the input and the output impedances. As shown in Fig. 8.38(a) and noted in Example 8.3, the input resistance of R_0 appears in *series* with its input port. We write $I_F = I_X - V_X/R_{in}$ and $(V_X/R_{in})R_0 g_{mF} = I_F$. Thus,

$$\frac{V_X}{I_X} = \frac{R_{in}}{1 + g_{mF}R_0} \tag{8.49}$$

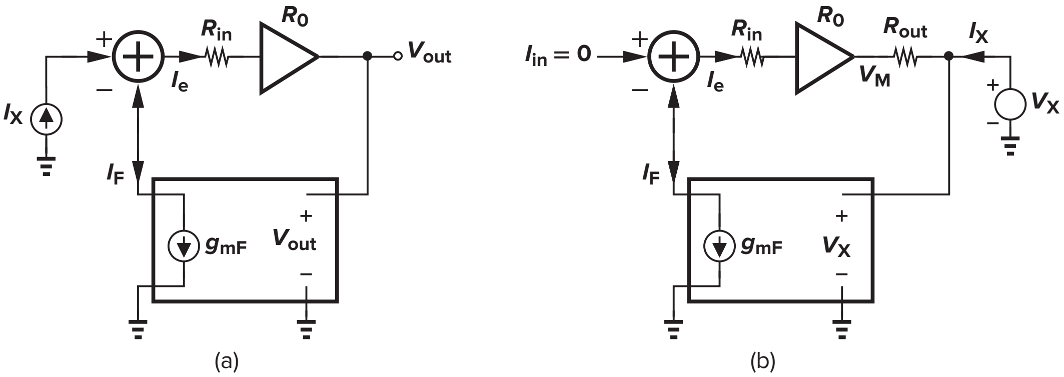


Figure 8.38 Calculation of (a) input and (b) output impedance of a voltage-current feedback amplifier.

Similarly, from Fig. 8.38(b), we have $I_F = V_X g_{mF}$, $I_e = -I_F$, and $V_M = -R_0 g_{mF} V_X$. Neglecting the input current of the feedback network, we write $I_X = (V_X - V_M)/R_{out} = (V_X + g_{mF}R_0V_X)/R_{out}$. That is

$$\frac{V_X}{I_X} = \frac{R_{out}}{1 + g_{mF}R_0} \tag{8.50}$$

▶ **Example 8.10**

Calculate the input and output impedances of the circuit shown in Fig. 8.39(a). For simplicity, assume that $R_F \gg R_D$.

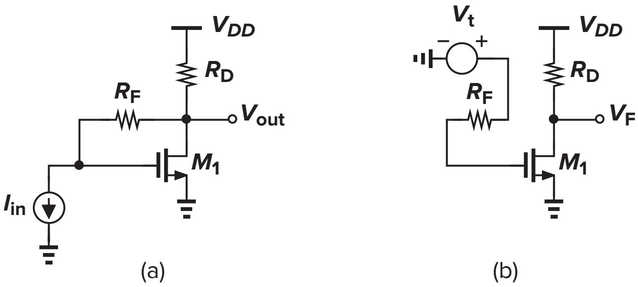


Figure 8.39

Solution

In this circuit, R_F senses the output voltage and returns a current to the input. Breaking the loop as depicted in Fig. 8.39(b), we calculate the loop gain as $g_m R_D$. Thus, the open-loop input impedance, R_F, is divided by $1 + g_m R_D$:

$$R_{in,closed} = \frac{R_F}{1 + g_m R_D} \tag{8.51}$$

Similarly,

$$R_{out,closed} = \frac{R_D}{1 + g_m R_D} \tag{8.52}$$

$$= \frac{1}{g_m} || R_D \tag{8.53}$$

Note that $R_{out,closed}$ is in fact the parallel combination of a diode-connected transistor and R_D.

The reduction of the input impedance agrees with Miller's prediction: since the voltage gain from the gate of M_1 to its drain is approximately equal to $-g_m R_D$, the feedback resistor equivalently produces a grounded resistance at the input equal to $R_F/(1 + g_m R_D)$.

An important application of amplifiers with *low* input impedance is in fiber optic receivers, where light received through a fiber is converted to a *current* by a reverse-biased photodiode. This current is typically converted to a voltage for further amplification and processing. Shown in Fig. 8.40(a), such conversion can be accomplished by a simple resistor, but at the cost of bandwidth because the diode suffers from a relatively large junction capacitance. For this reason, the feedback topology of Fig. 8.40(b) is usually employed, where R_1 is placed around the voltage amplifier A to form a "transimpedance amplifier" (TIA). The input impedance is $R_1/(1 + A)$ and the output voltage is approximately $-R_1 I_{D1}$. The bandwidth thus increases from $1/(2\pi R_1 C_{D1})$ to $(1 + A)/(2\pi R_1 C_{D1})$ if A itself is a wideband amplifier.

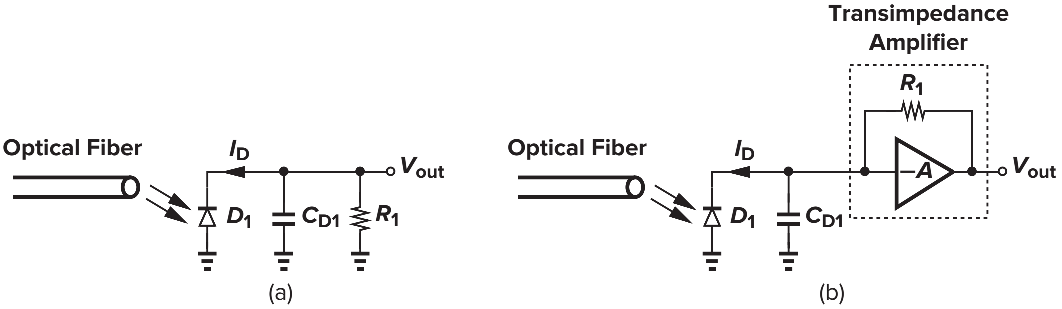


Figure 8.40 Detection of current produced by a photodiode by (a) resistor R_1 and (b) a transimpedance amplifier.

8.2.4 Current-Current Feedback

Figure 8.41 illustrates this type of feedback.[12] Here, the feedforward amplifier is characterized by a current gain, A_I, and the feedback network by a current ratio, β. In a fashion similar to the previous derivations, the reader can easily prove that the closed-loop current gain is equal to $A_I/(1 + \beta A_I)$, the input impedance is divided by $1 + \beta A_I$ and the output impedance is multiplied by $1 + \beta A_I$.

[12] This topology is also called "shunt-series" feedback, where the first term refers to the input connection and the second to the output connection.

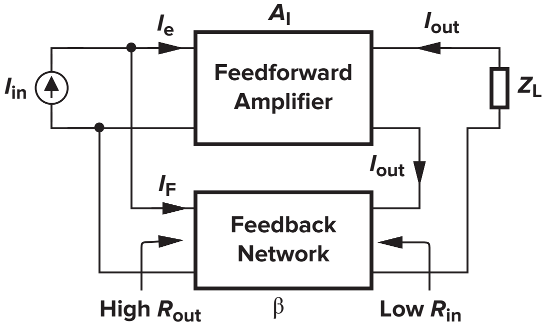


Figure 8.41 Current-current feedback.

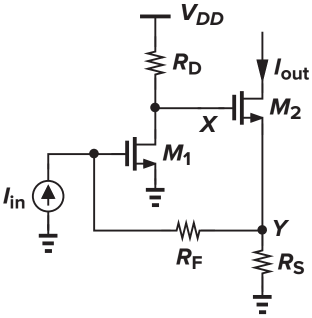


Figure 8.42

Figure 8.42 illustrates an example of current-current feedback. Here, since the source and drain currents of M_2 are equal (at low frequencies), resistor R_S is inserted in the source network to monitor the output current. Resistor R_F plays the same role as in Fig. 8.39.

8.3 ■ Effect of Feedback on Noise

Feedback does not improve the noise performance of circuits. Let us first consider the simple case illustrated in Fig. 8.43(a), where the open-loop voltage amplifier A_1 is characterized by only an input-referred noise voltage and the feedback network is noiseless. We have $(V_{in} - \beta V_{out} + V_n)A_1 = V_{out}$, and hence

$$V_{out} = (V_{in} + V_n)\frac{A_1}{1 + \beta A_1} \tag{8.54}$$

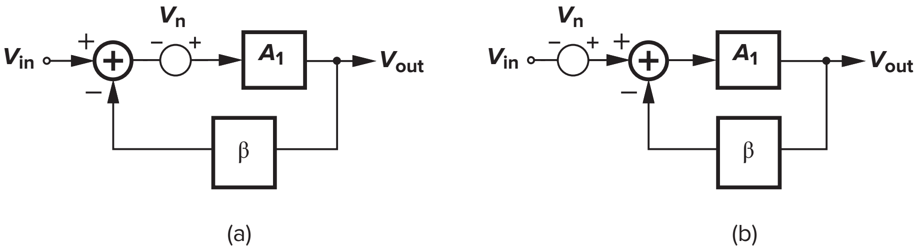


Figure 8.43 Feedback around a noisy circuit.

Thus, the circuit can be simplified as shown in Fig. 8.43(b), revealing that the input-referred noise of the overall circuit is still equal to V_n. This analysis can be extended to all four feedback topologies to prove that the input-referred noise voltage and current remain the same if the feedback network introduces no noise. In practice, the feedback network itself may contain resistors or transistors, degrading the overall noise performance.

It is important to note that in Fig. 8.43(a), the output of interest is the same as the quantity sensed by the feedback network. This need not always be the case. For example, in the circuit of Fig. 8.44, the output is provided at the drain of M_1 whereas the feedback network senses the voltage at the source of M_1. In such cases, the input-referred noise of the closed-loop circuit may not be equal to that of the open-loop circuit even if the feedback network is noiseless. As an example, let us consider the topology of Fig. 8.44 and, for simplicity, take only the noise of R_D, $V_{n,RD}$, into account. The reader can prove that the closed-loop voltage gain is equal to $-A_1 g_m R_D/[1+(1+A_1)g_m R_S]$ if $\lambda = \gamma = 0$, and hence the input-referred noise voltage due to R_D is

$$\left|V_{n,in,closed}\right| = \frac{|V_{n,RD}|}{A_1 R_D}\left[\frac{1}{g_m} + (1+A_1)R_S\right] \tag{8.55}$$

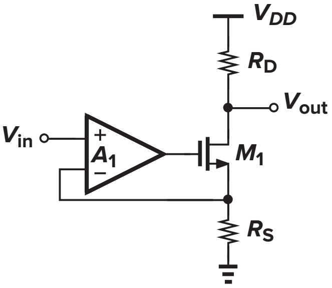


Figure 8.44 Noisy circuit with feedback sensing the source voltage.

For the open-loop circuit, on the other hand, the input-referred noise is

$$\left|V_{n,in,open}\right| = \frac{|V_{n,RD}|}{A_1 R_D}\left[\frac{1}{g_m} + R_S\right] \tag{8.56}$$

Interestingly, as $A_1 \to \infty$, $|V_{n,in,closed}| \to |V_{n,RD}|R_S/R_D$ whereas $|V_{n,in,open}| \to 0$.

8.4 ■ Feedback Analysis Difficulties

Our study of feedback systems has made some simplifying assumptions that may not hold in all circuits. In this section, we point out five difficulties that arise in the analysis of feedback circuits, and in subsequent sections, we deal with some of them.

The analysis approach described previously proceeds as follows: (a) break the loop and obtain the open-loop gain and input and output impedances, (b) determine the loop gain, βA_0, and hence the closed-loop parameters from their open-loop counterparts, and (c) use the loop gain to study properties such as stability (Chapter 10), etc. However, this approach faces issues in some circuits.

The first difficulty relates to breaking the loop and stems from the "loading" effects imposed by the feedback network upon the feedforward amplifier. For example, in the noninverting amplifier of Fig. 8.45(a) and its simple implementation shown in Fig. 8.45(b), the feedback branch consisting of R_1 and R_2 may draw a significant signal current from the op amp, reducing its *open-loop* gain. Figure 8.45(c) depicts another case, in which the open-loop gain of the forward CS stage falls if R_F is not very large. In both cases, this "output" loading results from the nonideal input impedance of the feedback network.

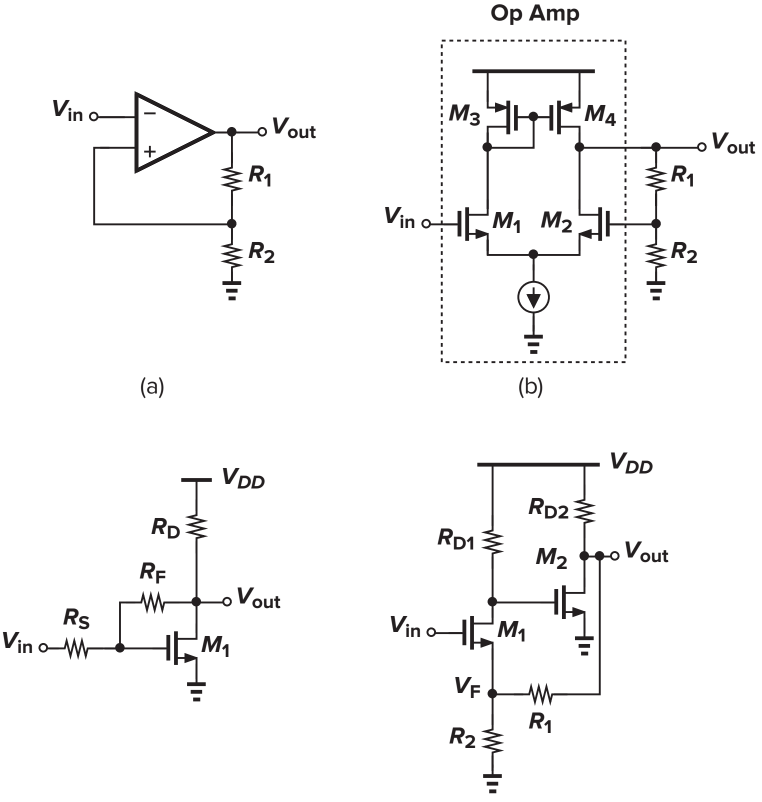


Figure 8.45 (a) Noninverting amplifier, (b) implementation using a differential pair, (c) implementation using a CS stage, and (d) implementation using a two-stage amplifier.

As another example, consider the arrangement shown in Fig. 8.45(d), where R_1 and R_2 sense V_{out} and return a voltage to the source of M_1. Since the *output* impedance of the feedback network may not be sufficiently small, we surmise that M_1 is degenerated appreciably even as far as the open-loop forward amplifier is concerned. This circuit exemplifies "input loading" due to the nonideal output impedance of the feedback network.

The important question that we must address with regard to loading is, how do we break the loop while properly including output and input loading effects?

▶ **Example 8.11**

Can the loop be broken at the gate of M_2 in Fig. 8.45(d) without concern for loading effects?

Solution

As illustrated in Fig. 8.46, such an attempt provides the loop gain while avoiding loading issues. However, we are also interested in the open-loop gain and the open-loop input and output impedances, which cannot be obtained from this configuration. We must therefore develop a methodical approach to constructing the *open-loop* system such that the loading effects are included.

◀

The second difficulty is that some circuits cannot be clearly decomposed into a forward amplifier and a feedback network. In the two-stage network of Fig. 8.47, it is unclear whether R_{D2} belongs to the

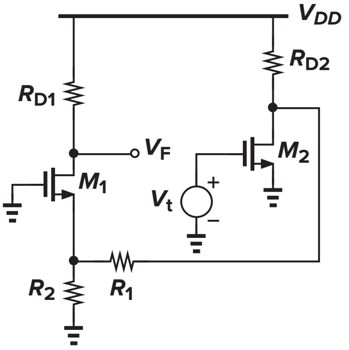


Figure 8.46

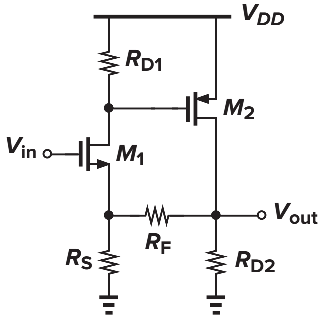


Figure 8.47 Feedback circuit without a clearly-distinguishable feedback network.

feedforward amplifier or the feedback network. We might choose the former case, reasoning that M_2 needs a load so as to operate as a voltage amplifier, but such a choice seems arbitrary.

The third difficulty in feedback analysis is that some circuits do not readily map to the four canonical topologies studied in the previous sections. For example, a simple degenerated common-source stage does contain feedback because the source resistance measures the drain current, converts it to voltage, and subtracts the result from the input [Fig. 8.48(a)]. However, it is not immediately clear which feedback topology represents this arrangement because the sensed quantity, I_{D1}, is different from the output of interest, V_{out} [Fig. 8.48(b)].

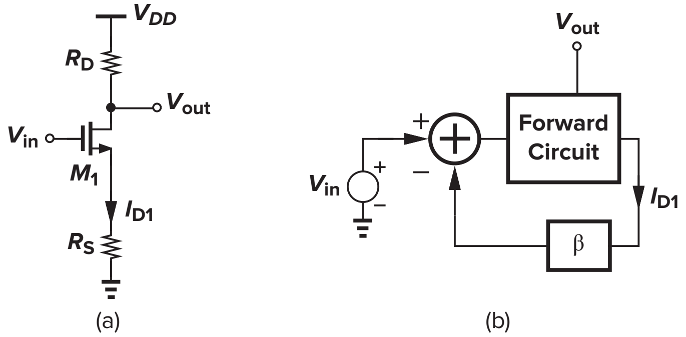


Figure 8.48 (a) CS stage and (b) block diagram showing the output and sense ports.

The fourth difficulty is that the general feedback system analyzed thus far assumes unilateral stages, i.e., signal propagation in only *one* direction around the loop. In practice, however, the loop may contain bilateral circuits, allowing signals to flow from the output toward the input through a path other than the

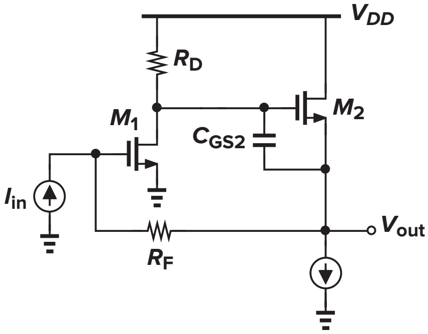
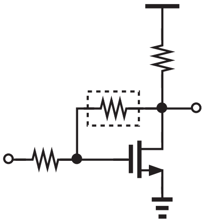
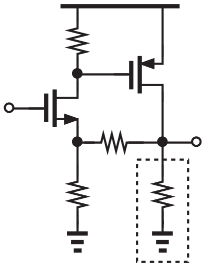
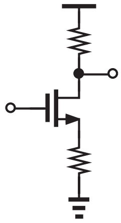
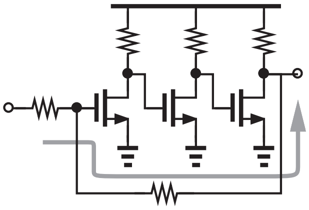
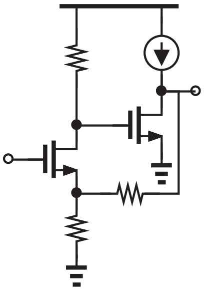


Figure 8.49 Example of circuit with more than one feedback mechanism.

nominal feedback path. In Fig. 8.47, for example, the signal leaks from the drain of M_2 to its gate through C_{GD2} at high frequencies.

The fifth difficulty arises in circuits containing multiple feedback mechanisms (loosely called "multiloop" circuits). In the topology of Fig. 8.49, for example, R_F provides feedback around the circuit, and C_{GS2} around M_2. We can also say that the source follower itself contains degeneration and hence feedback. We must then ask, which loop should be broken and what exactly do we mean by "loop gain" in this case? Table 8.1 summarizes the five issues described here.

Table 8.1 Feedback analysis difficulties.

Loading	Ambiguous Decomposition	Noncanonical Topologies	Nonunilateral Loop	Multiple Feedback Mechanisms
				

In this chapter, we introduce three methods of feedback circuit analysis. Outlined in Table 8.2, the first employs two-port models to analyze the four canonical topologies while including loading effects.

Table 8.2 Three methods of feedback analysis.

Two-Port Method	Bode's Method	Middlebrook's Method
• Computes open-loop and closed-loop quantities and the loop gain. • Includes loading effects. • Neglects feedforward through feedback network. • Can be applied recursively to multiple feedback mechanisms. • Does not apply to noncanonical topologies.	• Computes closed-loop quantities without breaking the loop. • Applies to any topology. • Provides loop gain only if one feedback mechanism is present.	• Computes closed-loop quantities without breaking the loop. • Applies to any topology. • Provides loop gain only if local and global loops are distinguishable. • Reveals effect of reverse loop gain in nonunilateral loops.

This method proves more efficient than direct analysis of the circuit (with no knowledge of feedback) if the loop is assumed unilateral, i.e., the forward propagation of the input signal through the feedback network is neglected, and so is the backward propagation of the signal through the forward amplifier. The other two methods do not attempt to break the loop and yield the closed-loop quantities exactly but with lengthier algebra.

8.5 ■ Effect of Loading

The problem of loading manifests itself when we need to break the feedback loop so as to identify the open-loop system, e.g., calculate the open-loop gain and the input and output impedances. To arrive at the proper procedure for including the feedback network terminal impedances, we first review models of two-port networks.

8.5.1 Two-Port Network Models

The simplified amplifier and feedback network models employed in the previous sections may not suffice in general. We must therefore resort to accurate two-port models. For example, the feedback network placed around the feedforward amplifier can be considered a two-port circuit sensing and producing voltages or currents. Recall from basic circuit theory that a two-port linear (and time-invariant) network can be represented by any of the four models shown in Fig. 8.50. The "Z model" in Fig. 8.50(a) consists of input and output impedances in series with current-dependent voltage sources, whereas the "Y model" in Fig. 8.50(b) comprises input and output admittances in parallel with voltage-dependent current sources. The "hybrid models" of Figs. 8.50(c) and (d) incorporate a combination of impedances and admittances and voltage sources and current sources. Each model is described by two equations. For the Z model, we have

$$V_1 = Z_{11}I_1 + Z_{12}I_2 \tag{8.57}$$

$$V_2 = Z_{21}I_1 + Z_{22}I_2 \tag{8.58}$$

Each Z parameter has a dimension of impedance and is obtained by leaving one port open, e.g., $Z_{11} = V_1/I_1$ when $I_2 = 0$. Similarly, for the Y model,

$$I_1 = Y_{11}V_1 + Y_{12}V_2 \tag{8.59}$$

$$I_2 = Y_{21}V_1 + Y_{22}V_2 \tag{8.60}$$

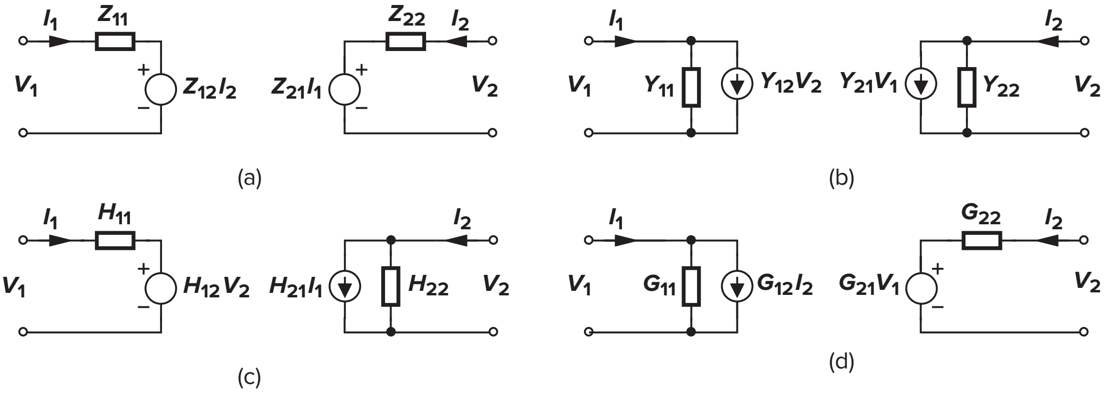


Figure 8.50 Linear two-port network models.

where each Y parameter is calculated by shorting one port, e.g., $Y_{11} = I_1/V_1$ when $V_2 = 0$. For the H model,

$$V_1 = H_{11}I_1 + H_{12}V_2 \tag{8.61}$$

$$I_2 = H_{21}I_1 + H_{22}V_2 \tag{8.62}$$

and for the G model,

$$I_1 = G_{11}V_1 + G_{12}I_2 \tag{8.63}$$

$$V_2 = G_{21}V_1 + G_{22}I_2 \tag{8.64}$$

Note that, for example, Y_{11} may not be equal to the inverse of Z_{11} because the two are obtained under different conditions: the output is shorted for the former but left open for the latter.

It is instructive to compare the general two-port models with the simplified amplifier representations that we have used in the previous sections. For example, let us consider the voltage amplifier model in Example 8.3 vis-à-vis the Z model. We observe that (1) absent in the former, $Z_{12}I_2$ represents the amplifier's *internal* feedback, e.g., due to C_{GD}; (2) if Z_{12} is zero, then Z_{11} is equal to Z_{in}, the input impedance calculated with the output left open; and (3) Z_{22} is not necessarily equal to Z_{out}: the former is computed with the input port left open and the latter with the input *shorted*.

The most important drawback of the Z model for our purposes is that its output generator, $Z_{21}I_1$, is controlled by the input *current* rather than the input voltage. For a MOS circuit with the input applied to the gate, this model becomes meaningless if the input capacitance is neglected. The H model entails the same difficulty.

Do any of the two-port models agree with our intuitive picture of voltage amplifiers? Yes, the G model is close. If the internal feedback, $G_{12}I_2$, is neglected, then G_{11} $(= I_1/V_1$ with $I_2 = 0)$ represents the inverse of the input impedance, and G_{22} $(= V_2/I_2$ with $V_1 = 0)$ the output impedance. The reader can try this exercise for the other three types of amplifiers.

8.5.2 Loading in Voltage-Voltage Feedback

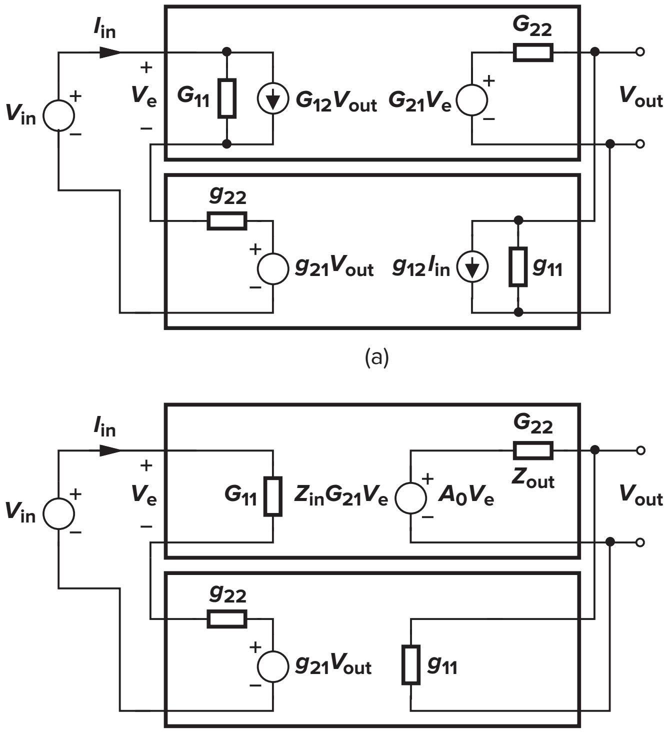
As mentioned before, the Z and H models fail to represent voltage amplifiers if the input current is very small—as in a simple CS stage. We therefore choose the G model here.[13] The complete equivalent circuit is shown in Fig. 8.51(a), where the forward and feedback network parameters are denoted by upper-case and lower-case letters, respectively. Since the *input* port of the feedback network is connected to the *output* port of the forward amplifier, g_{11} and $g_{12}I_{in}$ are tied to V_{out}.

It is possible to solve this circuit exactly, but we simplify the analysis by neglecting two quantities: the amplifier's internal feedback, $G_{12}V_{out}$, and the "forward" propagation of the input signal through the feedback network, $g_{12}I_{in}$. In other words, the loop is "unilateralized." Figure 8.51(b) depicts the resulting circuit with our intuitive amplifier notations (Z_{in}, Z_{out}, A_0) added to indicate equivalencies. Let us first directly compute the closed-loop voltage gain. Recognizing that g_{11} is an admittance and g_{22} an impedance, we write a KVL around the input network and a KCL at the output node:

$$V_{in} = V_e + g_{22}\frac{V_e}{Z_{in}} + g_{21}V_{out} \tag{8.65}$$

$$g_{11}V_{out} + \frac{V_{out} - A_0V_e}{Z_{out}} = 0 \tag{8.66}$$

[13] Though allowing simpler algebra, the Y model does not provide intuitive results.



(a)

(b)

Figure 8.51 Voltage-voltage feedback circuit with (a) feedback network represented by a G model and (b) a simplified G model.

Finding V_e from the latter equation and substituting the result in the former, we have

$$\frac{V_{out}}{V_{in}} = \frac{A_0}{(1+\frac{g_{22}}{Z_{in}})(1+g_{11}Z_{out})+g_{21}A_0} \tag{8.67}$$

It is desirable to express the closed-loop gain in the familiar form, $A_{v,open}/(1+\beta A_{v,open})$. To this end, we divide the numerator and the denominator by $(1+g_{22}/Z_{in})(1+g_{11}Z_{out})$:

$$\frac{V_{out}}{V_{in}} = \frac{\dfrac{A_0}{(1+\frac{g_{22}}{Z_{in}})(1+g_{11}Z_{out})}}{1+g_{21}\dfrac{A_0}{(1+\frac{g_{22}}{Z_{in}})(1+g_{11}Z_{out})}} \tag{8.68}$$

We can thus write

$$A_{v,open} = \frac{A_0}{(1+\frac{g_{22}}{Z_{in}})(1+g_{11}Z_{out})} \tag{8.69}$$

$$\beta = g_{21} \tag{8.70}$$

Let us now interpret these results. The equivalent open-loop gain contains a factor A_0, i.e., the original amplifier's voltage gain (before immersion in feedback). But this gain is attenuated by two factors, namely, $1 + g_{22}/Z_{in}$ and $1 + g_{11}Z_{out}$. Interestingly, we can write $1 + g_{22}/Z_{in} = (Z_{in} + g_{22})/Z_{in}$, concluding that A_0 is multiplied by $Z_{in}/(Z_{in} + g_{22})$, which reminds us of a voltage divider. Similarly, $1 + g_{11}Z_{out} = (g_{11}^{-1} + Z_{out})/g_{11}^{-1}$, whose inverse points to another voltage divider. The loaded forward amplifier now emerges as shown in Fig. 8.52. Note that this model excludes the two generators $G_{12}V_{out}$ and $g_{12}I_{in}$, which are generally not negligible.

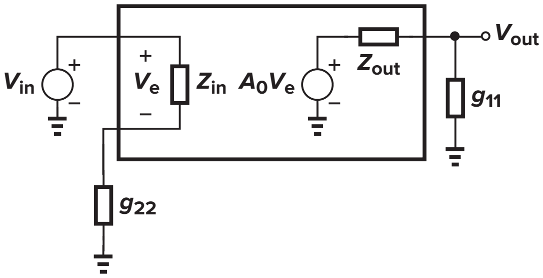


Figure 8.52 Proper method of including loading in a voltage-voltage feedback circuit.

The reader may wonder why we go to the trouble of finding the open-loop parameters while the closed-loop circuit in Fig. 8.51(a) can be solved exactly. The key principle here is that the rules depicted in Fig. 8.52 afford us a quick and intuitive understanding of the circuit that would not be possible from the direct analysis of Fig. 8.51(a). Specifically, we recognize that the finite input and output impedances of the feedback network reduce the output voltage and the voltage seen by the input of the main amplifier, respectively.

It is important to note that g_{11} and g_{22} in Fig. 8.50 are computed as follows:

$$g_{11} = \left.\frac{I_1}{V_1}\right|_{I2=0} \tag{8.71}$$

$$g_{22} = \left.\frac{V_2}{I_2}\right|_{V1=0} \tag{8.72}$$

Thus, as illustrated in Fig. 8.53, g_{11} is obtained by leaving the output of the feedback network open whereas g_{22} is calculated by *shorting* the input of the feedback network.

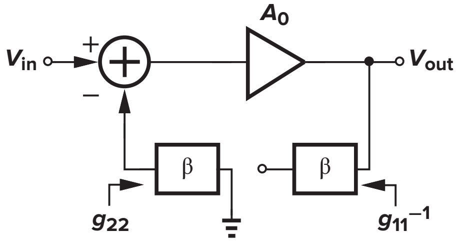


Figure 8.53 Conceptual view of opening a voltage-voltage feedback loop with proper loading.

Another important result of the foregoing analysis is that the loop gain, i.e., the second term in the denominator of (8.68), is simply equal to the loaded open-loop gain multiplied by g_{21}. Thus, a separate calculation of the loop gain is not necessary. Also, the open-loop input and output impedances obtained from Fig. 8.52 are scaled by $1 + g_{21}A_{v,open}$ to yield the closed-loop values. Again, we must bear in mind that this loop gain neglects the effect of $G_{12}V_{out}$ and $g_{12}I_{in}$.

▶ Example 8.12

For the circuit shown in Fig. 8.54(a), calculate the open-loop and closed-loop gains assuming $\lambda = \gamma = 0$.

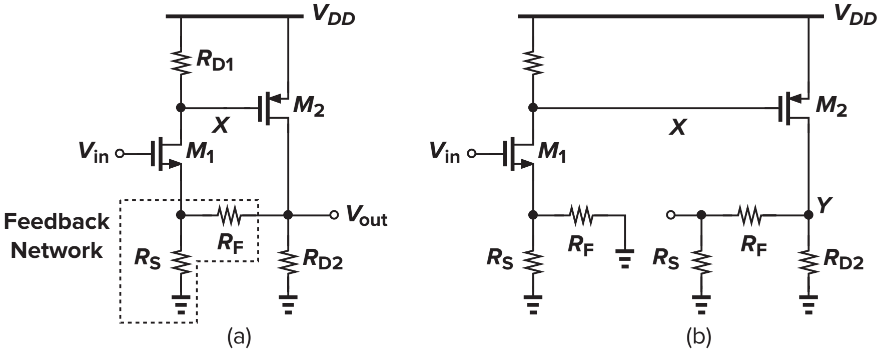


Figure 8.54

Solution

The circuit consists of two common-source stages, with R_F and R_S sensing the output voltage and returning a fraction thereof to the source of M_1. This transistor subtracts the returned voltage from V_{in}. The reader can prove that the feedback is indeed negative. Following the procedure illustrated in Fig. 8.53, we identify R_F and R_S as the feedback network and construct the open-loop circuit as shown in Fig. 8.54(b). Note that the loading effect in the input network is obtained by shorting the right terminal of R_F to ground and that in the output by leaving the left terminal of R_F open. Neglecting channel-length modulation and body effect for simplicity, we observe that M_1 is degenerated by the feedback network and

$$A_{v,open} = \frac{V_Y}{V_{in}} = \frac{-R_{D1}}{R_F \| R_S + 1/g_{m1}}\{-g_{m2}[R_{D2}\|(R_F + R_S)]\} \tag{8.73}$$

To compute the closed-loop gain, we first find the loop gain as $g_{21}A_{v,open}$. Recall from (8.64) that $g_{21} = V_2/V_1$ with $I_2 = 0$. For the voltage divider consisting of R_F and R_S, $g_{21} = R_S/(R_F + R_S)$. The closed-loop gain is simply equal to $A_{v,closed} = A_{v,open}/(1 + g_{21}A_{v,open})$.

Can we include R_{D2} in the feedback network rather than in the forward amplifier? Yes, we can ascribe a finite r_O to M_2 and proceed while considering R_{D2}, R_F, and R_S as the feedback network. The result is slightly different from that obtained above.

The above analysis neglects the forward amplifier's internal feedback (e.g., due to C_{GD2}) and the propagation of the input signal from the source of M_1 and through R_F to the output. (Transistor M_1 also operates as a source follower in this case.)

◀

▶ Example 8.13

A student eager to understand the approximations leading to the circuit in Fig. 8.51(b) decides to use an H model for the forward amplifier and obtain an exact solution. Perform this analysis and explain the results.

Solution

Illustrated in Fig. 8.55, this representation is attractive as it allows a simple series connection of voltages and impedances at the input and a parallel connection at the output. Writing a KVL and a KCL gives

$$V_{in} = I_{in}H_{11} + H_{12}V_{out} + I_{in}g_{22} + g_{21}V_{out} \tag{8.74}$$

$$H_{22}V_{out} + H_{21}I_{in} + g_{11}V_{out} + g_{12}I_{in} = 0 \tag{8.75}$$

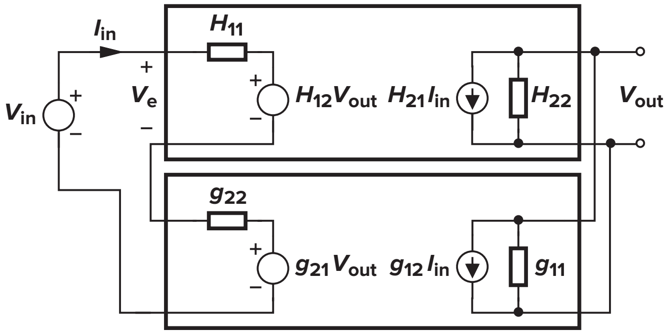


Figure 8.55

Finding I_{in} from the latter and replacing it in the former, we have

$$\frac{V_{out}}{V_{in}} = \frac{-\dfrac{H_{21}+g_{12}}{(H_{22}+g_{11})(H_{11}+g_{22})}}{1-(H_{12}+g_{21})\dfrac{H_{21}+g_{12}}{(H_{22}+g_{11})(H_{11}+g_{22})}} \tag{8.76}$$

We can thus define

$$A_{v,open} = -\frac{H_{21}+g_{12}}{(H_{22}+g_{11})(H_{11}+g_{22})} \tag{8.77}$$

$$\beta = H_{12}+g_{21} \tag{8.78}$$

If we assume that $g_{12} \ll H_{21}$ and $H_{12} \ll g_{21}$, then

$$A_{v,open} = \frac{-H_{21}}{(H_{22}+g_{11})(H_{11}+g_{22})} \tag{8.79}$$

$$\beta = g_{21} \tag{8.80}$$

and the attenuation factors $H_{22}+g_{11}$ and $H_{11}+g_{22}$ can be interpreted in the same manner as those in Eq. (8.69). This approach therefore explicitly reveals the simplifying approximations, namely, $g_{12} \ll H_{21}$ and $H_{12} \ll g_{21}$. Unfortunately, however, for a MOS gate input, H_{21} (the "current gain") approaches infinity, making the model difficult to use.

8.5.3 Loading in Current-Voltage Feedback

In this case, the feedback network appears in series with the output so as to sense the current. We represent the forward amplifier and the feedback network by Y and Z models, respectively (Fig. 8.56), neglecting the generators $Y_{12}V_{out}$ and $z_{12}I_{in}$. We wish to compute the closed-loop gain, I_{out}/V_{in}, and therefrom determine how the open-loop parameters can be obtained in the presence of loading. Noting that $I_{in} = Y_{11}V_e$ and $I_2 = I_{in}$, we write two KVLs:

$$V_{in} = V_e + Y_{11}V_e z_{22} + z_{21}I_{out} \tag{8.81}$$

$$-I_{out}z_{11} = \frac{I_{out}-Y_{21}V_e}{Y_{22}} \tag{8.82}$$

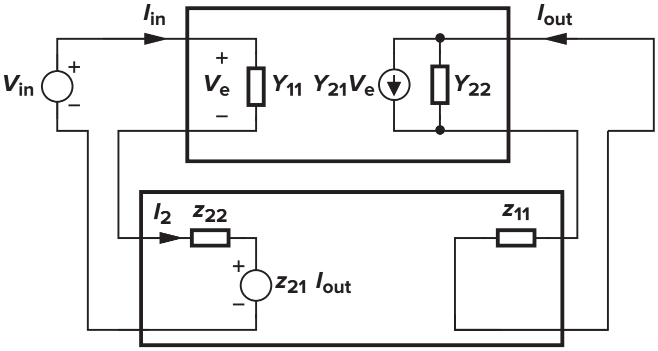


Figure 8.56 Current-voltage feedback circuit with loading.

Finding V_e from the latter and substituting in the former, we have

$$\frac{I_{out}}{V_{in}} = \frac{\dfrac{Y_{21}}{(1+z_{22}Y_{11})(1+z_{11}Y_{22})}}{1+z_{21}\dfrac{Y_{21}}{(1+z_{22}Y_{11})(1+z_{11}Y_{22})}} \tag{8.83}$$

We can thus visualize the open-loop gain and the feedback factor as

$$G_{m,open} = \frac{Y_{21}}{(1+z_{22}Y_{11})(1+z_{11}Y_{22})} \tag{8.84}$$

$$\beta = z_{21} \tag{8.85}$$

Note that Y_{21} is in fact the transconductance gain, G_m, of the original amplifier. The two attenuation factors $(1+z_{22}Y_{11})^{-1}$ and $(1+z_{11}Y_{22})^{-1}$ respectively correspond to voltage division at the input and current division at the output, allowing us to construct the loaded open-loop forward amplifier as shown in Fig. 8.57. Since $z_{22} = V_2/I_2$ with $I_1 = 0$ and $z_{11} = V_1/I_1$ with $I_2 = 0$, we arrive at the conceptual picture depicted in Fig. 8.58 for properly breaking the feedback. Note that the loop gain is equal to $z_{21}G_{m,open}$.

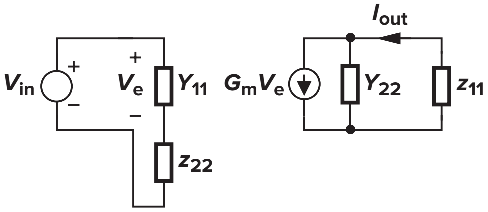


Figure 8.57 Current-voltage feedback circuit with proper loading of feedback network.

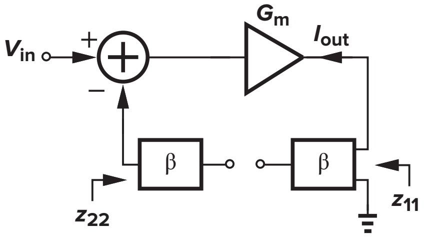


Figure 8.58 Conceptual view of opening the loop in current-voltage feedback.

▶ **Example 8.14**

A PMOS current source delivers a current to a load, e.g., the rechargeable battery in a cell phone [Fig. 8.59(a)]. We wish to make this current less PVT-dependent by means of negative feedback. As shown in Fig. 8.59(b), we convert the output current to voltage by a small series resistor, r_M, compare this voltage with a reference by means of an amplifier, and return the result to the gate of M_1. Determine the output current and the impedance seen by the load.

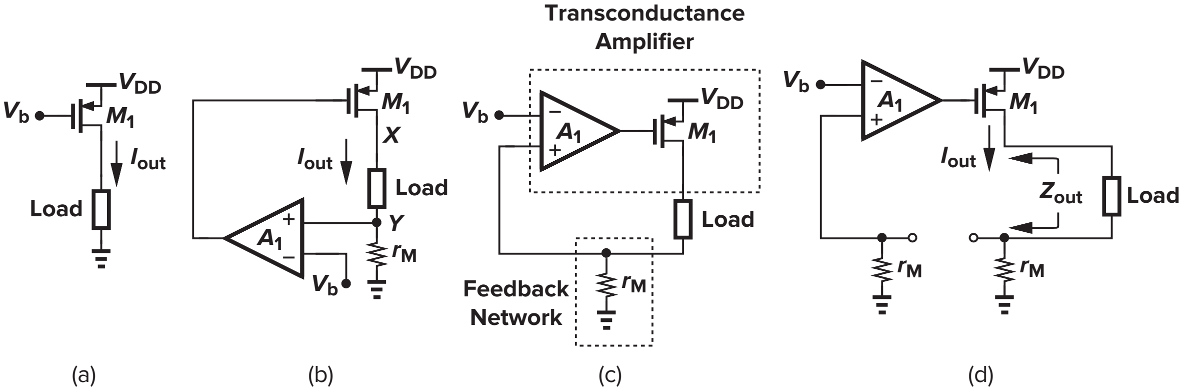


Figure 8.59

Solution

We view V_b as the input voltage and recognize that r_M sustains a voltage approximately equal to V_b if the loop gain is high. That is, $I_{out} \approx V_b/r_M$. But let us analyze this arrangement more accurately. Redrawing the circuit as in Fig. 8.59(c), we identify A_1 and M_1 as the forward transconductance amplifier and r_M as the feedback network. The procedure depicted in Fig. 8.58 leads to the open-loop topology of Fig. 8.59(d), and hence

$$G_{m,open} = \frac{I_{out}}{V_b} \tag{8.86}$$

$$\approx A_1 g_m \tag{8.87}$$

where the current flowing through r_O is neglected. The feedback factor $\beta = z_{21} = r_M$. Thus, the closed-loop output current is given by

$$I_{out} = \frac{A_1 g_m}{1 + A_1 g_m r_M} V_b \tag{8.88}$$

In the open-loop configuration, the load sees an impedance of $r_O + r_M$. Since feedback regulates the output current, the impedance seen by the load rises by a factor of $1 + A_1 g_m r_M$, reaching $Z_{out} = (1 + A_1 g_m r_M)(r_O + r_M)$.

A critical point emerging from this example is that the output impedance of a current-voltage feedback topology must be obtained by *breaking* the output current path and measuring the impedance between the resulting two nodes [e.g., X and Y in Fig. 8.59(b)]. In the above calculations, the "impedance seen by the load" is in fact computed by replacing the load with a voltage source and measuring the current through it. ◀

8.5.4 Loading in Voltage-Current Feedback

In this configuration, the forward (transimpedance) amplifier generates an output voltage in response to the input current and can thus be represented by a Z model. Sensing the output voltage and returning a proportional current, the feedback network lends itself to a Y model. The equivalent circuit is shown in Fig. 8.60, where the effect of Z_{12} and y_{12} is neglected. As in previous cases, we compute the closed-loop

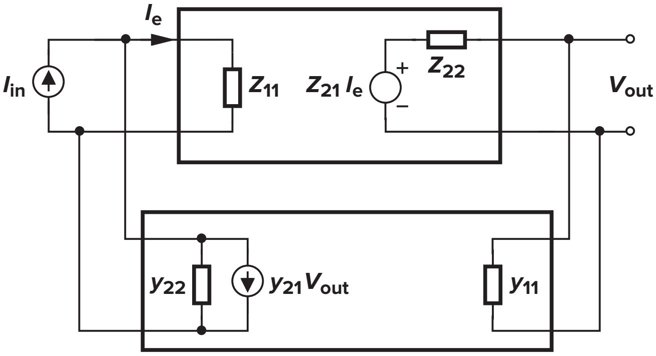


Figure 8.60 Voltage-current feedback circuit with loading.

gain, V_{out}/I_{in}, by writing two equations:

$$I_{in} = I_e + I_e Z_{11} y_{22} + y_{21} V_{out} \tag{8.89}$$

$$y_{11}V_{out} + \frac{V_{out} - Z_{21}I_e}{Z_{22}} = 0 \tag{8.90}$$

Eliminating I_e, we obtain

$$\frac{V_{out}}{I_{in}} = \frac{\dfrac{Z_{21}}{(1+y_{22}Z_{11})(1+y_{11}Z_{22})}}{1 + y_{21}\dfrac{Z_{21}}{(1+y_{22}Z_{11})(1+y_{11}Z_{22})}} \tag{8.91}$$

Thus, the equivalent open-loop gain and feedback factor are given by

$$R_{0,open} = \frac{Z_{21}}{(1+y_{22}Z_{11})(1+y_{11}Z_{22})} \tag{8.92}$$

$$\beta = y_{21} \tag{8.93}$$

Interpreting the attenuation factors in $R_{0,open}$ as current division at the input and voltage division at the output, we arrive at the conceptual view in Fig. 8.61. The loop gain is given by $y_{21}R_{0,open}$.

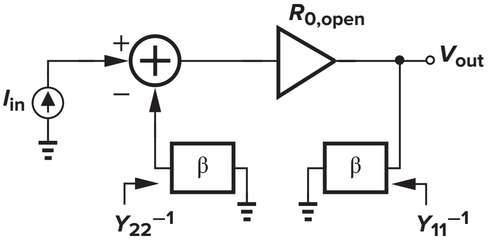


Figure 8.61 Conceptual view of opening the loop in voltage-current feedback.

▶ **Example 8.15**

Figure 8.62(a) shows a transimpedance amplifier topology commonly used in optical communication systems. Determine the circuit's gain and input and output impedances if $\lambda = 0$.

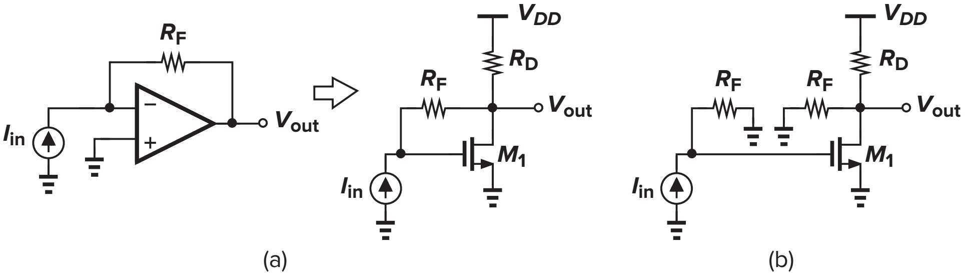


Figure 8.62

Solution

We can view the feedback resistor, R_F, as a network that senses the output voltage, converts it to current, and returns the result to the input. Following Figure 8.61, we construct the loaded open-loop amplifier as shown in Fig. 8.62(b), and express the open-loop gain as

$$R_{0,open} = -R_F g_m (R_F || R_D) \tag{8.94}$$

The feedback factor, y_{21} $(= I_2/V_1$ with $V_2 = 0)$ is equal to $-1/R_F$. It follows that the closed-loop gain is equal to

$$\frac{V_{out}}{I_{in}} = \frac{-R_F g_m (R_F || R_D)}{1 + g_m (R_F || R_D)} \tag{8.95}$$

which, if $g_m (R_F || R_D) \gg 1$, reduces to $-R_F$, an expected result (why?). The closed-loop input impedance is

$$R_{in} = \frac{R_F}{1 + g_m (R_F || R_D)} \tag{8.96}$$

which is approximately equal to $(1 + R_F/R_D)(1/g_m)$ if the above condition holds. Similarly, the closed-loop output impedance is given by

$$R_{out} = \frac{R_F || R_D}{1 + g_m (R_F || R_D)} \tag{8.97}$$

which amounts to $1/g_m$ if $g_m (R_F || R_D) \gg 1$. Note that if $\lambda > 0$, we can simply replace R_D with $R_D || r_O$ in all of the foregoing equations.

This transconductance amplifier is simple enough that we can solve it directly, and the reader is encouraged to do so. But we can readily identify two inconsistencies. First, breaking the loop at the gate of M_1 yields a loop gain of $g_m R_D$ rather than $g_m (R_D || R_F)$. Second, the closed-loop output impedance [with I_{in} set to zero in Fig. 8.62(a)] is simply equal to $R_D || (1/g_m) = R_D/(1 + g_m R_D)$. The value derived above can be expressed as $R_D/(1 + g_m R_D + R_D/R_F)$, revealing the extra term R_D/R_F. These errors arise from the approximate nature of the model.

▶ Example 8.16

Calculate the voltage gain of the circuit shown in Fig. 8.63(a).

Solution

What type of feedback is used in this circuit? Resistor R_F senses the output voltage and returns a proportional current to node X. Thus, the feedback can be considered as the voltage-current type. However, in the general representation of Fig. 8.60(a), the input signal is a current quantity, whereas in this example, it is a voltage quantity. For this reason, we replace V_{in} and R_S by a Norton equivalent [Fig. 8.63(b)] and view R_S as the input resistance of the main amplifier. Opening the loop according to Fig. 8.61 and neglecting channel-length modulation, we write the open-loop gain

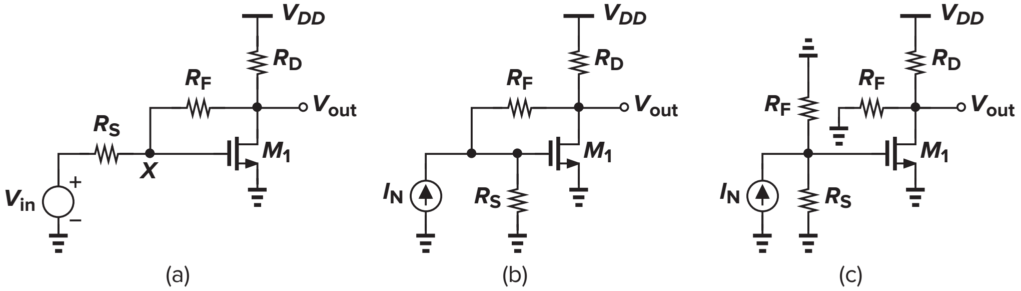


Figure 8.63

from Fig. 8.63(c) as

$$R_{0,open} = \frac{V_{out}}{I_N}\bigg|_{open} \tag{8.98}$$

$$= -(R_S \| R_F)g_m(R_F \| R_D) \tag{8.99}$$

where $I_N = V_{in}/R_S$. We also calculate the loop gain as $y_{21}R_{0,open}$. Thus, the circuit of Fig. 8.63(a) exhibits a voltage gain of

$$\frac{V_{out}}{V_{in}} = \frac{1}{R_S} \cdot \frac{-(R_S \| R_F)g_m(R_F \| R_D)}{1 + g_m(R_F \| R_D)R_S/(R_S + R_F)} \tag{8.100}$$

Interestingly, if R_F is replaced by a capacitor, this analysis does not yield a zero in the transfer function because we have neglected the reverse transmission of the feedback network (from the output of the feedback network to its input). The input and output impedances of the circuit are also interesting to calculate. This is left as an exercise for the reader. The reader is also encouraged to apply this solution to the circuit of Fig. 8.3(b).

8.5.5 Loading in Current-Current Feedback

The forward amplifier in this case generates an output current in response to the input current and can be represented by an H model, and so can the feedback network. Shown in Fig. 8.64 is the equivalent circuit with the H_{12} and h_{12} generators neglected. We write

$$I_{in} = I_e H_{11} h_{22} + h_{21} I_{out} + I_e \tag{8.101}$$

$$I_{out} = -I_{out} h_{11} H_{22} + H_{21} I_e \tag{8.102}$$

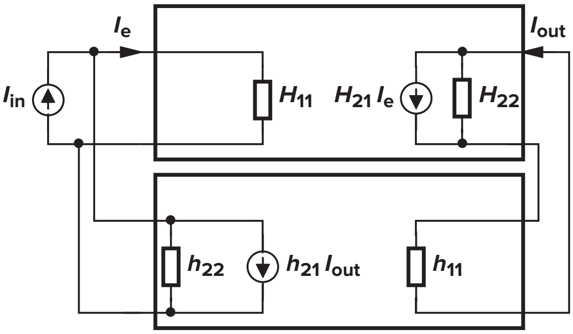


Figure 8.64 Equivalent circuit for current-current feedback.

and hence

$$\frac{I_{out}}{I_{in}} = \frac{\dfrac{H_{21}}{(1+h_{22}H_{11})(1+h_{11}H_{22})}}{1+h_{21}\dfrac{H_{21}}{(1+h_{22}H_{11})(1+h_{11}H_{22})}} \tag{8.103}$$

As in previous topologies, we define the equivalent open-loop current gain and the feedback factor as

$$A_{I,open} = \frac{H_{21}}{(1+h_{22}H_{11})(1+h_{11}H_{22})} \tag{8.104}$$

$$\beta = h_{21} \tag{8.105}$$

The conceptual view of the broken loop is depicted in Fig. 8.65, and the loop gain is equal to $h_{21}A_{I,open}$.

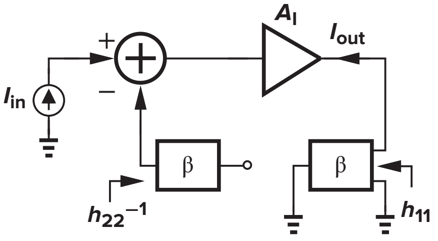


Figure 8.65 Conceptual view of loading in current-current feedback.

▶ Example 8.17

Calculate the open-loop and closed-loop gains of the circuit shown in Fig. 8.66(a). Assume that $\lambda = \gamma = 0$.

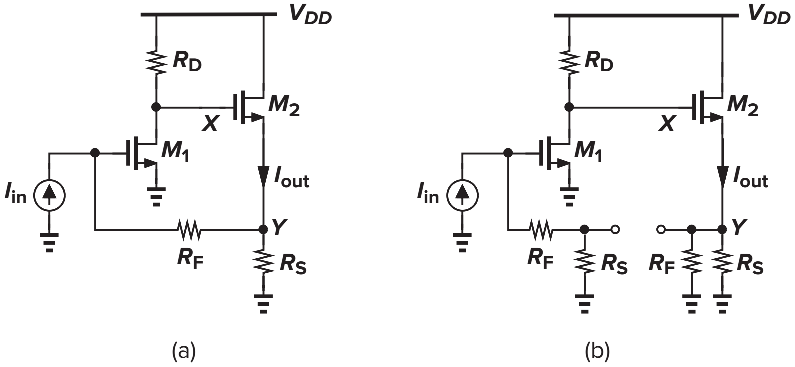


Figure 8.66

Solution

In this circuit, R_S and R_F sense the output current and return a fraction thereof to the input. Breaking the loop according to Fig. 8.65, we arrive at the circuit in Fig. 8.66(b), where we have

$$A_{I,open} = -(R_F + R_S)g_{m1}R_D\frac{1}{R_S\|R_F + 1/g_{m2}} \tag{8.106}$$

The loop gain is given by $h_{21}A_{I,open}$, where, from (8.62), $h_{21} = I_2/I_1$ with $V_2 = 0$. For the feedback network consisting of R_S and R_F, we have $h_{21} = -R_S/(R_S + R_F)$. The closed-loop gain equals $A_{I,open}/(1 + h_{21}A_{I,open})$. ◀

8.5.6 Summary of Loading Effects

The results of our study of loading are summarized in Fig. 8.67. The analysis is carried out in three steps: (1) open the loop with proper loading and calculate the open-loop gain, A_{OL}, and the open-loop input and output impedances; (2) determine the feedback ratio, β, and hence the loop gain, βA_{OL}; and (3) calculate the closed-loop gain and input and output impedances by scaling the open-loop values by a factor of $1 + \beta A_{OL}$. Note that in the equations defining β, the subscripts 1 and 2 refer to the input and output ports of the feedback network, respectively.

In this chapter, we have described two methods of obtaining the loop gain: (1) by breaking the loop at an arbitrary point, as shown in Fig. 8.5, and (2) by calculating A_{OL} and β, as illustrated in Fig. 8.67. The two methods may yield slightly different results due to the issues outlined in Table 8.1.

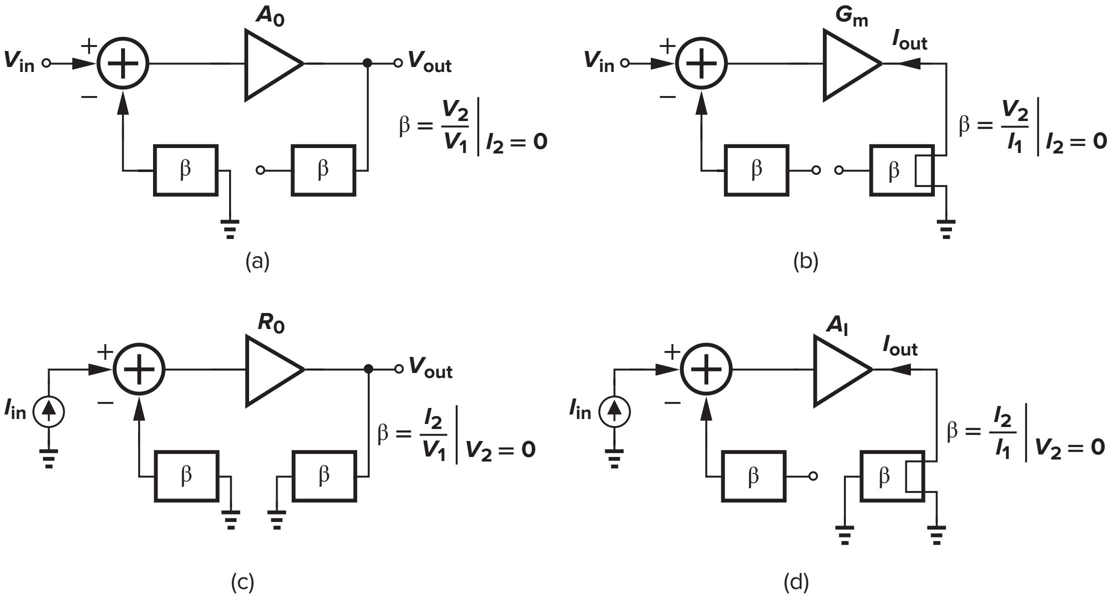


Figure 8.67 Summary of loading effects.

8.6 ■ Bode's Analysis of Feedback Circuits

Bode's approach provides a rigorous solution for a circuit's closed-loop parameters (whether it includes feedback or not), but it does not tell us much about the loop gain in the presence of multiple feedback mechanisms. The analysis presented in this section was originally described by Bode in his 1945 classic textbook *Network Analysis and Feedback Network Design*. Since this approach is somewhat less intuitive, we encourage the reader to be patient and read this section in several sittings.

8.6.1 Observations

Before delving into Bode's analysis, we should make two simple, yet new observations with regard to circuit equations.

First, consider the general circuit shown in Fig. 8.68(a), where one transistor is explicitly shown in its ideal form. We know from our small-signal gain and transfer function analyses in previous chapters that V_{out} can eventually be expressed as $A_v V_{in}$ or $H(s)V_{in}$. But, what happens if we denote the dependent current source by I_1 and do not make the substitution $I_1 = g_m V_1$ yet? Then, V_{out} is obtained as a function

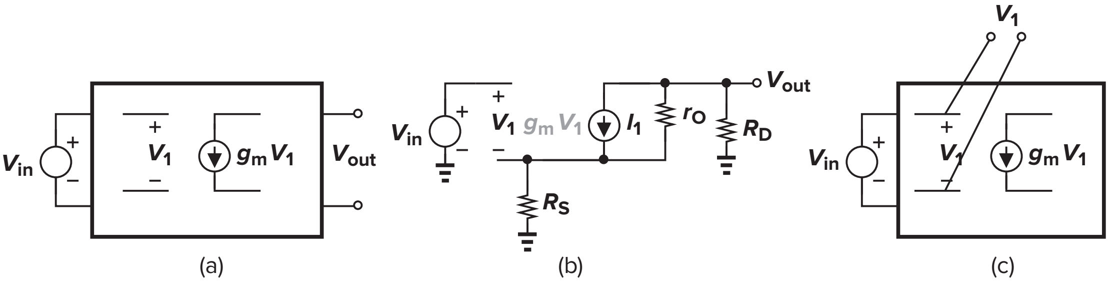


Figure 8.68 (a) Circuit containing a dependent source, (b) circuit example, and (c) V_1 as a signal of interest.

of both V_{in} and I_1:

$$V_{out} = AV_{in} + BI_1 \tag{8.107}$$

As an example, in the degenerated common-source stage of Fig. 8.68(b), we note that the current flowing upward through R_D (and downward through R_S) is equal to $-V_{out}/R_D$, and hence the voltage drop across r_O is given by $(-V_{out}/R_D - I_1)r_O$. A KVL around the output network thus yields

$$V_{out} = \left(-\frac{V_{out}}{R_D} - I_1\right) r_O - \frac{V_{out}}{R_D} R_S \tag{8.108}$$

and

$$V_{out} = \frac{-r_O}{1 + \dfrac{r_O + R_S}{R_D}} I_1 \tag{8.109}$$

In this case, $A = 0$ and $B = -r_O R_D/(R_D + r_O + R_S)$.

Second, let us return to the general circuit in Fig. 8.68(a) and consider V_1 as the signal of interest, i.e., we wish to compute V_1 as a function of V_{in} in the form of $A_v V_{in}$ or $H(s)V_{in}$. This is always possible by pretending that V_1 is the "output," as conceptually illustrated in Fig. 8.68(c). In a manner similar to Eq. (8.107), V_1 can be written as

$$V_1 = CV_{in} + DI_1 \tag{8.110}$$

if we temporarily forget that $I_1 = g_m V_1$. In Fig. 8.68(b), for example, we express the current though R_S (and R_D) as $(V_{in} - V_1)R_S$, subtract this current from I_1, and let the result flow through r_O. A KVL around the output network gives

$$V_{in} - V_1 - \left(I_1 - \frac{V_{in} - V_1}{R_S}\right) r_O = -\frac{V_{in} - V_1}{R_S} R_D \tag{8.111}$$

and hence

$$V_1 = V_{in} - \frac{r_O R_S}{R_D + r_O + R_S} I_1 \tag{8.112}$$

That is, $C = 1$ and $D = -r_O R_S/(R_D + r_O + R_S)$.

In summary, in a given circuit containing at least one transistor (whether there is feedback or not), we can eventually reach two equations that express V_{out} and V_1 in terms of V_{in} and I_1. To obtain V_{out}/V_{in}, we solve the two equations while applying the knowledge that I_1 is in fact equal to $g_m V_1$.

The foregoing developments and, in particular, Eqs. (8.107) and (8.110) appear unnecessarily tedious. After all, we can directly solve the circuit in Fig. 8.68(b) with less algebra. However, the interpretation of the coefficients A, B, C, and D affords a simple and elegant approach to feedback analysis.

8.6.2 Interpretation of Coefficients

We now focus on Eqs. (8.107) and (8.110) and ask whether the A–D coefficients can be directly calculated for a given circuit. We begin with A:

$$A = \frac{V_{out}}{V_{in}} \text{ with } I_1 = 0 \tag{8.113}$$

This result implies that A is obtained as the voltage gain of the circuit if the dependent current source is set to zero, which can be readily accomplished by "disabling" the transistor, i.e., by forcing the transistor's g_m to zero. We can consider V_{out} in this case as the "feedthrough" of the input signal (in the absence of the ideal transistor) [Fig. 8.69(a)]. In the CS example, $V_{out} = 0$ if $I_1 = 0$ because no current flows through R_S, r_O, and R_D. That is, $A = 0$.

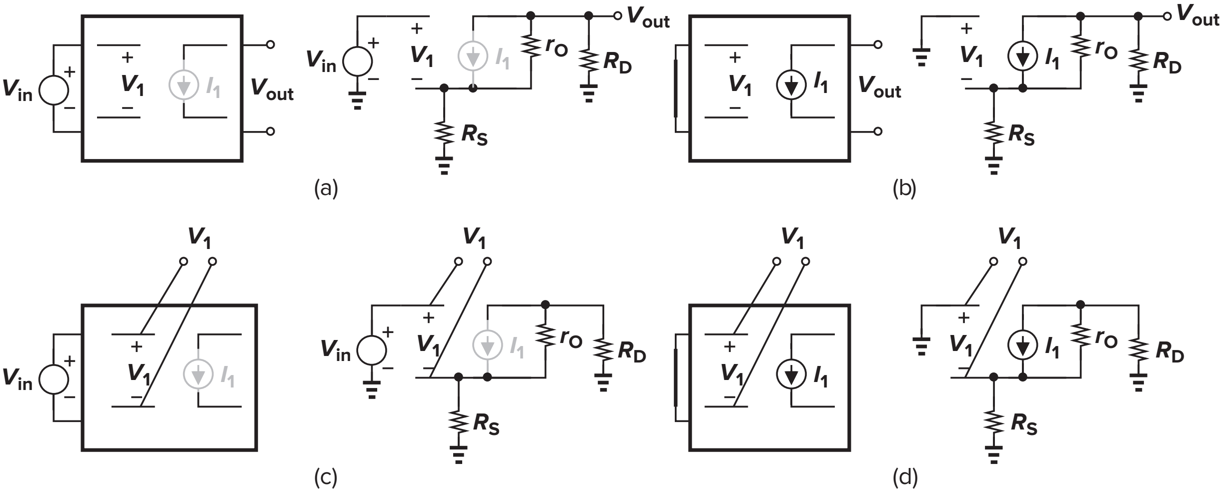


Figure 8.69 Setups for the calculation of (a) A, (b) B, (c) C, and (d) D.

As for the B coefficient in (8.107), we have

$$B = \frac{V_{out}}{I_1} \text{ with } V_{in} = 0 \tag{8.114}$$

That is, we set the input to zero and compute V_{out} as a result of I_1 [Fig. 8.69(b)], pretending that I_1 is an independent source.[14] In the CS example,

$$\left(-\frac{V_{out}}{R_D} - I_1\right) r_O - \frac{V_{out}}{R_D} R_S = V_{out} \tag{8.115}$$

[14] If I_1 is kept as a dependent source, the circuit has no external stimulus and, therefore, generates no voltage or current.

and hence

$$V_{out} = \frac{-r_O R_D}{R_D + r_O + R_S} I_1 \tag{8.116}$$

Thus, $B = -r_O R_D/(R_D + r_O + R_S)$.

The C coefficient in (8.110) is interpreted as

$$C = \frac{V_1}{V_{in}} \text{ with } I_1 = 0 \tag{8.117}$$

i.e., the transfer function from the input to V_1 with the transistor's g_m set to zero [Fig. 8.69(c)]. In the CS circuit, no current flows through R_S under this condition, yielding $V_1 = V_{in}$ and $C = 1$.

Finally, the D coefficient is obtained as

$$D = \frac{V_1}{I_1} \text{ with } V_{in} = 0 \tag{8.118}$$

which, as illustrated in Fig. 8.69(d), represents the transfer function from I_1 to V_1 with the input at zero. In the CS stage, the current flowing through R_S (and R_D) under this condition is equal to $-V_1/R_S$, producing a voltage drop of $(-V_1/R_S - I_1)r_O$ across r_O. A KVL around the output network yields

$$-V_1 - \left(\frac{V_1}{R_S} + I_1\right) r_O = \frac{V_1}{R_S} R_D \tag{8.119}$$

We therefore have

$$V_1 = -\frac{r_O R_S}{R_D + r_O + R_S} I_1 \tag{8.120}$$

and hence $D = -r_O R_S/(R_D + r_O + R_S)$.

In summary, the A–D coefficients are computed as shown in Fig. 8.70: (1) we disable the transistor by setting its g_m to zero and obtain A and C as the feedthroughs from V_{in} to V_{out} and to V_1, respectively, and (2) we set the input to zero and calculate B and D as the gain from I_1 to V_{out} and to V_1, respectively. From another perspective, the former step finds the responses to V_{in} with $g_m = 0$, and the latter, to I_1 with $V_{in} = 0$. We can even say that the circuit is excited each time by *one* input, either V_{in} or I_1, and generates *two* outputs of interest, V_{out} and V_1. The reader may still not see the reason for these derivations, but patience is a virtue!

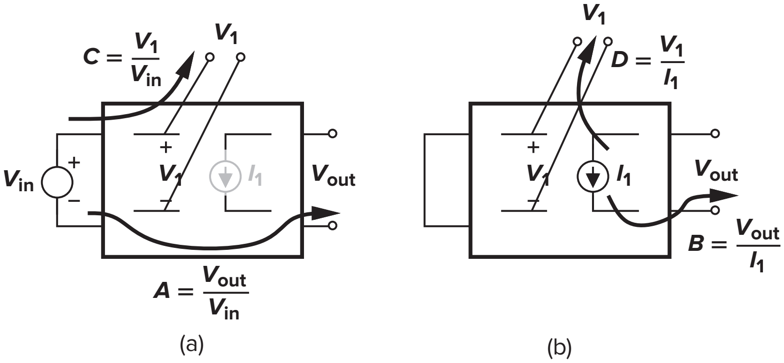


Figure 8.70 Summary of computations for A–D.

▶ Example 8.18

Compute the A–D coefficients for the circuit shown in Fig. 8.71(a).

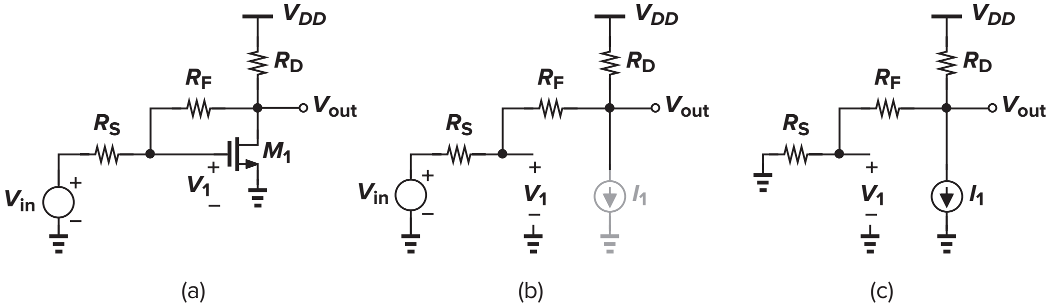


Figure 8.71

Solution

Following the procedures illustrated in Fig. 8.70, we first set I_1 (i.e., g_m) to zero and determine the feedthrough components V_{out}/V_{in} and V_1/V_{in}. From Fig. 8.71(b), we have

$$A = \frac{V_{out}}{V_{in}} \tag{8.121}$$

$$= \frac{R_D}{R_D + R_S + R_F} \tag{8.122}$$

and

$$C = \frac{V_1}{V_{in}} \tag{8.123}$$

$$= \frac{R_F + R_D}{R_D + R_S + R_F} \tag{8.124}$$

Next, we set V_{in} to zero and calculate the transfer functions from I_1 to V_{out} and to V_1 [Fig. 8.71(c)]:

$$B = \frac{V_{out}}{I_1} \tag{8.125}$$

$$= -R_D||(R_S + R_F) \tag{8.126}$$

$$= -\frac{R_D(R_S + R_F)}{R_D + R_S + R_F} \tag{8.127}$$

and

$$D = \frac{V_1}{I_1} \tag{8.128}$$

$$= \frac{R_S}{R_S + R_F}\frac{V_{out}}{I_1} \tag{8.129}$$

$$= -\frac{R_S R_D}{R_D + R_S + R_F} \tag{8.130}$$

◀

For our subsequent studies, we must refresh our memory about loop gain calculations.

▶ **Example 8.19**

Determine the exact loop gain for the circuit of Fig. 8.71(a).

Solution

We prefer to break the loop at a port that does not entail loading effects. Let us do so at the gate of M_1, as depicted in Fig. 8.72(a). Applying a test voltage, V_t, and calculating the feedback voltage, V_F, we have

$$\text{Loop Gain} = -\frac{V_F}{V_t} \tag{8.131}$$

$$= g_m[R_D||(R_S + R_F)]\frac{R_S}{R_S + R_F} \tag{8.132}$$

$$= \frac{g_m R_S R_D}{R_D + R_S + R_F} \tag{8.133}$$

Note that the loop gain and the D coefficient in (8.130) differ by only a factor of $-g_m$. We return to this point below.

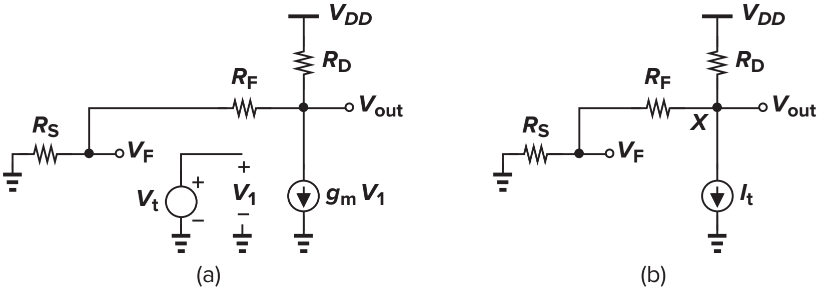


Figure 8.72

Alternatively, we can break the loop at the top terminal of the dependent current source. Illustrated in Fig. 8.72(b), the idea is to draw a test current, I_t, from node X and measure the resulting feedback voltage, V_F, recognizing that the ratio $-V_F/I_t$ must be multiplied by g_m to arrive at the loop gain:

$$V_F = -I_t[R_D||(R_S + R_F)]\frac{R_S}{R_S + R_F} \tag{8.134}$$

and thus

$$\text{Loop Gain} = -\frac{g_m V_F}{I_t} \tag{8.135}$$

$$= \frac{g_m R_S R_D}{R_D + R_S + R_F} \tag{8.136}$$

We see a similarity between the calculation of D in Fig. 8.69(d) and the calculation of the loop gain in Fig. 8.72(b). In both cases, we set the input to zero, apply I_1 or I_t, and measure the controlling voltage, V_1. We therefore surmise that D and the loop gain may be related. We will keep the reader in suspense for now. ◀

8.6.3 Bode's Analysis

We have seen in the previous section that the A–D coefficients can be computed relatively easily. We now express V_{out}/V_{in} in terms of these coefficients. Since

$$V_{out} = AV_{in} + BI_1 \tag{8.137}$$

$$V_1 = CV_{in} + DI_1 \tag{8.138}$$

and, in the actual circuit, $I_1 = g_m V_1$, we have

$$V_1 = \frac{C}{1 - g_m D} V_{in} \tag{8.139}$$

The closed-loop gain is therefore equal to

$$\frac{V_{out}}{V_{in}} = A + \frac{g_m BC}{1 - g_m D} \tag{8.140}$$

As expected, the first term represents the input-output feedthrough, manifesting itself when $g_m = 0$. We can also write

$$\frac{V_{out}}{V_{in}} = \frac{A + g_m(BC - AD)}{1 - g_m D} \tag{8.141}$$

In contrast to direct analysis of the closed-loop circuit, Bode's method decomposes the computation into several simpler steps. While our formulation has assumed a dependent current source, the results are applicable to dependent voltage sources as well. Let us solve some circuits using Bode's approach.

▶ **Example 8.20**

Determine the voltage gain of the degenerated CS stage shown in Fig. 8.69.

Solution

Utilizing the results obtained for Fig. 8.69 and noting that $A = 0$ and $C = 1$, we have

$$\frac{V_{out}}{V_{in}} = \frac{g_m \dfrac{-r_O R_D}{R_D + r_O + R_S}}{1 + g_m \dfrac{r_O R_S}{R_D + r_O + R_S}} \tag{8.142}$$

$$= \frac{-g_m r_O R_D}{R_D + r_O + (1 + g_m r_O) R_S} \tag{8.143}$$

The reader is encouraged to repeat this analysis in the presence of body effect. ◀

▶ **Example 8.21**

Determine the voltage gain of the feedback amplifier shown in Fig. 8.71(a) without breaking the loop.

Solution

With the aid of the results obtained in Example 8.18, we obtain

$$\frac{V_{out}}{V_{in}} = \frac{R_D}{R_D + R_S + R_F} + \frac{-g_m \dfrac{R_D(R_S + R_F)(R_F + R_D)}{(R_D + R_S + R_F)^2}}{1 + \dfrac{g_m R_S R_D}{R_D + R_S + R_F}} \tag{8.144}$$

$$= \frac{R_D}{R_D + R_S + R_F} + \frac{-g_m R_D(R_S + R_F)(R_F + R_D)}{(R_D + R_S + R_F + g_m R_S R_D)(R_D + R_S + R_F)} \tag{8.145}$$

Note that this result is exact, with the first term representing the circuit's direct feedthrough in the absence of transistor action ($g_m = 0$).

Under what condition does the above loop gain reduce to the familiar, ideal form $-R_F/R_S$? We may surmise that R_D must be small enough not to "feel" the loading effect of R_F. But the condition $R_D \ll R_F$ does not yield a voltage gain of $-R_F/R_S$. After all, this ideal value also presumes a high *open-loop* gain. Thus, we need two conditions, namely, $R_D \ll R_F$ and $g_m R_D \gg 1$ for the above result to reduce to $-R_F/R_S$.

Let us make a useful observation. If $A = 0$, Eq. (8.140) yields $V_{out}/V_{in} = g_m BC/(1 - g_m D)$, a result resembling the generic feedback equation $A_0/(1 + \beta A_0)$. We therefore loosely call $g_m BC$ the "open-loop" gain.

Return Ratio and Loop Gain As mentioned in Example 8.19, the quantity D ($= V_1/I_1$ with $V_{in} = 0$) and the loop gain appear to be related. In fact, the closed-loop gain expression in Eq. (8.141) may suggest that $1 - g_m D = 1 +$ loop gain, and hence loop gain $= -g_m D$. This is not a coincidence: in both cases, we set the main input to zero, break the loop by replacing the dependent source with an independent source, and compute the returned quantity.

In his original treatment of feedback, Bode introduces the term "return ratio" (RR) to refer to $-g_m D$ and ascribes it to a given dependent source in the circuit [1]. Thus, the return ratio, obtained by injecting a voltage in place of V_{GS} or a current in place of I_D, appears to be the same as the true loop gain[15] even if the loop cannot be completely broken. In fact, the return ratio is equal to the loop gain if the circuit contains only one feedback mechanism and the loop traverses the transistor of interest. We elaborate on this point later.

▶ Example 8.22

Determine the voltage gain of the source follower shown in Fig. 8.73(a) using Bode's method. Assume that $\lambda = \gamma = 0$.

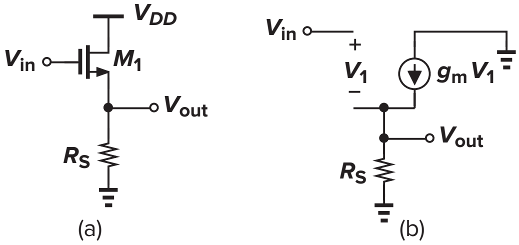


Figure 8.73

Solution

Figure 8.73(b) depicts the small-signal model. To compute the A and C coefficients, Fig. 8.70 suggests setting g_m to zero, which results in

$$A = \frac{V_{out}}{V_{in}} = 0 \tag{8.146}$$

$$C = \frac{V_1}{V_{in}} = 1 \tag{8.147}$$

For the B and and D coefficients, we set V_{in} to zero and apply a current source I_1 in lieu of $g_m V_1$:

$$B = \frac{V_{out}}{I_1} = R_S \tag{8.148}$$

$$D = \frac{V_1}{I_1} = -R_S \tag{8.149}$$

[15] By the true loop gain, we mean one that is obtained without any approximations, e.g., without neglecting loading or the propagation of the input signal through the feedback network to the main output.

From (8.140) or (8.141), we have

$$\frac{V_{out}}{V_{in}} = \frac{g_m R_S}{1 + g_m R_S} \tag{8.150}$$

The return ratio associated with the dependent source is equal to $-g_m D = g_m R_S$.

A peculiar result occurs here if R_S approaches an ideal current source: the return ratio, $g_m R_S$, goes to infinity, and so does B. Since (8.140) was obtained by dividing by B and D, in general it may give an incorrect value if B or D is infinite. In the case of the source follower, however, (8.140) produces a correct result.

▶ Example 8.23

Figure 8.74(a) shows a circuit in which one transistor, M_1, resides outside the feedback loop. Using Bode's method, compute V_{out}/V_{in}.

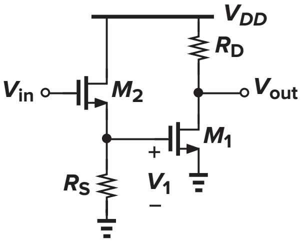


Figure 8.74

Solution

We first obtain A and C by setting g_{m1} to zero:

$$A = \frac{V_{out}}{V_{in}} = 0 \tag{8.151}$$

$$C = \frac{V_1}{V_{in}} = \frac{g_{m2} R_S}{1 + g_{m2} R_S} \tag{8.152}$$

Next, we set V_{in} to zero and apply I_1 in lieu of M_1:

$$B = \frac{V_{out}}{I_1} = -R_D \tag{8.153}$$

$$D = \frac{V_1}{I_1} = 0 \tag{8.154}$$

As expected, the return ratio for M_1 is zero. We thus have

$$\frac{V_{out}}{V_{in}} = g_{m1}(-R_D \frac{g_{m2} R_S}{1 + g_{m2} R_S}) \tag{8.155}$$

Alternatively, the gain can be obtained by treating M_2 as the dependent source of interest. The return ratio for M_2 is the same as that found for the source follower in the above example. Even though the circuit contains one feedback mechanism, the two return rations are unequal because the feedback loop does not traverse M_1.

▶ Example 8.24

Calculate the closed-loop gain of the circuit shown in Fig. 8.75(a). Assume that $\lambda = \gamma = 0$.

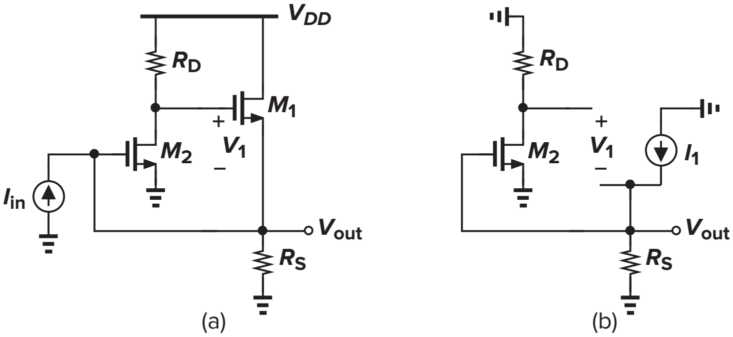


Figure 8.75

Solution

We calculate the A–D coefficients with the aid of the conceptual diagram in Fig. 8.70. We can select either transistor as the device of interest. Setting g_{m1} to zero, we obtain

$$A = \frac{V_{out}}{I_{in}} \text{ with } g_{m1} = 0 \tag{8.156}$$

$$= R_S \tag{8.157}$$

because, in the absence of M_1, I_{in} simply flows through R_S, producing a feedthrough component at the output. For C, we note that $V_1 = I_{in}R_S(-g_{m2}R_D) - I_{in}R_S$, and hence

$$C = \frac{V_1}{I_{in}} \text{ with } g_{m1} = 0 \tag{8.158}$$

$$= -(1 + g_{m2}R_D)R_S \tag{8.159}$$

We now set I_{in} to zero and inject an independent current source in place of M_1, as shown in Fig. 8.75(b). Since $V_{out} = I_1 R_S$,

$$B = \frac{V_{out}}{I_1} \text{ with } I_{in} = 0 \tag{8.160}$$

$$= R_S \tag{8.161}$$

Also, $V_1 = I_1 R_S(-g_{m2}R_D) - I_1 R_S = -I_1 R_S(1 + g_{m2}R_D)$ and

$$D = \frac{V_1}{I_1} \text{ with } I_{in} = 0 \tag{8.162}$$

$$= -R_S(1 + g_{m2}R_D) \tag{8.163}$$

Equation (8.140) thus gives

$$\frac{V_{out}}{I_{in}} = A + \frac{g_{m1}BC}{1 - g_{m1}D} \tag{8.164}$$

$$= R_S - \frac{g_{m1}(1 + g_{m2}R_D)R_S^2}{1 + g_{m1}R_S(1 + g_{m2}R_D)} \tag{8.165}$$

$$= \frac{R_S}{1 + g_{m1}R_S(1 + g_{m2}R_D)} \tag{8.166}$$

The reader is encouraged to repeat the derivation with M_2 as the dependent source of interest.

8.6.4 Blackman's Impedance Theorem

Continuing our effort to compute the closed-loop parameters of a feedback system without breaking the loop, we now study Blackman's theorem, which determines the impedance seen at any port of a general circuit. This theorem can be proved using Bode's approach.

Consider the general circuit depicted in Fig. 8.76(a), where the impedance between nodes P and Q is of interest. As in Bode's analysis, we have explicitly shown one of the transistors by its ideal model, the voltage-dependent current source I_1. Let us pretend that I_{in} is the input signal and V_{in} the *output* signal so that we can utilize Bode's results:

$$V_{in} = AI_{in} + BI_1 \tag{8.167}$$

$$V_1 = CI_{in} + DI_1 \tag{8.168}$$

It follows that

$$Z_{in} = \frac{V_{in}}{I_{in}} = A + \frac{g_m BC}{1 - g_m D} \tag{8.169}$$

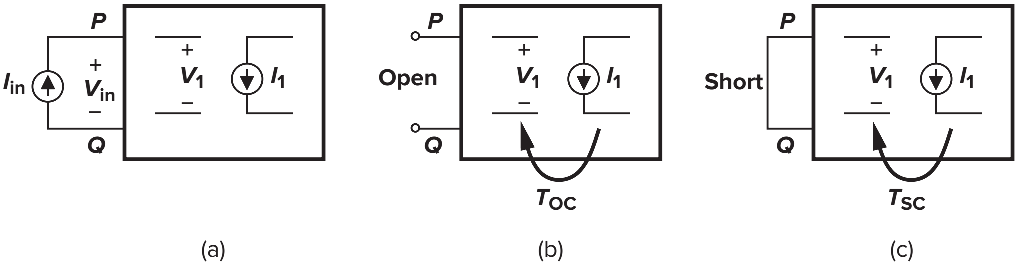


Figure 8.76 (a) Arrangement for calculating a port impedance, (b) calculation of T_{oc}, and (c) calculation of T_{sc}.

where g_m denotes the transconductance of the transistor modeled by I_1 in Fig. 8.76(a). We now manipulate this result in three steps so as to obtain a more intuitive expression. First, we recognize from (8.168) that, if $I_{in} = 0$, then $V_1/I_1 = D$. We call $-g_m D$ the "open-circuit loop gain" (because the port of interest is left *open*) and denote it by T_{oc} [Fig. 8.76(b)]. Second, we note from (8.167) that, if $V_{in} = 0$, then $I_{in} = (-B/A)I_1$, and hence, from (8.168),

$$\frac{V_1}{I_1} = \frac{AD - BC}{A} \tag{8.170}$$

We call $-g_m$ times this quantity the "short-circuit" loop gain (because $V_{in} = 0$) and denote it by T_{sc} [Fig. 8.76(c)]. Note that the circuit topology changes in these two cases. Both T_{oc} and T_{sc} can be viewed as return ratios associated with I_1 for the two circuit topologies. In summary,

$$T_{oc} = -g_m \frac{V_1}{I_1}|_{Iin=0} \tag{8.171}$$

$$T_{sc} = -g_m \frac{V_1}{I_1}|_{Vin=0} \tag{8.172}$$

In the third step, we use T_{oc} and T_{sc} to rewrite Eq. (8.169) as

$$Z_{in} = \frac{V_{in}}{I_{in}} = \frac{A - g_m(BC - AD)}{1 - g_m D} \tag{8.173}$$

$$= A\frac{1 + T_{sc}}{1 + T_{oc}} \tag{8.174}$$

Originally derived by Blackman [2], this result lends itself to a great deal of intuition if we recall that $A = V_{in}/I_{in}$ with $I_1 = 0$, i.e., when the transistor under consideration is disabled. We roughly view A as the "open-loop" impedance because it is obtained without the transistor in the feedback loop. In addition, we observe that (1) if $|T_{sc}| \ll 1$, then $Z_{in} \approx A/(1 + T_{oc})$; that is, the open-loop impedance is divided by $1 + T_{oc}$; and (2) if $|T_{oc}| \ll 1$, then $Z_{in} \approx A(1 + T_{sc})$; i.e., the open-loop impedance is multiplied by $1 + T_{sc}$. Reminiscent of closed-loop input and output impedances derived in previous sections, these two cases nonetheless reveal that, in general, the closed-loop impedance *cannot* be expressed as Z_{in} multiplied or divided by (1 + the loop gain).

▶ Example 8.25

Determine the output impedance of a degenerated CS stage [Fig. 8.77(a)]. Assume that $\gamma = 0$.

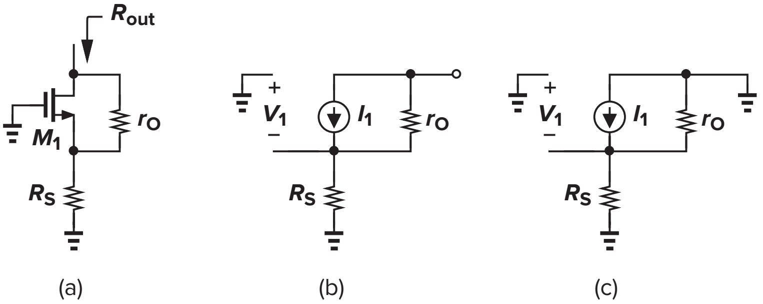


Figure 8.77

Solution

We must compute three quantities. First, with the transistor disabled,

$$A = r_O + R_S \tag{8.175}$$

Second, with the port of interest left open [Fig. 8.77(b)], we have

$$T_{oc} = -g_m\frac{V_1}{I_1} \tag{8.176}$$

$$= 0 \tag{8.177}$$

because no current flows through R_S. Third, with the port of interest shorted [Fig. 8.77(c)], we obtain

$$T_{sc} = -g_m\frac{V_1}{I_1} \tag{8.178}$$

$$= +g_m(R_S||r_O) \tag{8.179}$$

It follows from Eq. (8.174) that

$$Z_{out} = (r_O + R_S)[1 + g_m(R_S||r_O)] \tag{8.180}$$

$$= (1 + g_m r_O)R_S + r_O \tag{8.181}$$

The reader is encouraged to repeat the analysis while including body effect. ◀

▶ Example 8.26

Compute the output impedance of the circuit shown in Fig. 8.78(a). Assume that $\gamma = 0$.

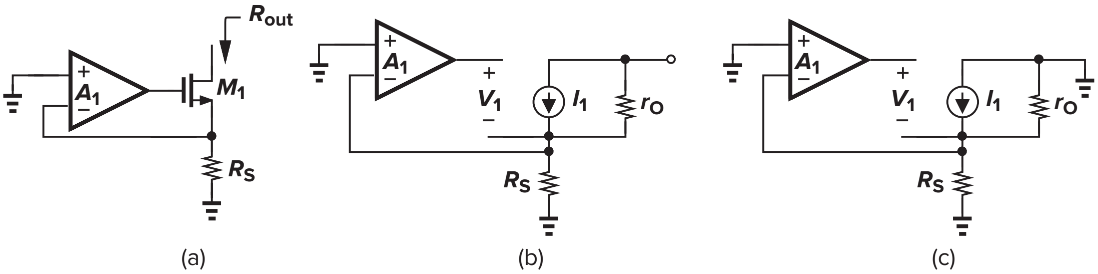


Figure 8.78

Solution

The difficulty with this circuit is that it does not map into one of the four canonical topologies: amplifier A_1 senses the voltage at the *source* of M_1 whereas the output is taken at the drain. Fortunately, Blackman's theorem is impervious to such departures. Again, we proceed in three steps. With the transistor disabled,

$$A = r_O + R_S \tag{8.182}$$

If the output is left open [Fig. 8.78(b)], no current flows through R_S, and hence $T_{oc} = 0$. With the output shorted [Fig. 8.78(c)],

$$T_{sc} = g_m(R_S||r_O)A_1 \tag{8.183}$$

Thus,

$$\begin{aligned} Z_{out} &= (r_O + R_S)[1 + g_m(R_S||r_O)A_1] && (8.184)\\ &= r_O + R_S + g_m r_O R_S A_1 && (8.185)\\ &= (1 + g_m r_O)A_1 R_S + r_O && (8.186)\end{aligned}$$

To the first order, the factor $1 + g_m r_O$ is "boosted" by another factor of A_1. ◀

▶ Example 8.27

Determine the output impedance of the source follower shown in Fig. 8.79(a). Assume that $\lambda = \gamma = 0$.

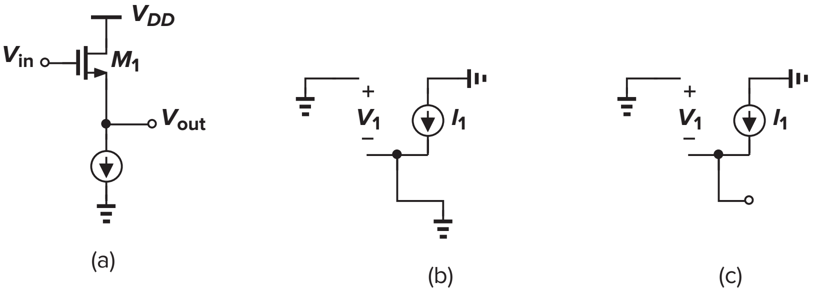


Figure 8.79

Solution

With $g_m = 0$, the output impedance, and hence A, are *infinity*. The two loop gains are obtained from Figs. 8.79(b) and (c) as $T_{sc} = 0$ and $T_{oc} = \infty$, respectively. These difficulties arise because the proof of Blackman's theorem divides by A, tacitly assuming that $A < \infty$. One can avoid this situation by placing a resistor in parallel with the port of interest and letting it approach infinity in the end result. This is left as an exercise for the reader.

▶ Example 8.28

Using Blackman's theorem, determine the input impedance of the circuit shown in Fig. 8.37(a). Assume that $\lambda = \gamma = 0$.

Solution

We set g_{m2} to zero to compute A, observing that $A = \infty$! Since the derivation of Blackman's expression relies on dividing by A, we know that $A = \infty$ may invalidate the result. This is one drawback of Blackman's approach. The situation becomes even more interesting if we attempt to compute T_{oc}. As depicted in Fig. 8.80, we apply an independent small-signal circuit source I_1 and seek V_1. The voltage at the gate of M_1 is equal to $-I_1 R_D C_1/(C_1+C_2)$, yielding a drain current of $-g_{m1} I_1 R_D C_1/(C_1 + C_2)$. This current must be equal to I_1, and hence

$$\left(1 + g_{m1} R_D \frac{C_1}{C_1 + C_2}\right) I_1 = 0 \tag{8.187}$$

This relation cannot hold because $g_{m1} R_D C_1/(C_1 + C_2)$ is not necessarily zero and I_1 itself is an external stimulus and nonzero. This nonsensical result arises because two ideal current sources, namely, I_1 and M_1, are placed in series. Similarly, V_1 cannot be calculated because the drain voltage of M_1 is not defined.

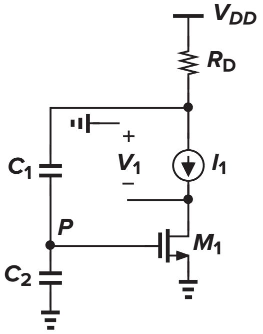


Figure 8.80

▶ Example 8.29

A student wrestling with the above example decides to attach a resistance from the drain of M_1 to ground and let its value go to infinity in the final result. Does this rescue Blackman's theorem?

Solution

As shown in Fig. 8.81, $A = R_T$. Moreover, we can now compute T_{oc} by writing a KCL at the drain of M_1:

$$-g_{m1} I_1 R_D \frac{C_1}{C_1 + C_2} - \frac{V_1}{R_T} = I_1 \tag{8.188}$$

and hence

$$T_{oc} = -g_{m2} \frac{V_1}{I_1} = g_{m2} \left(1 + g_{m1} R_D \frac{C_1}{C_1 + C_2}\right) R_T \tag{8.189}$$

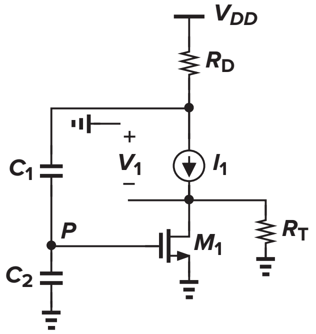


Figure 8.81

This result suggests that $T_{oc} \rightarrow \infty$ as $R_T \rightarrow \infty$. Since $T_{sc} = 0$ (why?), we have

$$R_{in} = A\frac{1 + T_{sc}}{1 + T_{oc}} \tag{8.190}$$

$$= R_T \frac{1}{1 + g_{m2}(1 + g_{m1}R_D\frac{C_1}{C_1 + C_2})R_T} \tag{8.191}$$

If $R_T \rightarrow \infty$, R_{in} approaches $1/g_{m2}$ divided by the loop gain.

It is peculiar that the return ratio of M_2 is not equal to that of M_1 even though the circuit appears to have only one feedback mechanism. But looks can be deceiving: M_2 is degenerated by R_T, experiencing local feedback. We can say M_2 sees infinite degeneration if $R_T = \infty$, and hence has an infinite return ratio.

▶ Example 8.30

Using Blackman's theorem, determine R_{in} in Fig. 8.82(a). Assume that $\gamma = 0$.

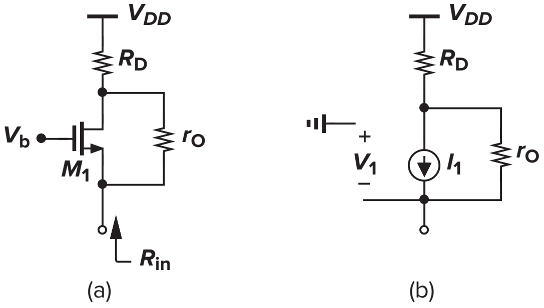


Figure 8.82

Solution

With $g_m = 0$, we have $A = R_D + r_O$. If the input port is shorted, no feedback is present and $T_{sc} = 0$. With the input port open [Fig. 8.82(b)], we observe that no current flows through R_D, I_1 generates a voltage of $-I_1 r_O$ across r_O, and $V_1 = -I_1 r_O$. That is, $T_{oc} = g_m r_O$. It follows that

$$R_{in} = \frac{R_D + r_O}{1 + g_m r_O} \tag{8.192}$$

as expected.

It is interesting to note that $T_{oc} > 0$ even though the feedback through r_O is *positive*. This occurs because the circuit contains *two* feedback mechanisms, one through r_O and another due to degeneration of M_1 by an infinite source resistance. In such a case, the sign of T_{oc} does not reveal the polarity of feedback. This point becomes clearer in the next example.

▶ Example 8.31

Determine the return ratios of M_1 and M_2 in Fig. 8.83(a), assuming $\lambda = \gamma = 0$.

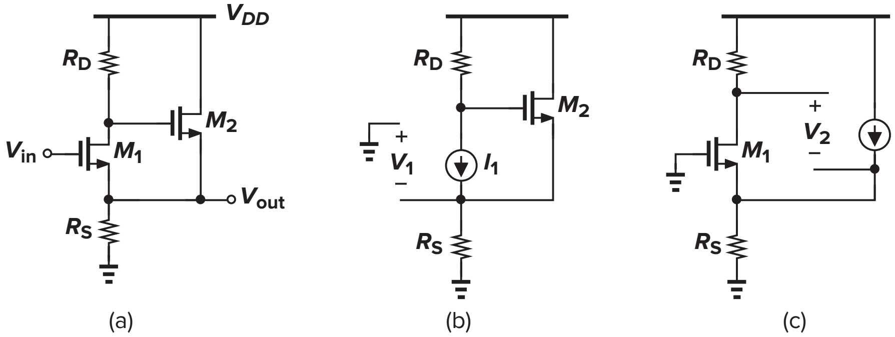


Figure 8.83

Solution

In this circuit, R_S degenerates both M_1 and M_2, and M_2 returns a voltage to the source of M_1 with positive feedback. Injecting a current as shown in Fig. 8.83(b), we note that R_S carries a current of $-V_1/R_S$, leading to $I_{D2} = -I_1 - V_1/R_S$, and hence $V_{GS2} = (-I_1 - V_1/R_S)/g_{m2}$. Adding the voltage drops across R_D and R_S to V_{GS2}, we have

$$I_1 R_D - \frac{I_1}{g_{m2}} - \frac{V_1}{g_{m2} R_S} - V_1 = 0 \tag{8.193}$$

and

$$RR_1 = -g_{m1} \frac{V_1}{I_1} \tag{8.194}$$

$$= \frac{1 - g_{m2} R_D}{1 + g_{m2} R_S} g_{m1} R_S \tag{8.195}$$

For RR_2, the arrangement in Fig. 8.83(c) yields $I_{D1} = -I_2 R_S/(R_S + 1/g_{m1}) = -I_2 g_{m1} R_S/(1 + g_{m1} R_S)$. Adding the voltage drops across R_D and R_S to V_2, we obtain

$$-\frac{I_2 g_{m1} R_S}{1 + g_{m1} R_S} R_D + V_2 + I_2 \frac{1/g_{m1}}{R_S + 1/g_{m1}} R_S = 0 \tag{8.196}$$

It follows that

$$RR_2 = \frac{1 - g_{m1} R_D}{1 + g_{m1} R_S} g_{m2} R_S \tag{8.197}$$

The return ratios are unequal and can assume positive or negative values independently.

8.7 ■ Middlebrook's Method

Middlebrook exploits the "Dissection Theorem" to derive the closed-loop transfer function without breaking the loop and while revealing the effect of backward (reverse) propagation in non-unilateral loops [5, 6]. This theorem states that any transfer function, $H(s)$, can be dissected into a product of the form

$$H(s) = H_\infty \frac{1 + \dfrac{1}{T_2}}{1 + \dfrac{1}{T_1}} \tag{8.198}$$

where H_{infty}, T_1, and T_2 are simpler transfer functions corresponding to special cases, e.g., with some signal in the loop forced to zero. These quantities are computed as follows. As shown in Fig. 8.84, we insert a voltage source, V_t, in series with a branch of the circuit and inject a current, I_t, to either side of V_t. We now have four new quantities, namely, V_1, V_2, I_1, and I_2. (Note the polarity of V_1.) The key point here is that the loop is not broken, and hence loading effects are immaterial. The "ideal" transfer function, H_∞, is obtained as follows:

$$H_\infty(s) = \frac{V_{out}}{V_{in}}|_{V1=0,I1=0} \tag{8.199}$$

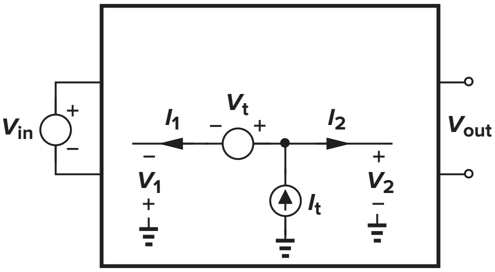


Figure 8.84 Illustration of Middlebrook's method.

i.e., we choose V_t and I_t such that V_1 and I_1 are forced to zero. The other two transfer functions are more involved. Middlebrook shows that

$$\frac{1}{T_1} = \frac{1}{T_i} + \frac{1}{T_v} + \frac{1}{T_i' T_v'} \tag{8.200}$$

where $V_{in} = 0$ and

$$T_i = \frac{I_1}{I_2}|_{V1=0} \quad \text{(Short-circuit forward current loop gain)} \tag{8.201}$$

$$T_v = \frac{V_1}{V_2}|_{I1=0} \quad \text{(Open-circuit forward voltage loop gain)} \tag{8.202}$$

$$\frac{1}{T_i'} = \frac{I_2}{I_1}|_{V2=0} \quad \text{(Short-circuit reverse current loop gain)} \tag{8.203}$$

$$\frac{1}{T_v'} = \frac{V_2}{V_1}|_{I2=0} \quad \text{(Open-circuit reverse voltage loop gain)} \tag{8.204}$$

The computation of T_2 is similar, except that it requires V_{out} (rather than V_{in}) to be forced to zero. We observe that Middlebrook's approach is generally more laborious than Bode's method.

Middlebrook's formulation provides insight regarding the forward (usually desirable) and reverse (usually undesirable) signal propagation around a nonunilateral loop. With no reverse propagation, we have $1/T_i' = 1/T_v' = 0$ and $T_1 = T_i||T_v$, e.g., the parallel combination of the two forward loop gains. Middlebrook denotes this quantity by T_{fwd}. In a similar fashion, we can define the total reverse loop gain as $T_{rev} = (1/T_i')||(1/T_v')$ and manipulate Eq. (8.200) to reach

$$T_1 = \frac{T_{fwd}}{1 + T_{rev}} \tag{8.205}$$

The interesting observation here is that the equivalent loop gain is degraded if the reverse loop gain, T_{rev}, becomes comparable to unity—even if it remains much less than T_{fwd}. Middlebrook, however, recognizes that this interpretation is valid only if (a) V_t and I_t are injected such that V_1 and I_1 are the "error" signal, a vague definition, and (b) V_t and I_t are injected inside the major loop and outside any minor loops, again a vague condition. For example, a degenerated CS stage with $\lambda > 0$ eludes both of these conditions.

8.8 ■ Loop Gain Calculation Issues

8.8.1 Preliminary Concepts

The loop gain plays a central role in feedback systems, as evidenced by the universal factor $1 + \beta A$ in the closed-loop expressions of gain, bandwidth, input and output impedances, and nonlinearity. In addition, if the poles and zeros in the loop are considered, then the loop gain [called the "loop transmission," $T(s)$, in this case] reveals the circuit's *stability* properties. For these reasons, we must often determine the loop gain even if we are not interested in the open-loop parameters of the circuit.

According to the procedure illustrated in Fig. 8.5, the loop gain calculation should be straightforward: we set the input to zero, break the loop at some point, apply a test signal, follow this signal around the loop (in the proper direction), and obtain the returned signal. However, in some cases, the situation is more complex, eliciting two questions: (1) Can we break the loop at any arbitrary point? (2) Should the test signal be a voltage or a current? We remind the reader that in such a test, the actual input and output disappear; i.e., the loop gain does not depend on where the main input and output ports are.

For example, consider the two-stage amplifier shown in Fig. 8.85(a), where the resistive divider consisting of R_1 and R_2 senses the output voltage and returns a fraction thereof to the source of M_1. As illustrated in Fig. 8.85(b), we set V_{in} to zero, break the loop at node X, apply a test signal to the right terminal of R_1, and measure the resulting V_F.[16] But is this test setup correct? First, we note that in Fig. 8.85(a), R_1 draws an ac current from R_{D2}, but in Fig. 8.85(b), it does not. That is, the gain associated with the second common-source stage has been altered. Second, why did we decide to apply a test *voltage*? Can we apply a test current and measure a returned current?

To address the first issue, we surmise that it is best to break the loop at the *gate* of a MOSFET. We can break the loop at the gate of M_2 [Fig. 8.85(c)] and thus not alter the gain associated with the first stage—at least at low frequencies. The reader is encouraged to derive the loop gain using Figs. 8.85(b) and (c) and show that they are not equal.

What if we must include C_{GS} of M_2 [Fig. 8.86(a)]? Then, we break the loop *after* C_{GS2} [Fig. 8.86(b)] to ensure that the load seen by M_1 remains unchanged. But is it always possible to break the loop at the

[16] It is clear that, upon breaking the loop, we must apply the test signal to R_1 and travel clockwise around the circuit. If we apply V_t to the drain of M_2 and travel counterclockwise, the result is meaningless.

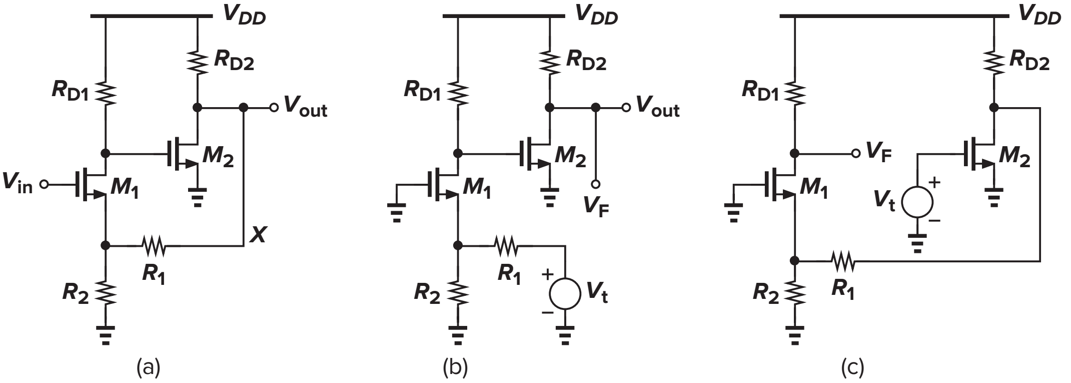


Figure 8.85 (a) Two-stage feedback amplifier, (b) breaking the loop at the left terminal of R_1, and (c) breaking the loop at the gate of M_2.

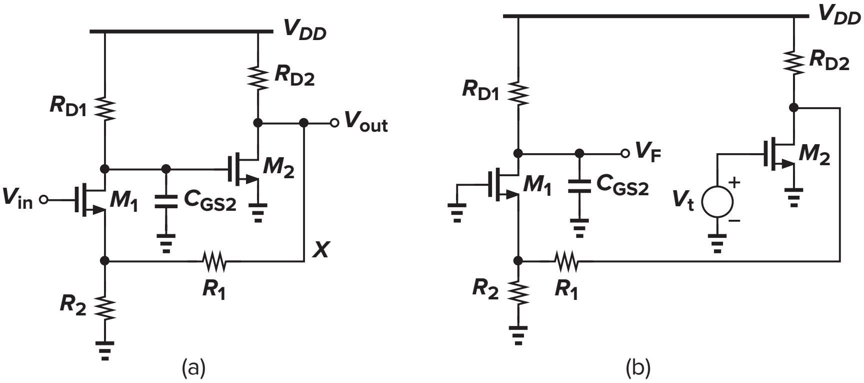


Figure 8.86 (a) Two-stage amplifier including C_{GS2}, and (b) breaking the loop at the gate of M_2.

gate of a MOSFET? Yes, indeed. For the feedback to be negative, the signal must be sensed by at least one gate in the loop because only the common-source topology inverts signals.

Let us now turn our attention to the second issue, namely, the *type* of the test signal. In the foregoing study, we naturally chose a test voltage, V_t, because we replaced the controlling voltage of a MOSFET with an independent source. Under what condition can we apply a test current? In Fig. 8.85(a), for example, we can break the loop at the drain of M_2, inject a current I_t, and measure the current returned by M_2 [Fig. 8.87(a)]. The reader can prove that I_F/I_t in this case is the same as V_F/V_t in Fig. 8.85(c).

But what exactly should we do with the drain node of M_2 in Fig. 8.87(a)? If tied to ac ground, this node does not experience the voltage excursions present in the closed-loop circuit—an issue when r_{O2} is taken into account. We can merge r_{O2} with R_{D2} in this case, but not if M_2 is degenerated. Thus, in general, we cannot inject I_t without altering some aspects of the circuit.

Not all hope is lost yet. Suppose we replace the controlled current source of M_2 with an independent current source, I_t, and compute the returned V_{GS} as V_F [Fig. 8.87(b)]. Since in the original circuit, the dependent source and V_{GS2} were related by a factor of g_{m2}, we can now write the loop gain as $(-V_F/I_t) \times g_{m2}$. This approach is feasible even if M_2 is degenerated. We recognize that this result is the same as the return ratio of M_2.

At low frequencies, the loop gain can be computed with the aid of the following observation. Since the circuit incorporates *negative* feedback, the loop must traverse the gate of a transistor (only the CS stage

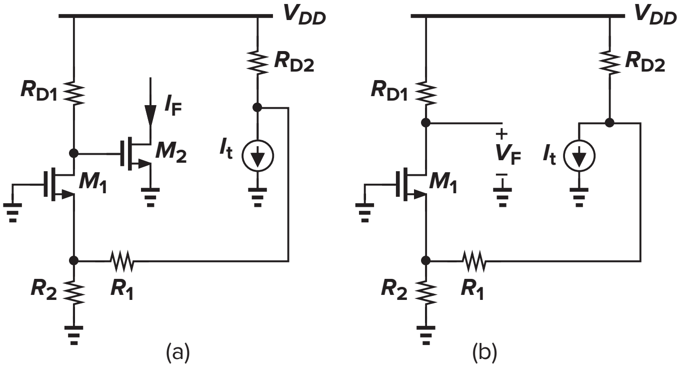


Figure 8.87 (a) Breaking the loop at the drain of M_2, and (b) replacing dependent source of M_2 with an independent source.

inverts).[17] We can therefore break the loop at this gate without the need for including loading effects. Of course, this method applies only if the loop has only one feedback mechanism.

In summary, the “best” place to break a feedback loop is (a) the gate-source of a MOSFET if voltage injection is desired, or (b) the dependent current source of a MOSFET if current injection is desired (provided that the returned quantity is V_{GS} of the MOSFET). Of course, these two methods are related because they differ by only a factor of g_m.

Unfortunately, the foregoing techniques face difficulties in some cases. For example, suppose we include C_{GD2} in Fig. 8.85(a). We inject a test voltage or current as before, but the issue is that C_{GD2} does not allow a “clean” break. As shown in Fig. 8.88, even though we provide the gate-source voltage by the independent source, V_t, C_{GD2} still creates “local” feedback from the drain of M_2 to its gate, raising the question of whether the loop gain should be obtained by nulling *all* feedback mechanisms. We should also mention that the method of current and voltage injection proposed by Middlebrook in [3] applies only if the loop is unilateral.

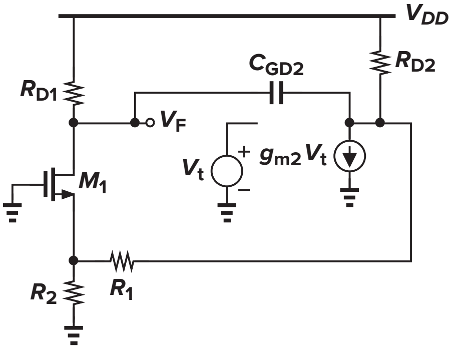


Figure 8.88 Two-stage amplifier including C_{D2}.

8.8.2 Difficulties with Return Ratio

Bode’s method enables us to compute the closed-loop transfer function in terms of four simpler transfer functions—without the need for breaking the loop. But we are also interested in the loop gain as it

[17] One exception are source-degenerated devices (in CS or follower stages).

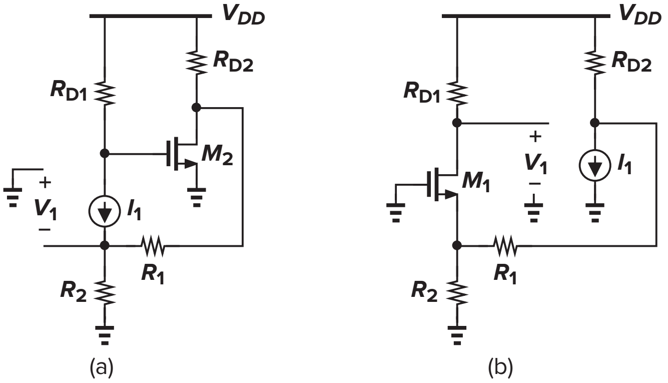


Figure 8.89 Equivalent circuits for the calculation of the return ratios for (a) M_1, and (b) M_2.

determines the consequences of applying feedback to a given circuit, e.g., the increase in the bandwidth, the reduction in the nonlinearity, and the stability behavior.

We may view the return ratio associated with a given dependent source as the loop gain, but circuits containing more than one feedback mechanism may exhibit different return ratios for different sources. As an example, we consider again the two-stage amplifier shown in Fig. 8.85(a), recognizing that R_1 and R_2 provide both "global" feedback and "local" feedback (by degenerating M_1). With the aid of the equivalent circuits shown in Fig. 8.89, the reader can show that the return ratios for M_1 and M_2 are respectively given by

$$\text{Return Ratio}|_{M1} = \frac{g_{m1}R_2(R_1 + R_{D2} + g_{m2}R_{D2}R_{D1})}{R_1 + R_2 + R_{D2}} \tag{8.206}$$

and

$$\text{Return Ratio}|_{M2} = \frac{g_{m1}g_{m2}R_2R_{D1}R_{D2}}{(1 + g_{m1}R_2)(R_1 + R_{D2}) + R_2} \tag{8.207}$$

If, as in our standard loop gain calculations, we break the loop at the gate of M_2, we obtain a value equal to the return ratio for M_2. It is unclear which return ratio should be considered the loop gain.

Why are the two return ratios different here? This is because disabling M_1 (by making I_1 an independent source) removes *both* feedback mechanisms whereas disabling M_2 still retains the degeneration experienced by M_1.

Another method of loop gain calculation is to inject a signal without breaking the loop, as shown in Fig. 8.90, and write $Y/W = 1/(1 + \beta A_0)$, and hence

$$\text{Loop Gain} = \left(\frac{Y}{W}\right)^{-1} - 1 \tag{8.208}$$

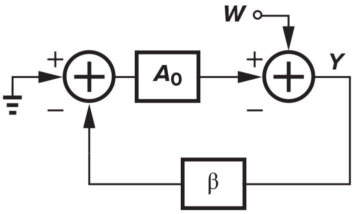


Figure 8.90 Another method of loop gain calculation.

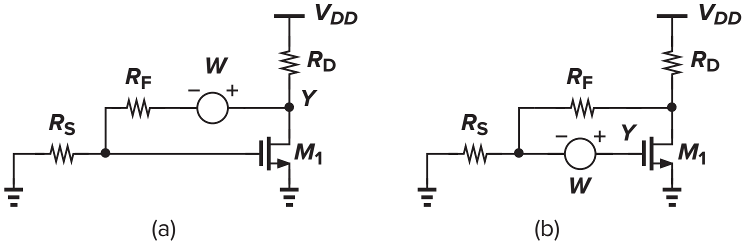


Figure 8.91 Different injection points in a nonunilateral circuit.

But this method tacitly assumes a unilateral loop, yielding different loop gains for different injection points if the loop is not unilateral. For example, the circuit of Fig. 8.71(a) can be excited as shown in Figs. 8.91(a) or (b), producing different values for $(Y/W)^{-1} - 1$.

The exact calculation of the loop gain for non-unilateral or multiloop circuits is beyond the scope of this book.

8.9 ■ Alternative Interpretations of Bode's Method

Bode's results can be manipulated to produce other forms that offer new insights.

Asymptotic Gain Form Let us return to $V_{out}/V_{in} = A + g_m BC/(1 - g_m D)$ and note that $V_{out}/V_{in} = A$ if $g_m = 0$ (the dependent source is disabled) and $V_{out}/V_{in} = A - BC/D$ if $g_m \to \infty$ (the dependent source is very "strong"). We denote these values of V_{out}/V_{in} by H_0 and H_∞, respectively, and $-g_m D$ by T. It is helpful to consider H_0 as the direct feedthrough and H_∞ as the "ideal" gain, i.e., if the dependent source were infinitely strong (or the loop gain were infinite). It follows that

$$\frac{V_{out}}{V_{in}} = H_0 + \frac{g_m BC}{1+T} \tag{8.209}$$

$$= H_0 \frac{1+T}{1+T} + \frac{g_m BC}{1+T} \tag{8.210}$$

$$= \frac{H_0}{1+T} + \frac{T(H_0 + g_m BC/T)}{1+T} \tag{8.211}$$

Since $H_0 + g_m BC/T = A - g_m BC/(g_m D) = A - BC/D = H_\infty$, we have

$$\frac{V_{out}}{V_{in}} = H_\infty \frac{T}{1+T} + H_0 \frac{1}{1+T} \tag{8.212}$$

Called the "asymptotic gain equation" [4], this form reveals that the gain consists of an ideal value multiplied by $T/(1+T)$ and a direct feedthrough multiplied by $1/(1+T)$. The calculations are somewhat simpler here if we recognize from $V_1 = CV_{in} + DI_1$ and $I_1 = g_m V_1$ that $V_1 = CV_{in}/(1 - g_m D) \to 0$ if $g_m \to \infty$ (provided that $V_{in} < \infty$). This is similar to how a virtual ground is created if the loop gain is large.

▶ **Example 8.32**

Calculate the voltage gain of the circuit shown in Fig. 8.92(a) using the asymptotic gain method. Assume that $\lambda = \gamma = 0$.

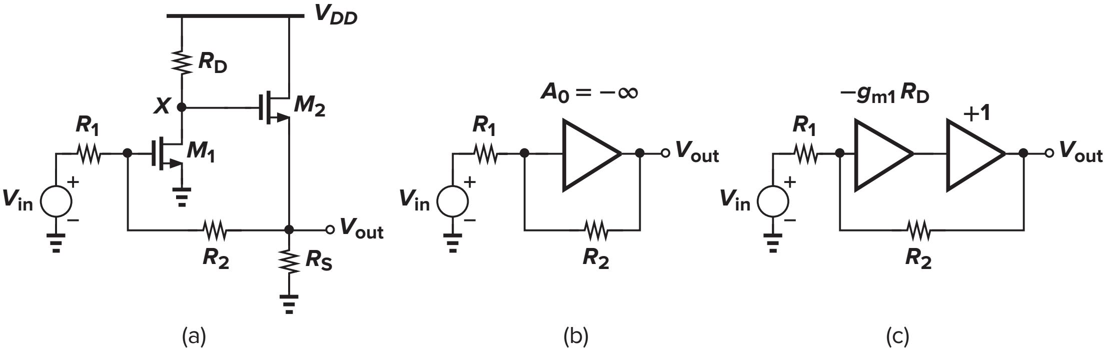


Figure 8.92

Solution

Suppose M_1 is the dependent source of interest. If $g_{m1} = 0$, then V_{in} propagates through R_1 and R_2 and sees an impedance of $(1/g_{m2})||R_S$ at the source of M_2. Thus,

$$H_0 = \frac{(1/g_{m2})||R_S}{(1/g_{m2})||R_S + R_1 + R_2} \tag{8.213}$$

If $g_{m1} = \infty$, then $V_{GS1} = 0$ (like a virtual ground), yielding a current of V_{in}/R_1 through R_1 and R_2. That is

$$H_\infty = -\frac{R_2}{R_1} \tag{8.214}$$

an expected result because M_1 and M_2 operate as an op amp with an infinite open-loop gain [Fig. 8.92(b)]. To determine the return ratio for M_1, we set V_{in} to zero, replace M_1's dependent source with an independent source, I_1, and express V_X as $-I_1 R_D$. Since M_2 sees a load resistance of $R_S||(R_1 + R_2)$, we have $V_{out} = -I_1 R_D[R_S||(R_1 + R_2)]/[1/g_{m2} + R_S||(R_1 + R_2)]$. The gate voltage of M_1 is equal to $V_{out}R_1/(R_1 + R_2)$, leading to

$$T_1 = g_{m1}R_D \frac{g_{m2}[R_S||(R_1 + R_2)]}{1 + g_{m2}[R_S||(R_1 + R_2)]} \frac{R_1}{R_1 + R_2} \tag{8.215}$$

We must now substitute for H_∞, T, and H_0 in Eq. (8.212) to obtain the closed-loop gain—a laborious task left for the reader. This example suggests that the direct analysis of the circuit (without knowledge of feedback) may in fact be simpler in some cases, as is true for this circuit.

It is instructive to repeat the foregoing calculations if M_2 is the dependent source of interest. For $g_{m2} = 0$, V_{in} is simply divided according to

$$H_0 = \frac{R_S}{R_S + R_1 + R_2} \tag{8.216}$$

For $g_{m2} = \infty$, we have $V_{GS2} = 0$, $V_X = V_{out}$, and hence a current of $-V_{out}/R_D$ flowing through M_1. It follows that $V_{GS1} = -V_{out}/(g_{m1}R_D)$ and $[V_{in} + V_{out}/(g_{m1}R_D)]/R_1 = [-V_{out}/(g_{m1}R_D) - V_{out}]/R_2$. We therefore have

$$H_\infty = \frac{-g_{m1}R_2R_D}{R_1 + R_2 + g_{m1}R_1R_D} \tag{8.217}$$

This result is also expected if we consider M_2 an ideal unity-gain buffer (due to its infinite g_m) and redraw the circuit as shown in Fig. 8.92(c).

The return ratio for M_2 can be found as

$$T_2 = \frac{g_{m2}R_S(g_{m1}R_1R_D + R_1 + R_2)}{R_S + R_1 + R_2} \tag{8.218}$$

Again, these values must be substituted in (8.212) to compute the closed-loop gain.

◀

Double-Null Method Blackman's impedance theorem raises an interesting question: Can we write the *transfer function* of a circuit in a form similar to $A(1 + T_{sc})/(1 + T_{oc})$? In other words, can we generalize the result to a case in which I_{in} is replaced by an arbitrary input and V_{in} by an arbitrary output? To understand the rationale for this question, let us observe that (1) T_{oc} is the return ratio with $I_{in} = 0$, i.e., T_{oc} denotes the RR with the *input* set to zero in Fig. 8.76(a); and (2) T_{sc} is the RR with $V_{in} = 0$, i.e., T_{sc} represents the return ratio with the *output* forced to zero. Figure 8.93 conceptually illustrates the setups for these two measurements, with one "nulling" the input and the other, the output. We make a slight change in our notation and postulate that the transfer function of a given circuit can be written as

$$\frac{V_{out}}{V_{in}} = A\frac{1 + T_{out,0}}{1 + T_{in,0}} \tag{8.219}$$

where $A = V_{out}/V_{in}$ with the dependent source set to zero, and $T_{out,0}$ and $T_{in,0}$ respectively denote the return ratios for $V_{out} = 0$ and $V_{in} = 0$.

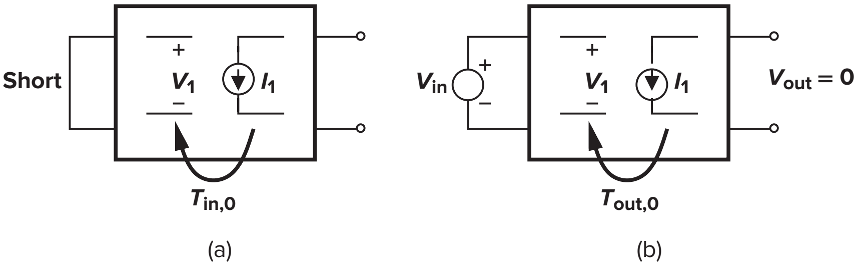


Figure 8.93 Conceptual illustration of $T_{in,0}$ and $T_{out,0}$.

The proof of (8.219) is similar to that of Blackman's theorem. Beginning from

$$V_{out} = AV_{in} + BI_1 \tag{8.220}$$

$$V_1 = CV_{in} + DI_1 \tag{8.221}$$

we recognize that, if $V_{in} = 0$, then $V_1/I_1 = D$, and hence $T_{in,0} = -g_m D$. On the other hand, if $V_{out} = 0$, then $V_{in} = (-B/A)I_1$, and hence $V_1/I_1 = (AD - BC)/A$, i.e., $T_{out,0} = -g_m(AD - BC)/A$. Combining these results indeed yields (8.219). Note that division by A in these calculations assumes that $A \neq 0$, a critical point revisited below.

Equation (8.219) offers interesting insights. The quantity $T_{out,0}$ reveals that, even though V_{in} and I_1 are chosen so as to drive V_{out} to zero, there is still an "internal" feedback loop emanating from I_1 and producing a finite value for V_1. The generic system of Fig. 8.1, on the other hand, does not lend itself to this perspective because its feedback network, $G(s)$, directly senses the output. The following example illustrates this point.

▶ Example 8.33

Determine V_{out}/V_{in} in Fig. 8.94(a), assuming $\lambda = \gamma = 0$. Note that the feedback network does not sense the main output here.

Solution

If M_1 is the dependent source of interest and $g_{m1} = 0$, then the voltage at the source of M_2 is equal to $V_{in}(R_S||g_{m2}^{-1})/(R_S||g_{m2}^{-1} + R_1 + R_2)$, yielding

$$A = g_{m2}R_{D2}\frac{R_S||g_{m2}^{-1}}{R_S||g_{m2}^{-1} + R_1 + R_2} \tag{8.222}$$

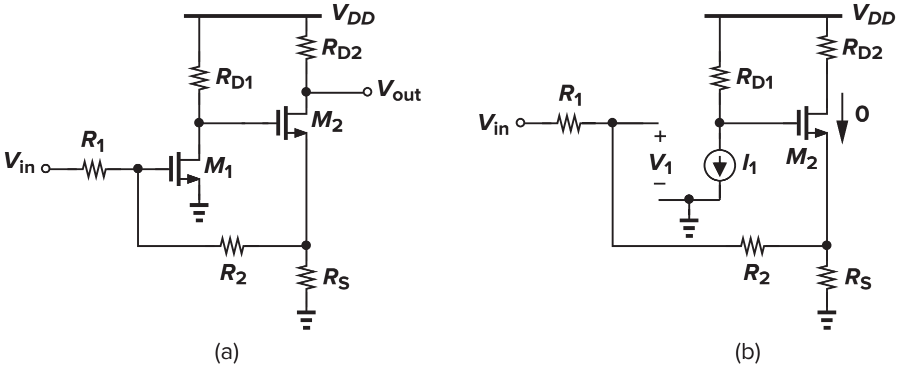


Figure 8.94

To obtain $T_{out,0}$, we choose V_{in} and I_1 so as to produce $V_{out} = 0$, and hence $V_{GS2} = 0$ and $I_{D2} = 0$ [Fig. 8.94(b)]. The source voltage of M_2 is therefore equal to $-I_1R_{D1}$ and also equal to $V_{in}R_S/(R_1 + R_2 + R_S)$. Similarly, $V_1 = V_{in}(R_2 + R_S)/(R_1 + R_2 + R_S)$ and

$$T_{out,0} = -g_{m1}\frac{V_1}{I_1} \tag{8.223}$$

$$= g_{m1}R_{D1}\frac{R_2 + R_S}{R_S} \tag{8.224}$$

The nonzero $T_{out,0}$ implies that I_1 still controls V_1 through an internal loop even though $V_{out} = 0$. The loop gain with $V_{in} = 0$ is given by Eq. (8.215).

What if M_2 is the dependent source of interest? Then, for $g_{m2} = 0$, we have $V_{out} = 0$, and hence $A = 0$. Equation (8.219) thus fails to hold because its derivation has assumed that $A \neq 0$. This shortcoming of the double-null method manifests itself in many CMOS circuits, even in a simple degenerated common-source stage.

References

[1] H. W. Bode, *Network Analysis and Feedback Amplifier Design* (New York, D. Van Nostrand, Inc., 1945).

[2] R. B. Blackman, "Effect of Feedback on Impedance," *Bell System Tech. J.*, vol. 23, pp. 269–277, October 1943.

[3] R. D. Middlebrook, "Measurement of Loop Gain in Feedback Systems," *Int. J. Electronics*, vol. 38, pp. 485–512, April 1975.

[4] S. Rosenstark, "A Simplified Method of Feedback Amplifier Analysis," *IEEE Trans. Education,* vol. 17, pp. 192–198, November 1974.

[5] R. D. Middlebrook, "The General Feedback Theorem: A Final Solution for Feedback Systems," *IEEE Microwave Magazine*, pp. 50–63, April 2006.

[6] R. D. Middlebrook, unpublished chapters available at www.rdmiddlebrook.com.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

8.1. Consider the circuit of Fig. 8.3(b), assuming that I_1 is ideal and $g_{m1}r_{O1}$ cannot exceed 50. If a gain error of less than 5% is required, what is the maximum closed-loop voltage gain that can be achieved by this topology? What is the low-frequency closed-loop output impedance under this condition?

8.2. In the circuit of Fig. 8.8(a), assume that $(W/L)_1 = 50/0.5$, $(W/L)_2 = 100/0.5$, $R_D = 2\ \text{k}\Omega$, and $C_2 = C_1$. Neglecting channel-length modulation and body effect, determine the bias current of M_1 and M_2 such that the input resistance at low frequencies is equal to 50 Ω.

8.3. Calculate the output impedance of the circuit shown in Fig. 8.9(a) at relatively low frequencies if R_D is replaced by an ideal current source.

8.4. Consider the example illustrated in Fig. 8.11. Suppose an overall voltage gain of 500 is required with maximum bandwidth. How many stages with what gain per stage must be placed in a cascade? (Hint: first find the 3-dB bandwidth of a cascade of n identical stages in terms of that of each stage.)

8.5. If in Fig. 8.22(b), amplifier A_0 exhibits an output impedance of R_0, calculate the closed-loop voltage gain and output impedance, taking into account loading effects.

8.6. Consider the circuit of Fig. 8.25(a), assuming that $(W/L)_{1,2} = 50/0.5$ and $(W/L)_{3,4} = 100/0.5$. If $I_{SS} = 1$ mA, what is the maximum closed-loop voltage gain that can be achieved if the gain error is to remain below 5%?

8.7. The circuit of Fig. 8.42 can operate as a transimpedance amplifier if I_{out} flows through a resistor, R_{D2}, connected to V_{DD}, producing an output voltage. Replacing R_S with an ideal current source and assuming that $\lambda = \gamma = 0$, calculate the transimpedance of the resulting circuit. Also, calculate the input-referred noise current per unit bandwidth.

8.8. For the circuit of Fig. 8.51(a), calculate the closed-loop gain without neglecting $G_{12}I_2$. Prove that this term can be neglected if $G_{12} \ll A_0 Z_{in}/Z_{out}$.

8.9. Calculate the loop gain of the circuit in Fig. 8.54 by breaking the loop at node X. Why is this result somewhat different from $G_{21}A_{v,open}$?

8.10. Using feedback techniques, calculate the input and output impedance and voltage gain of each circuit in Fig. 8.95.

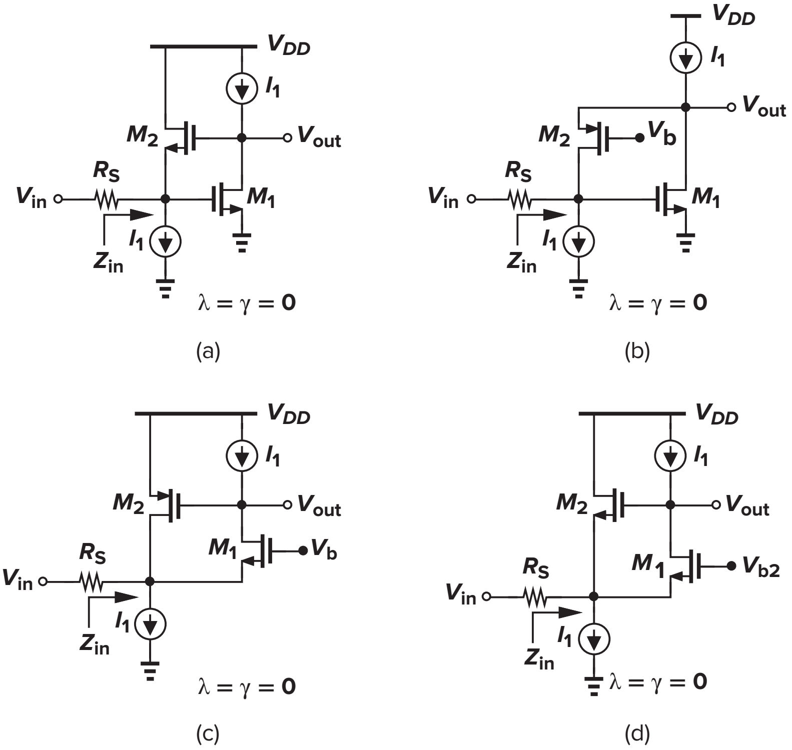


Figure 8.95

8.11. Using feedback techniques, calculate the input and output impedances of each circuit in Fig. 8.96.

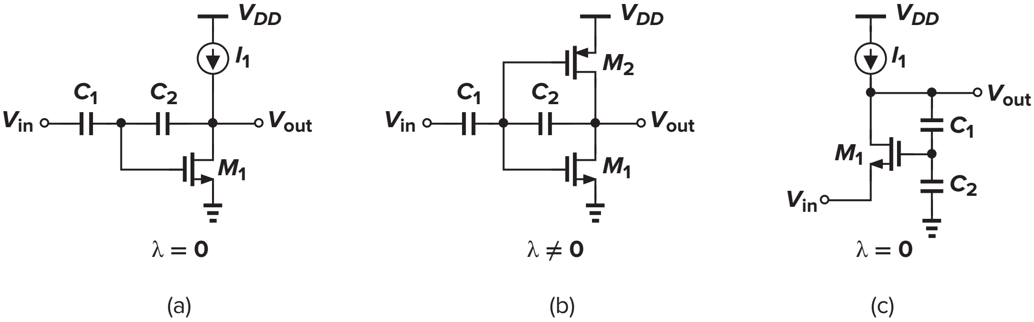


Figure 8.96

8.12. Consider the circuit of Fig. 8.54(a), assuming that $(W/L)_1 = (W/L)_2 = 50/0.5$, $\lambda = \gamma = 0$, and each resistor is equal to 2 kΩ. If $I_{D2} = 1$ mA, what is the bias current of M_1? What value of V_{in} gives such a current? Calculate the overall voltage gain.

8.13. Suppose the amplifier of the circuit shown in Fig. 8.22 has an open-loop transfer function $A_0/(1 + s/\omega_0)$ and an output resistance R_0. Calculate the output impedance of the closed-loop circuit and plot the magnitude as a function of frequency. Explain the behavior.

8.14. Calculate the input-referred noise voltage of the circuit shown in Fig. 8.25(a) at relatively low frequencies.

8.15. A differential pair with current-source loads can be represented as in Fig. 8.97(a), where $R_0 = r_{ON} \| r_{OP}$ and r_{ON} and r_{OP} denote the output resistance of NMOS and PMOS devices, respectively. Consider the circuit shown in Fig. 8.97(b), where G_{m1} and G_{m2} are placed in a negative feedback loop.

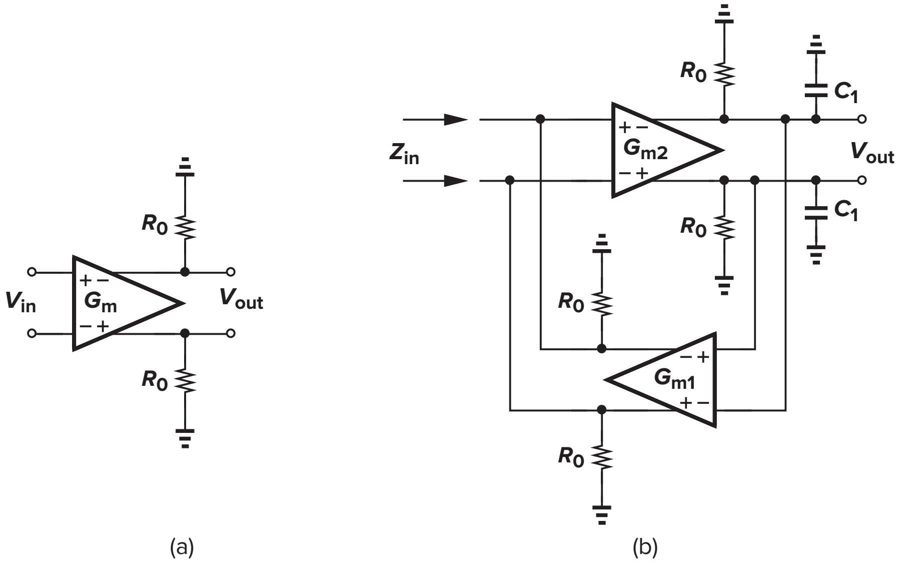


Figure 8.97

(a) Neglecting all other capacitances, derive an expression for Z_{in}. Sketch $|Z_{in}|$ versus frequency.
(b) Explain intuitively the behavior observed in part (a).
(c) Calculate the input-referred thermal noise voltage and current in terms of the input-referred noise voltage of each G_m stage.

8.16. In the circuit of Fig. 8.98, $(W/L)_{1-3} = 50/0.5$, $I_{D1} = |I_{D2}| = |I_{D3}| = 0.5$ mA, and $R_{S1} = R_F = R_{D2} = 3$ kΩ.

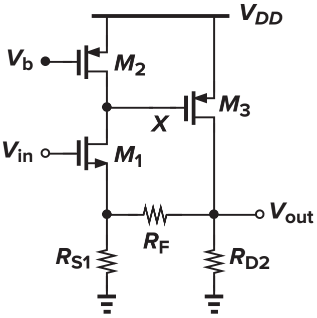


Figure 8.98

(a) Determine the input bias voltage required to establish the above currents.
(b) Calculate the closed-loop voltage gain and output resistance.

8.17. The circuit of Fig. 8.98 can be modified as shown in Fig. 8.99, where a source follower, M_4, is inserted in the feedback loop. Note that M_1 and M_4 can also be viewed as a differential pair. Assume that $(W/L)_{1-4} = 50/0.5$, $I_D = 0.5$ mA, for all transistors $R_{S1} = R_F = R_{D2} = 3$ kΩ, and $V_{b2} = 1.5$ V. Calculate the closed-loop voltage gain and output resistance, and compare the results with those obtained in the previous problem.

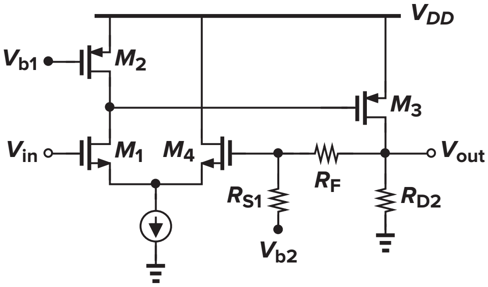


Figure 8.99

8.18. Consider the circuit of Fig. 8.100, where $(W/L)_{1-4} = 50/0.5$, $|I_{D1-4}| = 0.5$ mA, and $R_2 = 3$ kΩ.

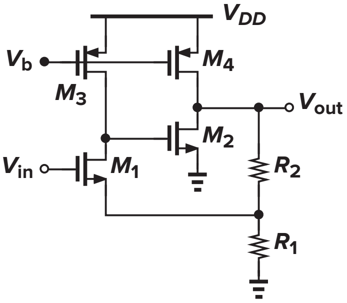


Figure 8.100

(a) For what range of R_1 are the above currents established while M_2 remains in saturation? What is the corresponding range of V_{in}?
(b) Calculate the closed-loop gain and output impedance for R_1 in the middle of the range obtained in part (a).

8.19. In the circuit of Fig. 8.101, suppose all resistors are equal to 2 kΩ and $g_{m1} = g_{m2} = 1/(200\ \Omega)$. Assuming that $\lambda = \gamma = 0$, calculate the closed-loop gain and output impedance.

8.20. A CMOS inverter can be used as an amplifier with or without feedback (Fig. 8.102). Assume that $(W/L)_{1,2} = 50/0.5$, $R_1 = 1$ kΩ, $R_2 = 10$ kΩ, and the dc levels of V_{in} and V_{out} are equal.
(a) Calculate the voltage gain and the output impedance of each circuit.
(b) Calculate the sensitivity of each circuit's output with respect to the supply voltage. That is, calculate the small-signal "gain" from V_{DD} to V_{out}. Which circuit exhibits less sensitivity?

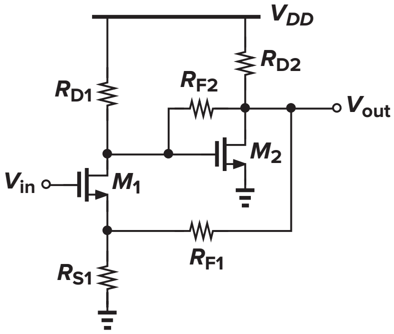


Figure 8.101

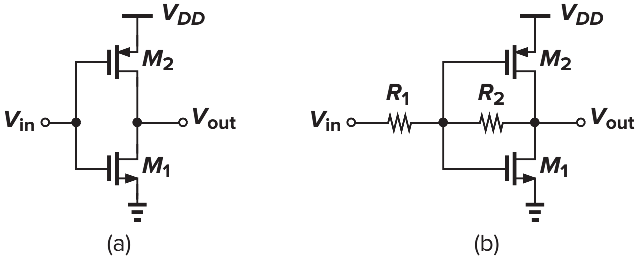


Figure 8.102

8.21. Calculate the input-referred thermal noise voltage of the circuits shown in Fig. 8.102.

8.22. The circuit shown in Fig. 8.103 employs positive feedback to produce a negative input capacitance. Using feedback analysis techniques, determine Z_{in} and identify the negative capacitance component. Assume that $\lambda = \gamma = 0$.

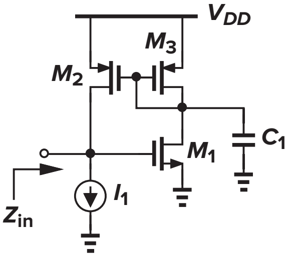


Figure 8.103

8.23. In the circuit of Fig. 8.104, assume that $\lambda = 0$, $g_{m1,2} = 1/(200\ \Omega)$, $R_{1-3} = 2\ \mathrm{k}\Omega$, and $C_1 = 100$ pF. Neglecting other capacitances, estimate the closed-loop voltage gain at very low and very high frequencies.

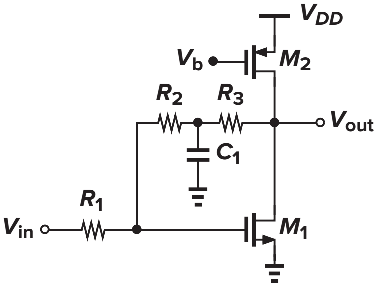


Figure 8.104

CHAPTER

9

Operational Amplifiers

Operational amplifiers (op amps) are an integral part of many analog and mixed-signal systems. Op amps with vastly different levels of complexity are used to realize functions ranging from dc bias generation to high-speed amplification or filtering. The design of op amps continues to pose a challenge as the supply voltage and transistor channel lengths scale down with each generation of CMOS technologies.

This chapter deals with the analysis and design of CMOS op amps. Following a review of performance parameters, we describe simple op amps such as telescopic and folded-cascode topologies. Next, we study two-stage and gain-boosting configurations and the problem of common-mode feedback. Finally, we introduce the concept of slew rate and analyze the effect of supply rejection and noise in op amps. The reader is encouraged to read this chapter before dealing with more advanced designs in Chapter 11.

9.1 ■ General Considerations

We loosely define an op amp as a "high-gain differential amplifier." By "high," we mean a value that is adequate for the application, typically in the range of 10^1 to 10^5. Since op amps are usually employed to implement a feedback system, their open-loop gain is chosen according to the precision required of the closed-loop circuit.

Up to three decades ago, most op amps were designed to serve as "general-purpose" building blocks, satisfying the requirements of many different applications. Such efforts sought to create an "ideal" op amp, e.g., with a very high voltage gain (several hundred thousand), high input impedance, and low output impedance, but at the cost of many other aspects of the performance, e.g., speed, output voltage swings, and power dissipation.

By contrast, today's op amp design proceeds with the recognition that the trade-offs between the parameters eventually require a multi dimensional compromise in the overall implementation, making it necessary to know the *adequate* value that must be achieved for each parameter. For example, if the speed is critical while the gain error is not, a topology is chosen that favors the former, possibly sacrificing the latter.

9.1.1 Performance Parameters

In this section, we describe a number of op amp design parameters, providing an understanding of why and where each may become important. For this discussion, we consider the differential cascode circuit

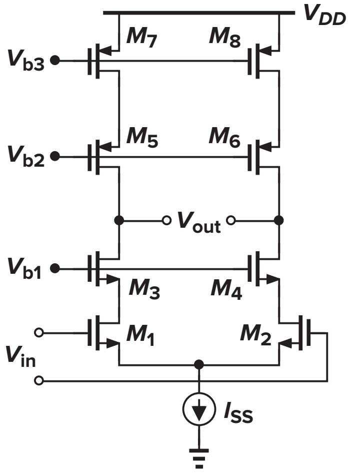


Figure 9.1 Cascode op amp.

shown in Fig. 9.1 as a representative op amp design.[1] The voltages $V_{b1}-V_{b3}$ are generated by the current mirror techniques described in Chapter 5.

Gain The open-loop gain of an op amp determines the precision of the feedback system employing the op amp. As mentioned before, the required gain may vary by four orders of magnitude according to the application. Trading with such parameters as speed and output voltage swings, the minimum required gain must therefore be known. As explained in Chapter 14, a high open-loop gain may also be necessary to suppress nonlinearity.

▶ **Example 9.1**

The circuit of Fig. 9.2 is designed for a nominal gain of 10, i.e., $1 + R_1/R_2 = 10$. Determine the minimum value of A_1 for a gain error of 1%.

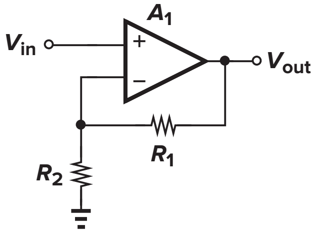


Figure 9.2

Solution

The closed-loop gain is obtained from Chapter 8 as

$$\frac{V_{out}}{V_{in}} = \frac{A_1}{1 + \dfrac{R_2}{R_1 + R_2} A_1} \tag{9.1}$$

$$= \frac{R_1 + R_2}{R_2} \frac{A_1}{\dfrac{R_1 + R_2}{R_2} + A_1} \tag{9.2}$$

[1] Since op amps of this type have a high output impedance, they are sometimes called “operational transconductance amplifiers” (OTAs). In the limit, the circuit can be represented by a single voltage-dependent current source and called a “G_m stage.”

Predicting that $A_1 \gg 10$, we approximate (9.2) as

$$\frac{V_{out}}{V_{in}} \approx \left(1+\frac{R_1}{R_2}\right)\left(1-\frac{R_1+R_2}{R_2}\frac{1}{A_1}\right) \tag{9.3}$$

The term $(R_1+R_2)/(R_2A_1) = (1+R_1/R_2)/A_1$ represents the relative gain error. To achieve a gain error less than 1%, we must have $A_1 > 1000$.

It is instructive to compare the circuit of Fig. 9.2 with an open-loop implementation such as that in Fig. 9.3. While it is possible to obtain a nominal gain of $g_m R_D = 10$ by a common-source stage, it is extremely difficult to guarantee an error less than 1%. The variations in the mobility and gate-oxide thickness of the transistor and the value of the resistor typically yield an error greater than 20%.

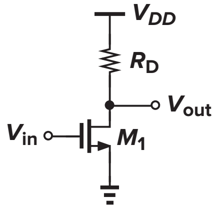


Figure 9.3 Simple common-source stage.

Small-Signal Bandwidth The high-frequency behavior of op amps plays a critical role in many applications. For example, as the frequency of operation increases, the open-loop gain begins to drop (Fig. 9.4), creating larger errors in the feedback system. The small-signal bandwidth is usually defined as the "unity-gain" frequency, f_u, which can reach several gigahertz in today's CMOS op amps. The 3-dB frequency, $f_{3\text{-dB}}$, may also be specified to allow easier prediction of the closed-loop frequency response.

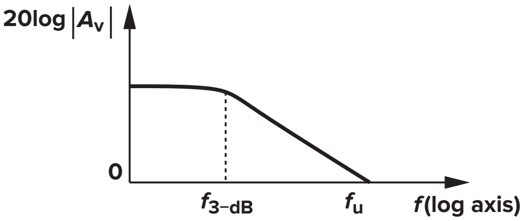


Figure 9.4 Gain roll-off with frequency.

▶ **Example 9.2**

In the circuit of Fig. 9.5, assume that the op amp is a single-pole voltage amplifier. If V_{in} is a small step, calculate the time required for the output voltage to reach within 1% of its final value. What unity-gain bandwidth must the

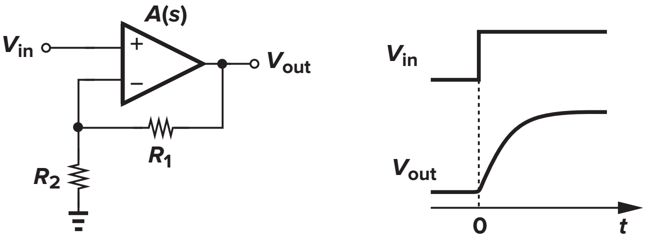


Figure 9.5

op amp provide if $1 + R_1/R_2 \approx 10$ and the settling time is to be less than 5 ns? For simplicity, assume that the low-frequency gain is much greater than unity.

Solution

Since

$$\left(V_{in} - V_{out}\frac{R_2}{R_1 + R_2}\right) A(s) = V_{out} \tag{9.4}$$

we have

$$\frac{V_{out}}{V_{in}}(s) = \frac{A(s)}{1 + \dfrac{R_2}{R_1 + R_2} A(s)} \tag{9.5}$$

For a one-pole system, $A(s) = A_0/(1+s/\omega_0)$, where ω_0 is the 3-dB bandwidth and $A_0\omega_0$ the unity-gain bandwidth. Thus,

$$\frac{V_{out}}{V_{in}}(s) = \frac{A_0}{1 + \dfrac{R_2}{R_1 + R_2}A_0 + \dfrac{s}{\omega_0}} \tag{9.6}$$

$$= \frac{\dfrac{A_0}{1 + \dfrac{R_2}{R_1 + R_2}A_0}}{1 + \dfrac{s}{\left(1 + \dfrac{R_2}{R_1 + R_2}A_0\right)\omega_0}} \tag{9.7}$$

indicating that the closed-loop amplifier is also a one-pole system with a time constant equal to

$$\tau = \frac{1}{\left(1 + \dfrac{R_2}{R_1 + R_2}A_0\right)\omega_0} \tag{9.8}$$

Recognizing that the quantity $R_2A_0/(R_1 + R_2)$ is the low-frequency loop gain and usually much greater than unity, we have

$$\tau \approx \left(1 + \frac{R_1}{R_2}\right)\frac{1}{A_0\omega_0} \tag{9.9}$$

The output step response for $V_{in} = au(t)$ can now be expressed as

$$V_{out}(t) \approx a\left(1 + \frac{R_1}{R_2}\right)\left(1 - \exp\frac{-t}{\tau}\right)u(t) \tag{9.10}$$

with the final value $V_F \approx a(1 + R_1/R_2)$. For 1% settling, $V_{out} = 0.99V_F$, and hence

$$1 - \exp\frac{-t_{1\%}}{\tau} = 0.99, \tag{9.11}$$

yielding $t_{1\%} = \tau \ln 100 \approx 4.6\tau$. For a 1% settling of 5 ns, $\tau \approx 1.09$ ns, and from (9.9), $A_0\omega_0 \approx (1+R_1/R_2)/\tau = 9.21$ Grad/s (1.47 GHz).

The key point in the above example is that the bandwidth is dictated by both the required settling accuracy (e.g., $V_{out} = 0.99V_F$) and the closed-loop gain $(1 + R_1/R_2)$.

Example 9.3

A student mistakenly swaps the inverting and non-inverting inputs of the op amp in Fig. 9.5. Explain how the circuit behaves.

Solution

Positive feedback may destabilize the circuit. For a one-pole op amp, we have

$$\left(V_{out}\frac{R_2}{R_1+R_2}-V_{in}\right)\frac{A_0}{1+\dfrac{s}{\omega_0}}=V_{out} \tag{9.12}$$

and hence

$$\frac{V_{out}}{V_{in}}(s)=\frac{\dfrac{A_0}{1-\dfrac{R_2}{R_1+R_2}A_0}}{1-\dfrac{s}{(1+\dfrac{R_2}{R_1+R_2}A_0)\omega_0}} \tag{9.13}$$

Interestingly, the closed-loop amplifier contains a pole in the *right half* plane, exhibiting a step response that grows exponentially with time:

$$V_{out}(t)\approx a\left(1+\frac{R_1}{R_2}\right)\left(\exp\frac{t}{\tau}-1\right)u(t) \tag{9.14}$$

This growth continues until the op amp output saturates.

Large-Signal Behavior In many of today's applications, op amps must operate with large transient signals. Under these conditions, nonlinear phenomena make it difficult to characterize the speed merely by small-signal properties such as the open-loop response shown in Fig. 9.4. As an example, suppose the feedback circuit of Fig. 9.5 incorporates a realistic op amp (i.e., with finite output impedance) while driving a large load capacitance. How does the circuit behave if we apply a 1-V step at the input? Since the output voltage cannot change instantaneously, the voltage difference sensed by the op amp itself at $t \geq 0$ is equal to 1 V. Such a large difference momentarily drives the op amp into a nonlinear region of operation. (Otherwise, with an open-loop gain of, say, 1000, the op amp would produce 1000 V at the output.)

As explained in Sec. 9.9, the large-signal behavior is usually quite complex, calling for careful simulations.

Output Swing Most systems employing op amps require large voltage swings to accommodate a wide range of signal amplitudes. For example, a high-quality microphone that senses the music produced by an orchestra may generate instantaneous voltages that vary by more than four orders of magnitude, demanding that subsequent amplifiers and filters handle large swings (and/or achieve a low noise).

The need for large output swings has made fully differential op amps popular. Similar to the circuits described in Chapter 4, such op amps generate "complementary" outputs, roughly doubling the available swing. Nonetheless, as mentioned in Chapters 3 and 4 and explained later in this chapter, the maximum voltage swing trades with device size and bias currents and hence speed. Achieving large swings is the principal challenge in today's op amp design.

Linearity Open-loop op amps suffer from substantial nonlinearity. In the circuit of Fig. 9.1, for example, the input pair M_1–M_2 exhibits a nonlinear relationship between its differential drain current and its input voltage. As explained in Chapter 14, the issue of nonlinearity is tackled by two approaches: using fully differential implementations to suppress even-order harmonics and allowing sufficient open-loop gain for the closed-loop feedback system to achieve adequate linearity. It is interesting to note that in many feedback circuits, the linearity requirement, rather than the gain error requirement, governs the choice of the open-loop gain.

Noise and Offset The input noise and offset of op amps determine the minimum signal level that can be processed with reasonable quality. In a typical op amp topology, several devices contribute noise and offset, necessitating large dimensions or bias currents. For example, in the circuit of Fig. 9.1, M_1–M_2 and M_7-M_8 contribute the most.

We should also recognize a trade-off between noise and *output swing*. For a given bias current, as the overdrive voltage of M_7 and M_8 in Fig. 9.1 is lowered to allow larger swings at the output, their transconductance increases and so does their drain noise current.

Supply Rejection Op amps are often employed in mixed-signal systems and sometimes connected to noisy digital supply lines. Thus, the performance of op amps in the presence of supply noise, especially as the noise frequency increases, is important. For this reason, fully differential topologies are preferred.

9.2 ■ One-Stage Op Amps

9.2.1 Basic Topologies

All of the differential amplifiers studied in Chapters 4 and 5 can be considered op amps. Figure 9.6 shows two such topologies with single-ended and differential outputs. The small-signal, low-frequency gain of both circuits is equal to $g_{mN}(r_{ON}\|r_{OP})$, where the subscripts N and P denote NMOS and PMOS, respectively. This value hardly exceeds 10 in nanometer technologies. The bandwidth is usually determined by the load capacitance, C_L. Note that the circuit of Fig. 9.6(a) exhibits a mirror pole (Chapter 6) whereas that of Fig. 9.6(b) does not, a critical difference in terms of the stability of feedback systems using these topologies (Chapter 10).

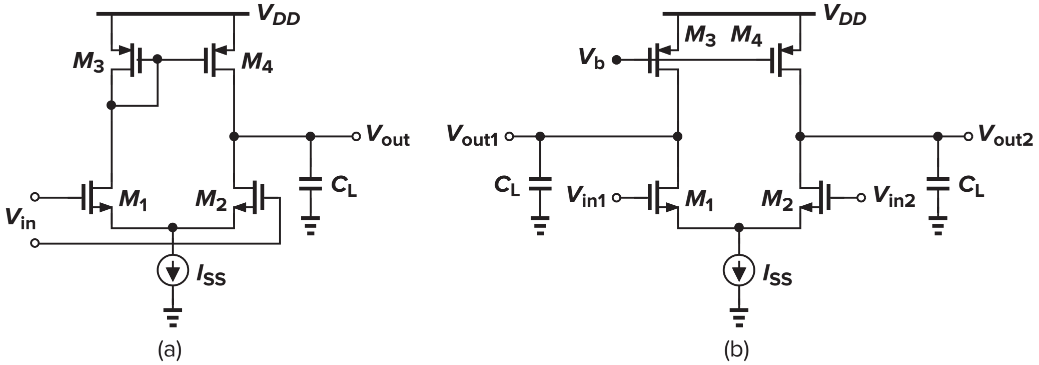


Figure 9.6 Simple op amp topologies.

The circuits of Fig. 9.6 suffer from noise contributions of M_1–M_4, as calculated in Chapter 7. Interestingly, in all op amp topologies, at least four devices contribute to the input noise: two input transistors and two “load” transistors.

▶ **Example 9.4**

Calculate the input common-mode voltage range and the closed-loop output impedance of the unity-gain buffer depicted in Fig. 9.7.

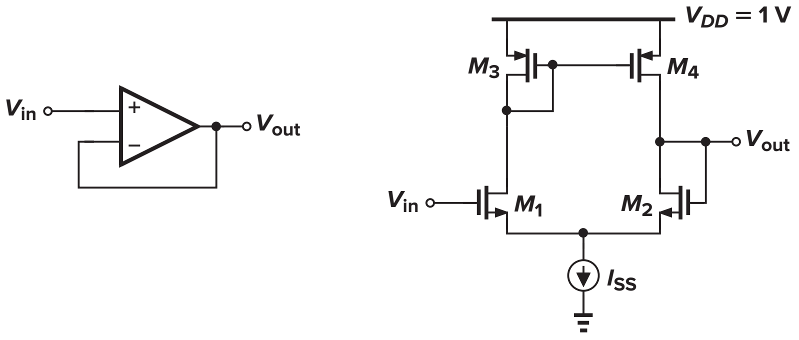


Figure 9.7

Solution

The minimum allowable input voltage is equal to $V_{ISS} + V_{GS1}$, where V_{ISS} is the voltage required across the current source. The maximum voltage is given by the level that places M_1 at the edge of the triode region: $V_{in,max} = V_{DD} - |V_{GS3}| + V_{TH1}$. For example, if each device (including the current source) has a threshold voltage of 0.3 V and an overdrive of 0.1 V, then $V_{in,min} = 0.1 + 0.1 + 0.3 = 0.5$ V and $V_{in,max} = 1 - (0.1 + 0.3) + 0.3 = 0.9$ V. Thus, the input CM range equals 0.4 V with a 1-V supply.

Since the circuit employs voltage feedback at the output, the output impedance is equal to the open-loop value, $r_{OP} \| r_{ON}$, divided by one plus the loop gain, $1 + g_{mN}(r_{OP} \| r_{ON})$. In other words, for large open-loop gain, the closed-loop output impedance is approximately equal to $(r_{OP} \| r_{ON})/[g_{mN}(r_{OP} \| r_{ON})] = 1/g_{mN}$.

It is interesting to note that the closed-loop output impedance is relatively *independent* of the open-loop output impedance. This is an important observation, allowing us to design high-gain op amps by *increasing* the open-loop output impedance while still achieving a relatively low closed-loop output impedance. We also observe that, if driving a load capacitance of C_L, the op amp incurs a closed-loop output pole approximately given by g_{mN}/C_L.

In order to achieve a high gain, the differential cascode topologies of Chapters 4 and 5 can be used. Shown in Figs. 9.8(a) and (b) for single-ended and differential output generation, respectively, such circuits display a gain on the order of $g_{mN}[(g_{mN} r_{ON}^2) \| (g_{mP} r_{OP}^2)]$, but at the cost of output swing and

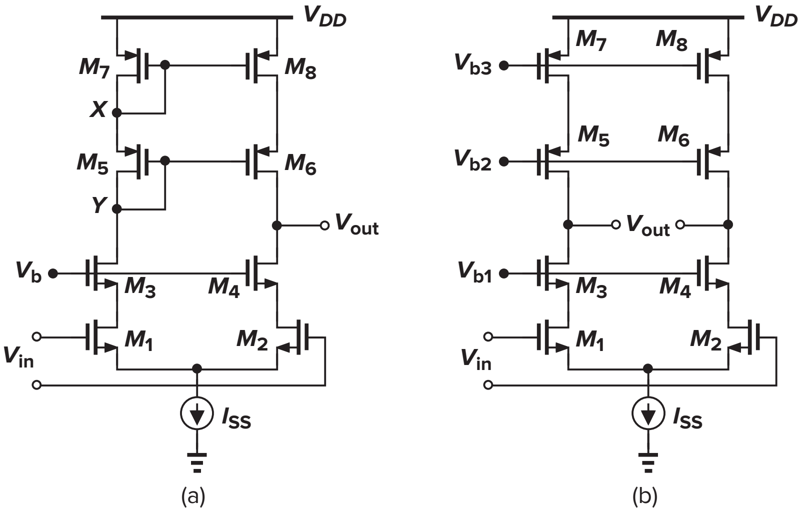


Figure 9.8 Cascode op amps.

additional poles. These configurations are also called "telescopic" cascode op amps to distinguish them from another cascode op amp described below. The circuit providing a single-ended output suffers from a mirror pole at node X (and a pole at Y), creating stability issues (Chapter 10).

As calculated in Chapter 4 , the output swings of telescopic op amps are relatively limited. In the fully differential version of Fig. 9.8(b), for example, the output swing is given by $2[V_{DD} - (V_{OD1} + V_{OD3} + V_{ISS} + |V_{OD5}| + |V_{OD7}|)]$, where V_{ODj} denotes the overdrive voltage of M_j and V_{ISS} the minimum allowable voltage across I_{SS}. We must recognize the three conditions necessary for allowing this much swing: (1) the input CM level, $V_{in,CM}$, is chosen *low* enough and equal to $V_{GS1} + V_{ISS}$, (2) V_{b1} is also chosen low enough and equal to $V_{GS3} + (V_{in,CM} - V_{TH1})$, placing M_1 at the edge of saturation, and (3) V_{b2} is chosen high enough and equal to $V_{DD} - |V_{OD7}| - |V_{GS5}|$, placing M_7 at the edge of saturation. Thus, $V_{in,CM}$ (and V_{b1} and V_{b2}) must be controlled tightly, a serious issue.

Another drawback of telescopic cascodes is the difficulty in shorting their inputs and outputs, e.g., to implement a unity-gain buffer similar to the circuit of Fig. 9.7. To understand the issue, let us consider the unity-gain feedback topology shown in Fig. 9.9. Under what conditions are both M_2 and M_4 in saturation? We must have $V_{out} \leq V_X + V_{TH2}$ and $V_{out} \geq V_b - V_{TH4}$. Since $V_X = V_b - V_{GS4}$, $V_b - V_{TH4} \leq V_{out} \leq V_b - V_{GS4} + V_{TH2}$. Depicted in Fig. 9.9, this voltage range is simply equal to $V_{max} - V_{min} = V_{TH4} - (V_{GS4} - V_{TH2})$ (one threshold minus one overdrive), maximized by minimizing the overdrive of M_4 but always less than V_{TH2}.

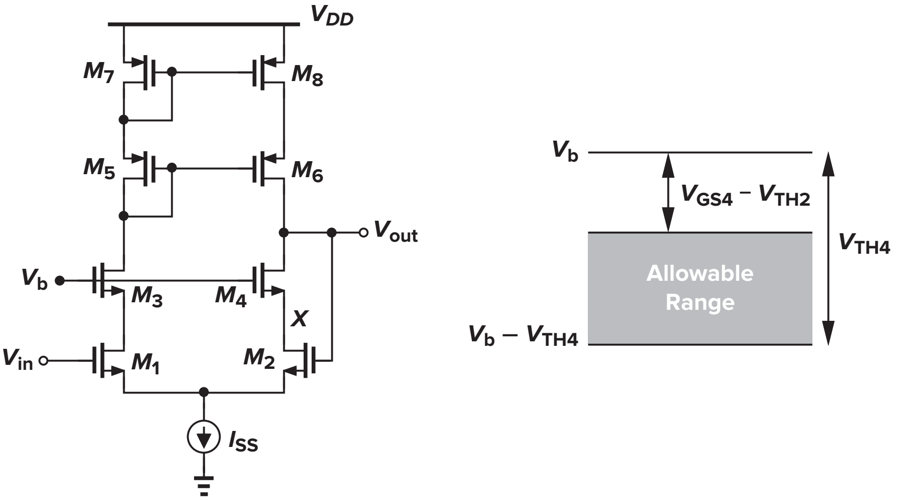


Figure 9.9 Telescopic cascode op amp with input and output shorted.

▶ **Example 9.5**

For the circuit of Fig. 9.9, explain in which region each transistor operates as V_{in} varies from below $V_b - V_{TH4}$ to above $V_b - V_{GS4} + V_{TH2}$.

Solution

Since the op amp attempts to force V_{out} to be equal to V_{in}, for $V_{in} < V_b - V_{TH4}$, we have $V_{out} \approx V_{in}$, and M_4 is in the triode region while other transistors are saturated. Under this condition, the open-loop gain of the op amp is reduced.

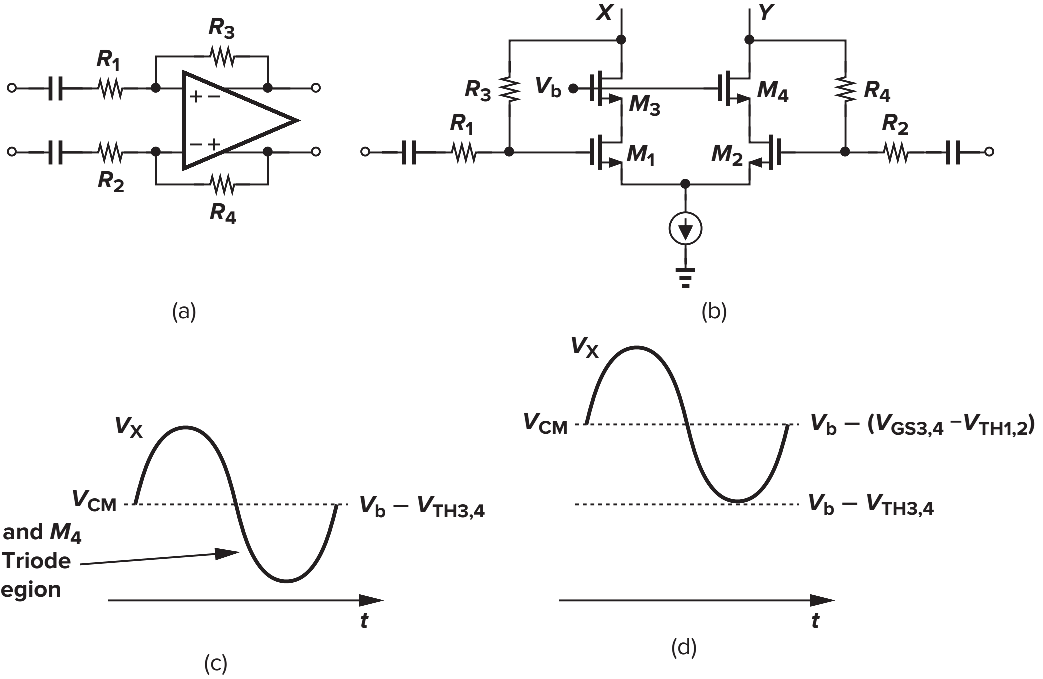
As V_{in} and hence V_{out} exceed $V_b - V_{TH4}$, M_4 enters saturation and the open-loop gain reaches a maximum. For $V_b - V_{TH4} < V_{in} < V_b - (V_{GS4} - V_{TH2})$, both M_2 and M_4 are saturated, and for $V_{in} > V_b - (V_{GS4} - V_{TH2})$, M_2 and M_1 enter the triode region, degrading the gain.

◀

While a cascode op amp is rarely used as a unity-gain buffer, some other topologies (such as the switched-capacitor circuits of Chapter 13) reduce to the configuration shown in Fig. 9.9 for part of their operation period, as illustrated by the following example.

▶ **Example 9.6**

Figure 9.10(a) shows a closed-loop amplifier utilizing a telescopic op amp.[2] Assuming that the op amp has a high open-loop gain, determine the maximum allowable output voltage swing.



(a)

(b)

(c)

(d)

Figure 9.10

Solution

Let us draw the circuit as shown in Fig. 9.10(b), noting that its input and output common-mode levels are equal (why?). Recall from the foregoing discussion that the voltage at the drains of M_3 and M_4 is bounded by $V_b - V_{TH3,4}$ to keep M_3 and M_4 in saturation and $V_b - (V_{GS3,4} - V_{TH1,2})$ to keep M_1 and M_2 in saturation. How should we set the output CM level, V_{CM}, in this range to maximize the output swing? If $V_{CM} = V_b - V_{TH3,4}$, then M_3 and M_4 reside at the edge of the triode region and cannot tolerate any *downward* swing [Fig. 9.10(c)]. On the other hand, if we select $V_{CM} = V_b - (V_{GS3,4} - V_{TH1,2})$ (placing M_1 and M_2 at the edge), then V_X or V_Y can fall to $V_b - V_{TH3,4}$ while maintaining M_3 and M_4 in saturation [Fig. 9.10(d)].

With the latter choice, how *high* can V_X or V_Y go? If the gain of the op amp is large, the gate voltages of M_1 and M_2 swing negligibly. Thus, V_X and V_Y can arbitrarily rise from $V_{CM} = V_b - (V_{GS3,4} - V_{TH1,2})$ without driving M_1 and M_2 into the triode region. (Of course, the PMOS loads constrain the upswing.) For symmetric up- and downswings, therefore, the circuit allows a voltage excursion of ±(one threshold − one overdrive) around V_{CM}.

◀

[2] The input capacitors ensure that the bias conditions are not disturbed by the preceding stage.

9.2.2 Design Procedure

At this point, the reader may wonder how exactly we design an op amp. With so many devices and performance parameters, it may not be clear where the starting point is and how the numbers are chosen. Indeed, the actual design methodology of an op amp somewhat depends on the specifications that the circuit must meet. For example, a high-gain op amp may be designed quite differently from a low-noise op amp. Nevertheless, in most cases, some aspects of the performance, e.g., output voltage swings and open-loop gain, are of primary concern, pointing to a specific design procedure. We will deal extensively with five parameters for each transistor: I_D, $V_{GS} - V_{TH}$, W/L, g_m, and r_O.

In the design of op amps (and many other circuits), it is helpful to begin with a power budget, even if none is specified. As seen later in this section, the resulting design can readily be "scaled" for lower or higher power dissipations. We describe a simple design here and deal with nanometer op amps in Chapter 11.

▶ **Example 9.7**

Design a fully differential telescopic op amp with the following specifications: $V_{DD} = 3$ V, peak-to-peak differential output swing = 3 V, power dissipation = 10 mW, voltage gain = 2000. Assume that $\mu_n C_{ox} = 60\ \mu$A/V^2, $\mu_p C_{ox} = 30\ \mu$A/V^2, $\lambda_n = 0.1$ V^{-1}, $\lambda_p = 0.2$ V^{-1} (for an effective channel length of 0.5 μm), $\gamma = 0$, and $V_{THN} = |V_{THP}| = 0.7$ V.

Solution

Figure 9.11 shows the op amp topology along with two current mirrors defining the drain currents of M_7–M_9. We begin with the power budget, allocating 3 mA to M_9 and the remaining 330 μA to M_{b1} and M_{b2}. Thus, each cascode branch of the op amp carries a current of 1.5 mA. Next, we consider the required output swings. Each of nodes X and Y must be able to swing by 1.5 V_{pp} without driving M_3–M_6 into the triode region. With a 3-V supply, therefore, the total voltage available for M_9 and each cascode branch is equal to 1.5 V, i.e., $|V_{OD7}| + |V_{OD5}| + V_{OD3} + V_{OD1} + V_{OD9} = 1.5$ V.

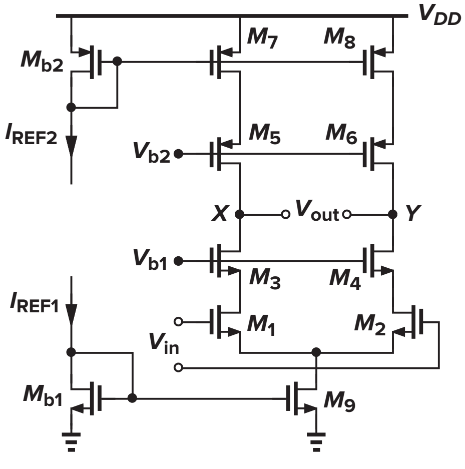


Figure 9.11

Since M_9 carries the largest current, we choose $V_{OD9} \approx 0.5$ V, leaving 1 V for the four transistors in the cascode. Moreover, since M_5–M_8 suffer from low mobility, we allocate an overdrive of approximately 300 mV to each, obtaining 400 mV for $V_{OD1} + V_{OD3}$. As an initial guess, $V_{OD1} = V_{OD3} = 200$ mV.

With the bias current and overdrive voltage of each transistor known, we can easily determine the aspect ratios from $I_D = (1/2)\mu C_{ox}(W/L)(V_{GS} - V_{TH})^2$ or simulated I/V characteristics. To minimize the device capacitances, we choose the minimum length for each transistor, obtaining a corresponding width. We then have $(W/L)_{1-4} = 1250$, and $(W/L)_{5-8} = 1111$, and $(W/L)_9 = 400$.

The reader may think that the above choice of overdrives is arbitrary and leads to a wide design space. However, we must emphasize that each of the overdrives has but a small range. For example, we can change the allocated values by only a few tens of millivolts before the device dimensions become disproportionately large.

The design has thus far satisfied the swing, power dissipation, and supply voltage specifications. But, how about the gain? Using $A_v \approx g_{m1}[(g_{m3}r_{O3}r_{O1})\|(g_{m5}r_{O5}r_{O7})]$ and assuming minimum channel length for all of the transistors, we have $A_v = 1416$, quite a lot lower than the required value.

In order to increase the gain, we recognize that $g_m r_O = \sqrt{2\mu C_{ox}(W/L)I_D}/(\lambda I_D)$. Now, recall that $\lambda \propto 1/L$, and hence $g_m r_O \propto \sqrt{WL/I_D}$. We can therefore increase the width or length or *decrease* the bias current of the transistors. In practice, speed or noise requirements may dictate the bias current, leaving only the dimensions as the variables. Of course, the width of each transistor must at least scale with its length so as to maintain a constant overdrive voltage.

Which transistors in the circuit of Fig. 9.11 should be made longer? Since M_1–M_4 appear in the signal path, it is desirable to keep their capacitances to a minimum. The PMOS devices, M_5–M_8, on the other hand, affect the signal to a much lesser extent and can therefore have larger dimensions.[3] Doubling the (effective) length and width of each of these transistors in fact *doubles* their $g_m r_O$ because g_m remains constant while r_O increases by a factor of 2. Choosing $(W/L)_{5-8} = 2222\ \mu\text{m}/1.0\ \mu\text{m}$ and hence $\lambda_p = 0.1\ \text{V}^{-1}$, we obtain $A_v \approx 4000$. Thus, the PMOS dimensions can be somewhat smaller. Note that with such large dimensions for PMOS transistors, we may revisit our earlier distribution of the overdrive voltages, possibly reducing that of M_9 by 100 to 200 mV and allocating more to the PMOS devices.

In the op amp of Fig. 9.11, the input CM level and the bias voltages V_{b1} and V_{b2} must be chosen so as to allow maximum output swings. The minimum allowable input CM level equals $V_{GS1} + V_{OD9} = V_{TH1} + V_{OD1} + V_{OD9} = 1.4$ V. The minimum value of V_{b1} is given by $V_{GS3} + V_{OD1} + V_{OD9} = 1.6$ V, placing M_1–M_2 at the edge of the triode region. Similarly, $V_{b2,max} = V_{DD} - (|V_{GS5}| + |V_{OD7}|) = 1.7$ V. In practice, some margin must be included in the value of V_{b1} and V_{b2} to allow for process variations. Also, the increase in the threshold voltages due to body effect must be taken into account. Finally, we should remark that this op amp requires common-mode feedback (CMFB) (Section 9.7).

◀

9.2.3 Linear Scaling

How do we modify the above design if the power budget is different but all other specifications remain the same? Suppose we are allowed to double the power dissipation and hence the bias current of each transistor. The key concept behind "linear scaling" is to double the widths of all of the transistors in the circuit while keeping the lengths constant. Returning to our five device design parameters, we observe that, in this example, (1) I_D is doubled, (2) W/L is doubled, (3) $V_{GS} - V_{TH}$ is *constant*, and so are the allowable voltage swings, (4) g_m is *doubled* because both the bias current and the width are doubled (as if two identical transistors were placed in parallel), and (5) r_O is halved (for the same reason that g_m is doubled). We therefore conclude that linear scaling by adjusting the transistor widths simply scales the power dissipation while retaining the gain and swing values. This concept is used in Chapter 11 to optimize the performance of op amps.

▶ **Example 9.8**

An engineer seeking a low-power op amp design scales down the transistor widths in Example 9.7 by a factor of 10. Explain what aspects of the performance degrade.

Solution

Since the g_m of each transistor falls by a factor of 10, two aspects are sacrificed: (1) the speed of the op amp in driving a capacitive load (e.g., the output pole in Example 9.4) degrades proportionally, and (2) the input-referred noise voltage of the op amp rises by a factor of $\sqrt{10}$ (Sec. 9.12).

◀

[3] This point is studied in Chapter 10.

In nanometer technologies, op amp design can still follow the above procedure, but with greater reliance on simulated device characteristics. Unfortunately, the lower supply voltage severely limits the output swing, making the telescopic cascode less attractive. We return to these points in Chapter 11.

The gate bias voltages V_{b1} and V_{b2} in the telescopic cascode of Fig. 9.11 must be generated with some precision. We note that if, for example, V_{b1} is less than its nominal value, then M_1 and M_2 enter the triode region. The same occurs even if V_{b1} is fixed, but the input CM level is slightly higher than expected. To ensure that V_{b1} "tracks" the input CM level, we can generate V_{b1} as shown in Fig. 9.12(a). Here, a small current I_1 flows through the diode-connected device, M_{b1}, producing $V_{b1} = V_P + V_{GS,b1}$. Since V_P tracks the input CM level ($V_P = V_{in,CM} - V_{GS1,2}$), we have

$$V_{b1} = V_{in,CM} - V_{GS1,2} + V_{GS,b1} \tag{9.15}$$

which should be chosen equal to $V_{in,CM} - V_{TH1,2} + V_{GS3,4}$ to allow M_1 and M_2 to operate in saturation. It follows that

$$V_{GS,b1} = (V_{GS1,2} - V_{TH1,2}) + V_{GS3,4} \tag{9.16}$$

indicating that M_{b1} must be "weak" enough to sustain a V_{GS} equal to one overdrive plus the gate-source voltage of M_3 and M_4. This is accomplished by choosing M_{b1} to be a narrrow, long device.

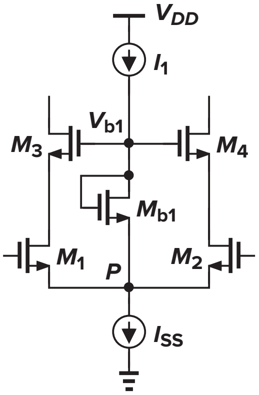


Figure 9.12 Generation of cascode gate voltage.

9.2.4 Folded-Cascode Op Amps

In order to alleviate the drawbacks of telescopic cascode op amps, namely, limited output swings and difficulty in choosing equal input and output CM levels, a "folded-cascode" op amp can be used. As described in Chapter 3 and illustrated in Fig. 9.13, in an NMOS or PMOS cascode amplifier, the input device is replaced by the opposite type while still converting the input voltage to a current. In the four circuits shown in Fig. 9.13, the small-signal current generated by M_1 flows through M_2 and subsequently the load, producing an output voltage approximately equal to $g_{m1}R_{out}V_{in}$. The primary advantage of the folded structure lies in the choice of the voltage levels because it does not "stack" the cascode transistor on top of the input device. We will return to this point later.

The folding idea depicted in Fig. 9.13 can easily be applied to differential pairs, and hence to operational amplifiers as well. Shown in Fig. 9.14, the resulting circuit replaces the input NMOS pair with a PMOS counterpart. Note two important differences between the two circuits. (1) In Fig. 9.14(a), one bias current, I_{SS}, provides the drain current of both the input transistors and the cascode devices, whereas in Fig. 9.14(b), the input pair requires an additional bias current. In other words, $I_{SS1} = I_{SS}/2 + I_{D3} = I_{SS}/2 + I_1$. Thus, the folded-cascode configuration generally consumes more power. (2) In Fig. 9.14(a), the input CM level

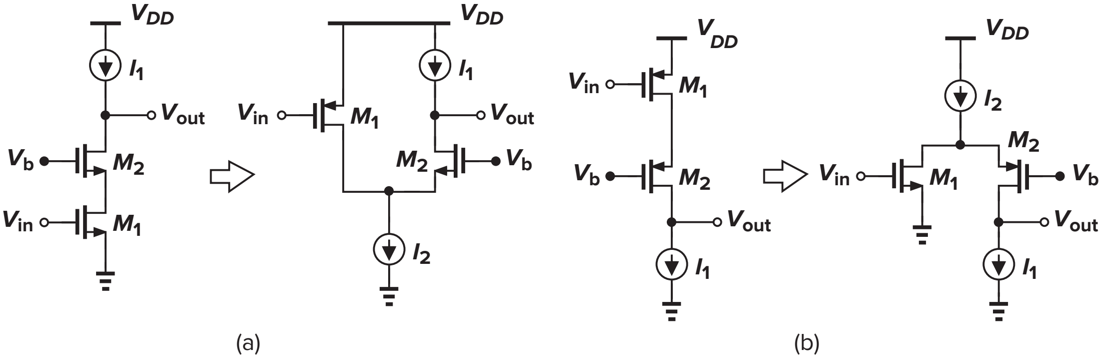


Figure 9.13 Folded-cascode amplifiers.

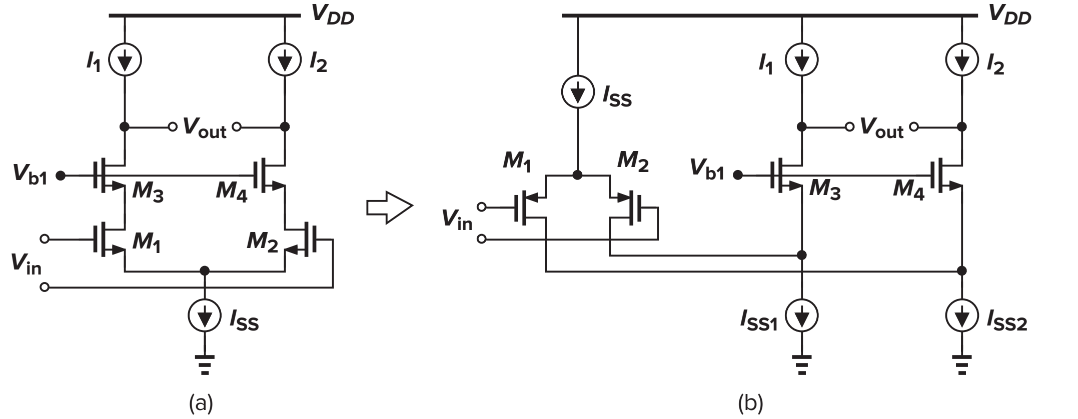


Figure 9.14 (a) Telescopic and (b) folded-cascode op amp topologies.

cannot exceed $V_{b1} - V_{GS3} + V_{TH1}$, whereas in Fig. 9.14(b), it cannot be *less* than $V_{b1} - V_{GS3} - |V_{THP}|$. It is therefore possible to design the latter to allow shorting its input and output terminals with negligible swing limitation. This is in contrast to the behavior depicted in Fig. 9.9. In Fig. 9.14(b), it is possible to tie the n-wells of M_1 and M_2 to their common source point. We return to this idea in Chapters 14 and 19.

Let us now calculate the maximum output voltage swing of the folded-cascode op amp shown in Fig. 9.15, where M_5–M_{10} replace the ideal current sources of Fig. 9.14(b). With proper choice of V_{b1} and V_{b2}, the lower end of the swing is given by $V_{OD3} + V_{OD5}$ and the upper end by $V_{DD} - (|V_{OD7}| + |V_{OD9}|)$. Thus, the peak-to-peak swing on each side is equal to $V_{DD} - (V_{OD3} + V_{OD5} + |V_{OD7}| + |V_{OD9}|)$. In the telescopic cascode of Fig. 9.14(a), on the other hand, the swing is less by the overdrive of the tail current source. We should nonetheless note that, carrying a large current, M_5 and M_6 in Fig. 9.15 may require a high overdrive voltage if their capacitance contribution to nodes X and Y is to be minimized.

We now determine the small-signal voltage gain of the folded-cascode op amp of Fig. 9.15. Using the half circuit depicted in Fig. 9.16(a) and writing $|A_v| = G_m R_{out}$, we must calculate G_m and R_{out}. As shown in Fig. 9.16(b), the output short-circuit current is approximately equal to the drain current of M_1 because the impedance seen looking into the source of M_3, that is, $(g_{m3} + g_{mb3})^{-1} \| r_{O3}$, is typically much lower than $r_{O1} \| r_{O5}$. Thus, $G_m \approx g_{m1}$. To calculate R_{out}, we use Fig. 9.16(c), with $R_{OP} \approx (g_{m7} + g_{mb7}) r_{O7} r_{O9}$, to write $R_{out} \approx R_{OP} \| [(g_{m3} + g_{mb3}) r_{O3} (r_{O1} \| r_{O5})]$. It follows that

$$|A_v| \approx g_{m1}\{[(g_{m3} + g_{mb3}) r_{O3} (r_{O1} \| r_{O5})] \| [(g_{m7} + g_{mb7}) r_{O7} r_{O9}]\} \tag{9.17}$$

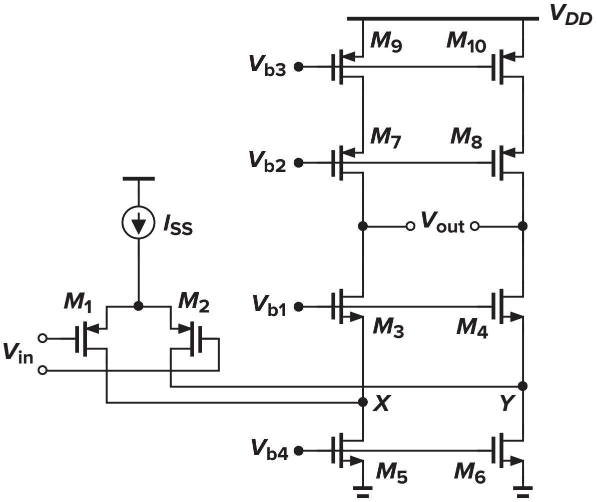


Figure 9.15 Folded-cascode op amp with cascode PMOS loads.

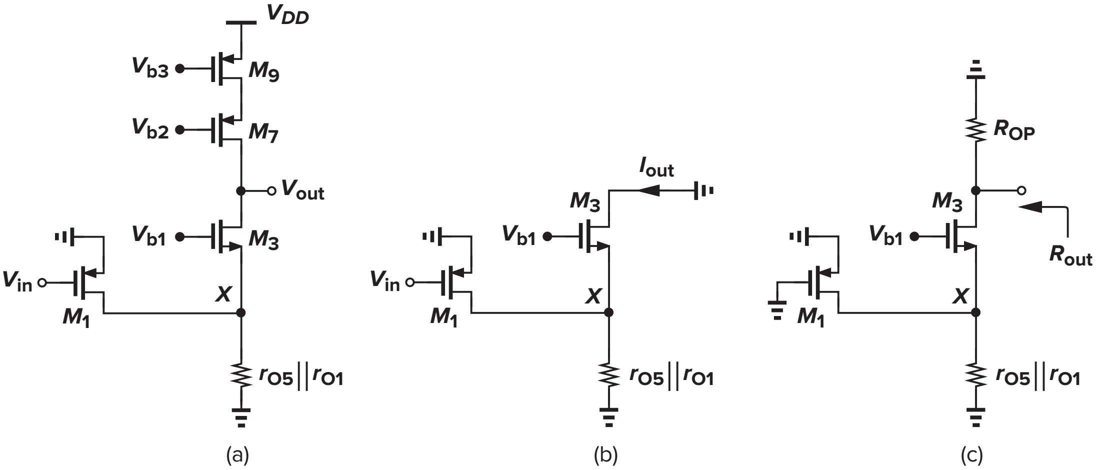


Figure 9.16 (a) Half circuit of folded cascode op amp, (b) equivalent circuit for G_m calculation, and (c) equivalent circuit for R_{out} calculation.

The reader is encouraged to repeat this calculation without neglecting the current drawn by $r_{O5}||r_{O1}$ in Fig. 9.16(b).

How does this value compare with the gain of a telescopic op amp? For comparable device dimensions and bias currents, the PMOS input differential pair exhibits a lower transconductance than does an NMOS pair. Furthermore, r_{O1} and r_{O5} appear in parallel, reducing the output impedance, especially because M_5 carries the currents of both the input device and the cascode branch. As a consequence, the gain in (9.17) is usually two to three times lower than that of a comparable telescopic cascode.

It is also worth noting that the pole at the "folding point," i.e., the sources of M_3 and M_4, is quite closer to the origin than that associated with the source of cascode devices in a telescopic topology. In Fig. 9.17(a), C_{tot} arises from C_{GS3}, C_{SB3}, C_{DB1}, and C_{GD1}. By contrast, in Fig. 9.17(b), C_{tot} contains additional contributions due to C_{GD5} and C_{DB5}, typically significant components because M_5 must be wide enough to carry a large current with a small overdrive.

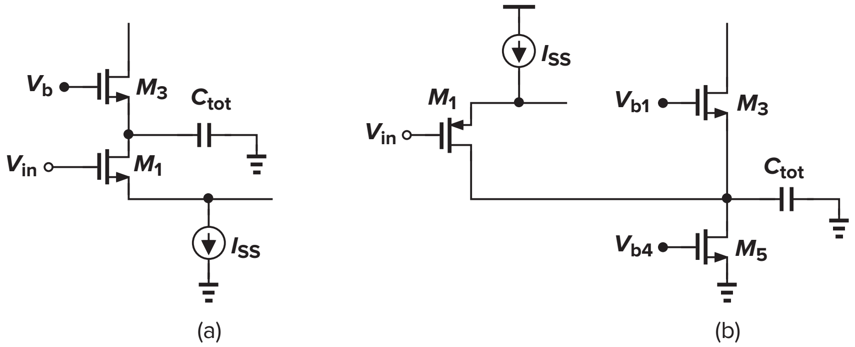


Figure 9.17 Effect of device capacitance on the nondominant pole in telescopic and folded-cascode op amps.

A folded-cascode op amp may incorporate NMOS input devices and PMOS cascode transistors. Illustrated in Fig. 9.18, such a circuit potentially provides a higher gain than the op amp of Fig. 9.15 because of the greater mobility of NMOS devices, but at the cost of lowering the pole at the folding points. To understand why, note that the pole at node X is given by the product of $1/(g_{m3}+g_{mb3})$ and the total capacitance at this node (if the output pole is dominant). The magnitude of both of these components is relatively high: M_3 suffers from a low transconductance, and M_5 contributes substantial capacitance because it must be wide enough to carry the drain currents of both M_1 and M_3. In fact, for comparable bias currents, M_5–M_6 in Fig. 9.18 may be several times wider than M_5–M_6 in Fig. 9.15. For applications sensitive to flicker noise, the PMOS-input op amp is preferable (Sec. 9.12).

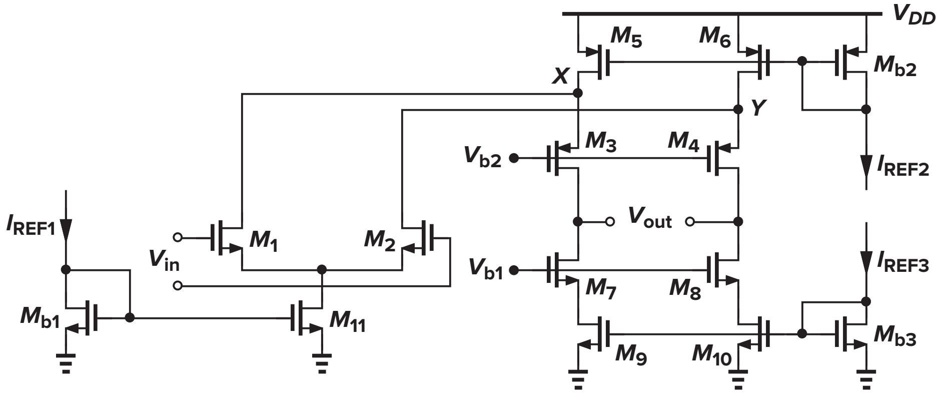


Figure 9.18 Realization of a folded-cascode op amp.

9.2.5 Folded-Cascode Properties

Our study thus far suggests that the overall voltage swing of a folded-cascode op amp is only slightly higher than that of a telescopic configuration. This advantage comes at the cost of higher power dissipation, lower voltage gain, lower pole frequencies, and, as explained in Sec. 9.12, higher noise. Nonetheless, folded-cascode op amps are used more widely for two reasons: (1) their input and output CM levels can be chosen equal without limiting the output swings, and (2) compared to telescopic cascodes, they can accommodate a wider input CM range. Let us elaborate on these properties.

Consider the closed-loop amplifier of Fig. 9.19(a), assuming a folded-cascode op amp. We can draw the circuit as shown in Fig. 9.19(b) or Fig. 9.19(c), noting that the input and output CM levels are equal. With a high open-loop gain, the gate voltages of M_1 and M_2 swing negligibly while V_X and V_Y can reach within two overdrives of ground or V_{DD}. This should be compared with the swings in Fig. 9.10.

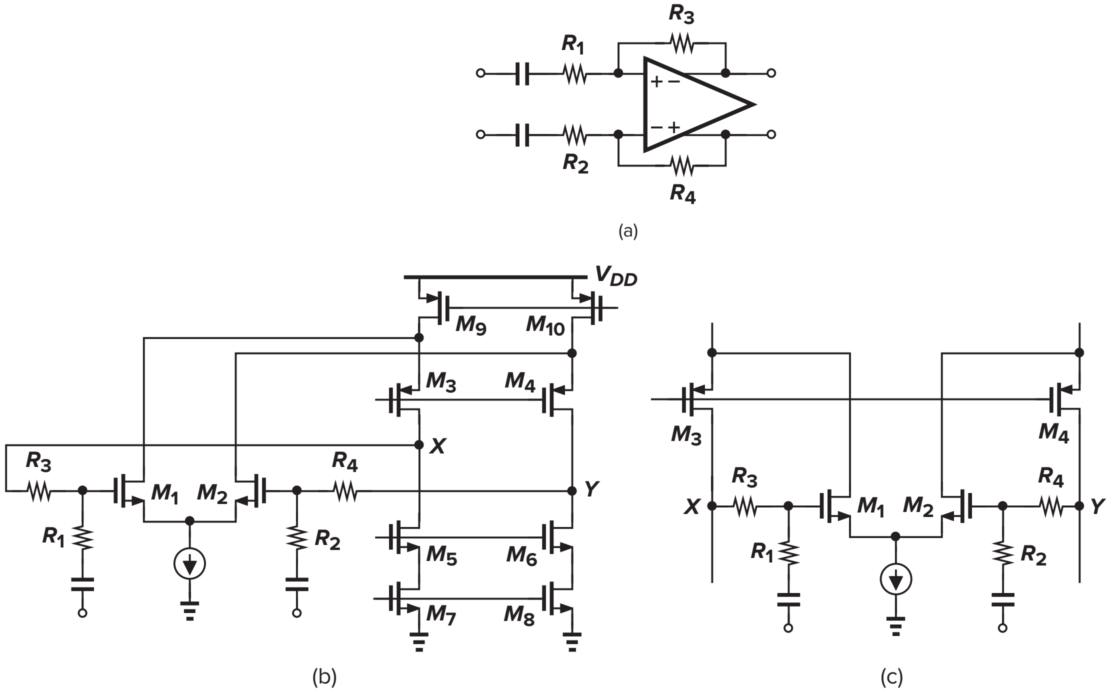


Figure 9.19 (a) Feedback amplifier, (b) implementation using a folded-cascode op amp, and (c) alternative drawing to find allowable swings.

In feedback topologies where the input and output CM levels need not be equal, the folded cascode allows a wider input CM range than does the telescopic cascode. In Fig. 9.18, for example, $V_{in,CM}$ must exceed $V_{GS1,2} + (V_{GS11} - V_{TH11})$, but it can be as high as $V_{b2} + |V_{GS3}| + V_{TH1,2}$ before M_1 and M_2 enter the triode region. Note that this upper bound can be *greater* than V_{DD} (why?). Similarly, a PMOS-input configuration can handle input CM levels as low as zero.

9.2.6 Design Procedure

We now deal with the design of folded-cascode op amps to reinforce the foregoing concepts.

▶ **Example 9.9**

Design a folded-cascode op amp with an NMOS input pair (Fig. 9.18) to satisfy the following specifications: $V_{DD} = 3$ V, differential output swing $= 3$ V, power dissipation $= 10$ mW, and voltage gain $= 2000$. Use the same device parameters as in Example 9.5.

Solution

As with the telescopic cascode of the previous example, we begin with the power and swing specifications. Allocating 1.5 mA to the input pair, 1.5 mA to the two cascode branches, and the remaining 330 μA to the three current mirrors, we first consider the devices in each cascode branch. Since M_5 and M_6 must each carry 1.5 mA, we allow an overdrive of 500 mV for these transistors so as to keep their width to a reasonable value. To M_3–M_4, we allocate 400 mV and to M_7–M_{10}, 300 mV. Thus, $(W/L)_{5,6} = 400$, $(W/L)_{3,4} = 313$, and $(W/L)_{7-10} = 278$. Since the minimum and maximum output levels are equal to 0.6 V and 2.1 V, respectively, the optimum output common-mode level is 1.35 V.

The minimum dimensions of M_1–M_2 are dictated by the minimum input common-mode level, $V_{GS1} + V_{OD11}$. For example, if the input and the output CM levels are equal (Fig. 9.20), then $V_{GS2} + V_{OD11} = 1.35$ V. With

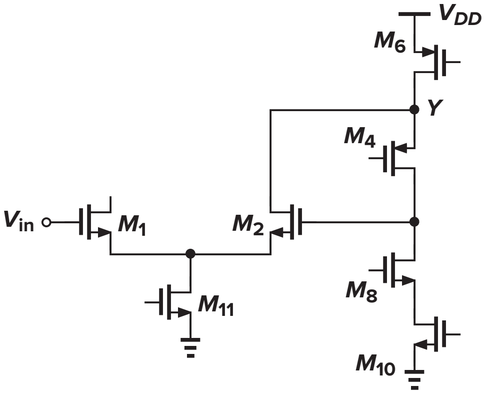


Figure 9.20 Folded-cascode op amp with input and output shorted.

$V_{OD11} = 0.4$ V as an initial guess, we have $V_{GS1} = 0.95$ V, obtaining $V_{OD1,2} = 0.95 - 0.7 = 0.25$ V, and hence $(W/L)_{1,2} = 400$. The maximum dimensions of M_1 and M_2 are determined by the tolerable input capacitance and the capacitance at nodes X and Y in Fig. 9.18.

We now calculate the small-signal gain. Using $g_m = 2I_D/(V_{GS} - V_{TH})$, we have $g_{m1,2} = 0.006$ A/V, $g_{m3,4} = 0.0038$ A/V, and $g_{m7,8} = 0.05$ A/V. For $L = 0.5\ \mu$m, $r_{O1,2} = r_{O7-10} = 13.3$ kΩ, and $r_{O3,4} = 2r_{O5,6} = 6.67$ kΩ. It follows that the impedance seen looking into the drain of M_7 (or M_8) is equal to 8.8 MΩ whereas, owing to the limited intrinsic gain of M_3 (or M_4), that seen looking into the drain of M_3 is equal to 66.5 kΩ. The overall gain is therefore limited to about 400.

In order to increase the gain, we first observe that $r_{O5,6}$ is quite lower than $r_{O1,2}$. Thus, the length of M_5–M_6 must be increased. Also, the transconductance of M_1–M_2 is relatively low and can be increased by widening these transistors. Finally, we may decide to double the intrinsic gain of M_3 and M_4 by doubling both their length and their width, but at the cost of increasing the capacitance at nodes X and Y. We leave the exact choice of the device dimensions as an exercise for the reader. Note that the op amp must incorporate common-mode feedback (Sec. 9.7).

Telescopic and folded-cascode op amps can also be designed to provide a single-ended output. Shown in Fig. 9.21(a) is an example, where a PMOS cascode current mirror converts the differential currents of M_3

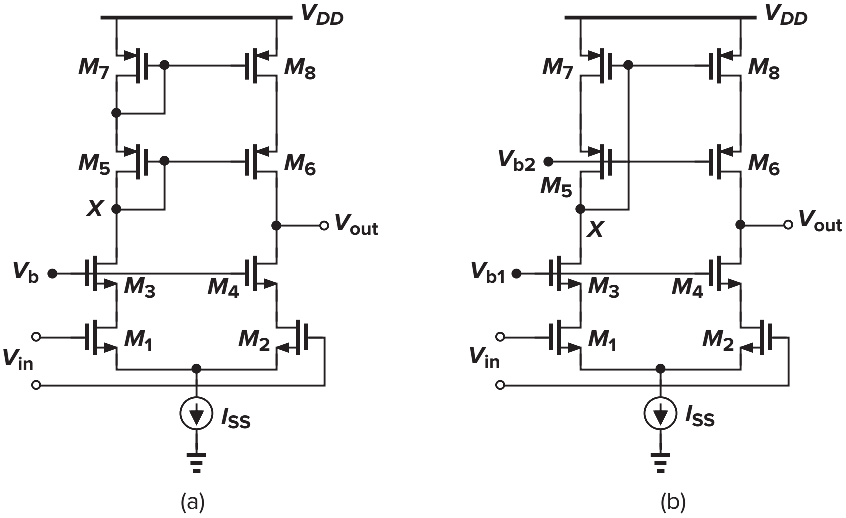


Figure 9.21 Cascode op amps with single-ended output.

and M_4 to a single-ended output voltage. In this implementation, however, $V_X = V_{DD} - |V_{GS5}| - |V_{GS7}|$, limiting the maximum value of V_{out} to $V_{DD} - |V_{GS5}| - |V_{GS7}| + |V_{TH6}|$ and "wasting" one PMOS threshold voltage in the swing (Chapter 5). To resolve this issue, the PMOS load can be modified to a low-voltage cascode (Chapter 5), as shown in Fig. 9.21(b), so that M_7 and M_8 are biased at the edge of the triode region. Similar ideas apply to folded-cascode op amps as well.

The circuit of Fig. 9.21(a) suffers from two disadvantages with respect to its differential counterpart in Fig. 9.8(b). First, it provides only half the output voltage swing. Second, it contains a mirror pole at node X (Chapter 5), thus limiting the speed of feedback systems employing such an amplifier. It is therefore preferable to use the differential topology, although it requires a feedback loop to define the output common-mode level (Sec. 9.7).

9.3 ■ Two-Stage Op Amps

The op amps studied thus far exhibit a "one-stage" nature in that they allow the small-signal current produced by the input pair to flow directly through the output impedance, i.e., they perform voltage-to-current conversion only once. The gain of these topologies is therefore limited to the product of the input pair transconductance and the output impedance. We have also observed that cascoding in such circuits increases the gain while limiting the output swings.

In some applications, the gain and/or the output swings provided by cascode op amps are not adequate. For example, a modern op amp must operate with supply voltages as low as 0.9 V while delivering single-ended output swings as large as 0.8 V. In such cases, we resort to "two-stage" op amps, with the first stage providing a high gain and the second, large swings (Fig. 9.22). In contrast to cascode op amps, a two-stage configuration isolates the gain and swing requirements.

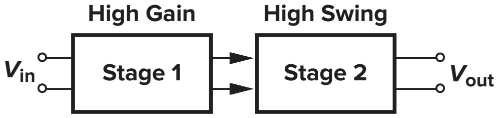


Figure 9.22 Two-stage op amp.

Each stage in Fig. 9.22 can incorporate various amplifier topologies studied in previous sections, but the second stage is typically configured as a simple common-source stage so as to allow maximum output swings. Figure 9.23 shows an example, where the first and second stages exhibit gains equal to $g_{m1,2}(r_{O1,2}\|r_{O3,4})$ and $g_{m5,6}(r_{O5,6}\|r_{O7,8})$, respectively. The overall gain is therefore comparable to that

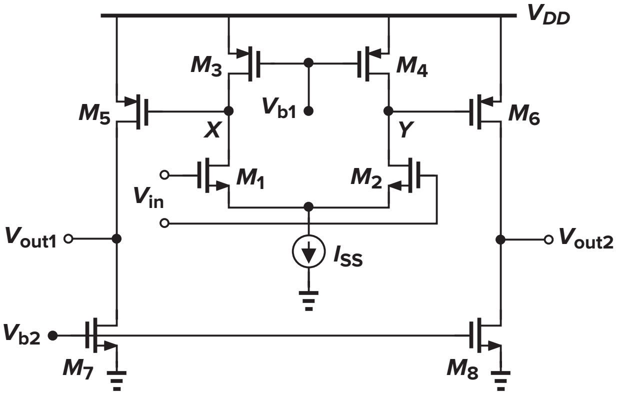


Figure 9.23 Simple implementation of a two-stage op amp.

of a cascode op amp, but the swing at V_{out1} and V_{out2} is equal to $V_{DD} - |V_{OD5,6}| - V_{OD7,8}$, the highest possible value.[4]

To obtain a higher gain, the first stage can incorporate cascode devices, as depicted in Fig. 9.24. With a gain of, say, 10 in the output stage, the voltage swings at X and Y are quite small, allowing optimization of M_1–M_8 for higher gain. The overall voltage gain can be expressed as

$$\begin{aligned} A_v \approx & \{g_{m1,2}[(g_{m3,4} + g_{mb3,4})r_{O3,4}r_{O1,2}] \| [(g_{m5,6} + g_{mb5,6})r_{O5,6}r_{O7,8}]\} \\ & \times [g_{m9,10}(r_{O9,10} \| r_{O11,12})] \end{aligned} \tag{9.18}$$

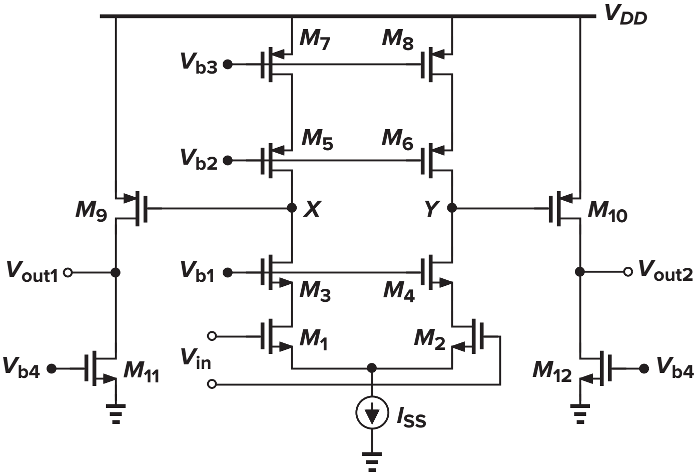


Figure 9.24 Two-stage op amp employing cascoding.

A two-stage op amp can provide a single-ended output. One method is to convert the differential currents of the two output stages to a single-ended voltage. Illustrated in Fig. 9.25, this approach maintains the differential nature of the first stage, using only the current mirror M_7–M_8 to generate a single-ended output.

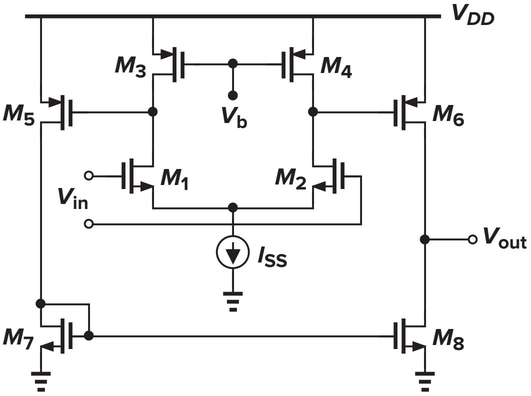


Figure 9.25 Two-stage op amp with single-ended output.

[4] One can replace M_7 and M_8 with resistors to allow greater swings, but the gain would be limited.

Can we cascade more than two stages to achieve a higher gain? As explained in Chapter 10, each gain stage introduces at least one pole in the open-loop transfer function, making it difficult to guarantee stability in a feedback system using such an op amp. For this reason, op amps having more than two stages are rarely used. Exceptions are described in [1, 2, 3].

9.3.1 Design Procedure

The design of two-stage op amps is somewhat more complex. We present a simple example here and more detailed designs in Chapter 11.

▶ **Example 9.10**

Design the two-stage op amp of Fig. 9.23 for $V_{DD} = 1$ V, $P = 1$ mW, a differential output swing of 1 V$_{pp}$, and a gain of 100. Use the same device parameters as in Example 9.7, but assume that $V_{THN} = 0.3$ V and $V_{THP} = -0.35$ V.

Solution

We allocate a bias current of 960 μA to M_1–M_8, leaving 40 μA for the bias branches that generate V_{b1} and V_{b2}. Let us split the current budget equally between the first and second stages, i.e., assume that $I_{D1} = \cdots = I_8 = 120\ \mu$A.

Since the second stage is likely to provide a voltage gain of 5 to 10, the output swing of the *first* stage need not be large. Specifically, if the second stage is designed for a gain of 5 and a single-ended output swing of 0.5 V$_{pp}$, the first stage need only sustain 0.1 V$_{pp}$ at X (or Y). The choice of overdrive voltages for M_1–M_4 and I_{SS} is therefore quite relaxed, i.e., $|V_{OD3}| + |V_{OD1}| + V_{ISS} = 1\text{ V} - 0.1\text{ V} = 0.9$ V. But we must consider two points: (1) recall from Chapter 7 that the noise contributed by current sources M_3 and M_4 is minimized by maximizing their overdrive voltage, and (2) the gain (and noise) requirements dictate a high g_m for M_1 and M_2 and, inevitably, a low overdrive voltage. In fact, the latter point typically translates to subthreshold operation for the input devices, yielding a maximum g_m of $I_D/(\xi V_T) \approx (325\ \Omega)^{-1}$ with $\xi = 1.5$. But, we ignore subthreshold operation in this example.

How large can the overdrive of M_3 and M_4 be? Since $V_{DS3,4} = V_{GS5,6}$ in this case, the upper bound may be imposed by M_5 and M_6 rather than by the first stage. For example, if the design of the second stage eventually yields $|V_{GS5,6}| = 400$ mV, and if V_X (or V_Y) can rise by 50 mV (for a 100-mV$_{pp}$ swing), then M_3 and M_4 experience a minimum $|V_{DS}|$ of 350 mV. We must therefore revisit this allocation after the second stage is designed.

For a single-ended output swing of 0.5 V$_{pp}$, we can choose 200 mV and 300 mV for the overdrives of the output NMOS and PMOS devices, respectively. With $I_D = 120\ \mu$A, we then compute the W/L values of these transistors. However, this allocation faces two issues: (1) the large overdrive of M_5 and M_6 may translate to an inadequately low $g_m = 2I_D/(V_{GS} - V_{TH})$, and (2) the small overdrive of M_7 and M_8 gives them a high noise current. For these reasons, we swap the overdrive allocation, giving 300 mV to M_7 and M_8 and 200 mV to M_5 and M_6. The penalty is the larger W/L of the latter pair and hence a greater capacitance at X and Y.

We begin the calculations from the output stage. With $|I_D| = 120\ \mu$A and the above overdrives, we have $g_{m5,6} = 2|I_D/(V_{GS} - V_{TH})| = (833\ \Omega)^{-1}$, $r_{O5,6} = 1/(\lambda|I_D|) = 42$ kΩ, and $r_{O7,8} = 83$ kΩ (for the minimum channel length of 0.5 μm). The second stage thus provides a gain of about 33, allowing even smaller voltage swings for the first stage. The corresponding device dimensions are $(W/L)_{5,6} = 200$ and $(W/L)_{7,8} = 44$.

Returning to the first stage in Fig. 9.23, we note that $V_{DS3,4} = |V_{GS5,6}| = 550$ mV. Transistors M_3 and M_4 can therefore operate with an overdrive as high as 500 mV (if we still assume V_X or V_Y can rise by 50 mV from the bias value) but require a $|V_{GS}|$ of 500 mV + $|V_{THP}| = 850$ mV, and hence $V_{b1} = 150$ mV. Such a low V_{b1} may cause difficulty in the design of the current mirror driving M_3 and M_4. Instead, we choose $|V_{GS3,4} - V_{THP}| = 400$ mV, obtaining $(W/L)_{3,4} = 50$, $g_{m3,4} = 1/(1.7\text{ k}\Omega)$, and $r_{O3,4} = 83$ kΩ (for $L = 0.5\ \mu$m).

The input transistors, M_1 and M_2, exhibit an output resistance of 83 kΩ (with $L = 0.5\ \mu$m) and can have an overdrive as large as 0.5 V. However, with such an overdrive, $g_{m1,2}/g_{m3,4} = |V_{GS3,4} - V_{THP}|/(V_{GS1,2} - V_{THN}) = 4/5$, implying that the PMOS devices contribute substantial noise. For this reason, we choose an overdrive of 100 mV for M_1 and M_2, arriving at $g_{m1,2} = 1/(420\ \Omega)$, $(W/L)_{1,2} = 400$, and a voltage gain of $g_{m1,2}(r_{O1}||r_{O3}) = 66$ for the first stage.

This design provides an overall gain of more than 2,000, primarily because of the low bias current and the use of an older technology. As explained in Chapter 11, nanometer two-stage op amps suffer from much lower gains. ◀

9.4 ■ Gain Boosting

9.4.1 Basic Idea

The limited gain of the one-stage op amps studied in Sec. 9.2 and the difficulties in using two-stage op amps at high speeds have motivated extensive work on new topologies. Recall that in one-stage op amps, such as telescopic and folded-cascode topologies, the objective is to maximize the output impedance so as to attain a high voltage gain. The idea behind gain boosting is to further increase the output impedance without adding more cascode devices [4, 5]. We neglect body effect for simplicity, but it can be readily included at the end.

First Perspective Suppose a transistor is preceded by an ideal voltage amplifier as shown in Fig. 9.26(a).

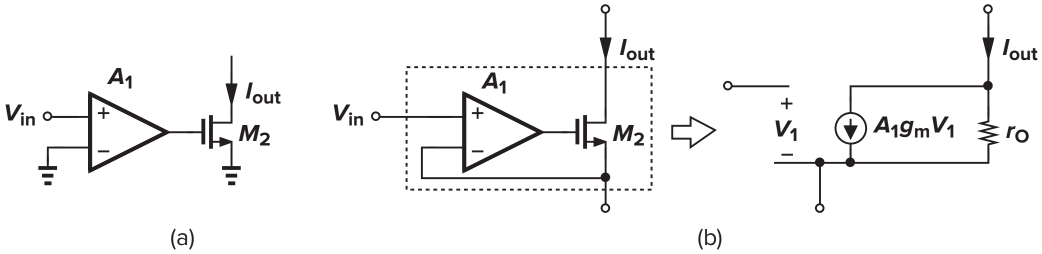


Figure 9.26 (a) Transistor preceded by a voltage amplifier, and (b) equivalent circuit.

We note that the overall circuit exhibits a transconductance of A_1g_m and a voltage gain of $-A_1g_mr_O$ (why?). We thus surmise that this arrangement can be viewed as a three-terminal device (a "supertransistor") having a transconductance of A_1g_m and an output resistance of r_O [Fig. 9.26(b)]. We neglect body effect in this section.

Let us now incorporate this new device in a familiar topology and examine the circuit's behavior. We begin with the degenerated stage depicted in Fig. 9.27(a) and wish to compute its transconductance (with the output shorted to ac ground). Since R_S carries I_{out}, the small-signal gate voltage is given by $(V_{in} - R_SI_{out})A_1$, yielding a gate-source voltage of $(V_{in} - R_SI_{out})A_1 - R_SI_{out}$ and hence $I_{out} = g_m[(V_{in} - R_SI_{out})A_1 - R_SI_{out}]$. It follows that

$$\frac{I_{out}}{V_{in}} = \frac{A_1g_m}{1 + (A_1 + 1)g_mR_S} \tag{9.19}$$

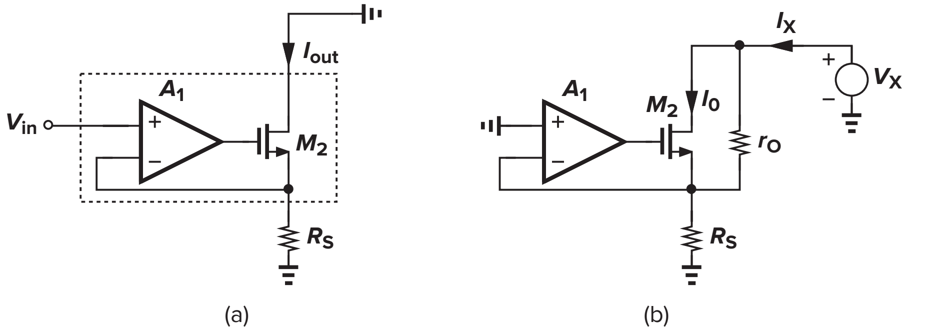


Figure 9.27 Arrangements for calculation of (a) transconductance, and (b) output resistance.

Without A_1, the transconductance would be equal to $g_m/(1+g_m R_S)$. Interestingly, the equivalent transconductance has risen by a factor of A_1 in the numerator and $A_1 + 1$ in the denominator, revealing that the model shown in Fig. 9.26(b) is not quite correct. However, since in practice $A_1 \gg 1$, the error introduced by this model is acceptably low.

How about the output resistance of the degenerated stage? From the setup in Fig. 9.27(b), we can express the voltage drop across R_S as $I_X R_S$ and the gate voltage of M_2 as $-A_1 I_X R_S$. That is, $I_0 = (-A_1 R_S I_X - R_S I_X) g_m$. Also, r_O carries a current equal to $(V_X - R_S I_X)/r_O$. We now have

$$I_X = (-A_1 R_S - R_S) g_m I_X + \frac{V_X - R_S I_X}{r_O} \tag{9.20}$$

and

$$R_{out} = r_O + (A_1 + 1) g_m r_O R_S + R_S \tag{9.21}$$

Without A_1, the output resistance would be equal to $r_O + g_m r_O R_S + R_S$.

Equation (9.21) is a remarkable result, suggesting that the output resistance of the circuit is "boosted," as if the transconductance of M_2 were raised by a factor of $A_1 + 1$. This increase in R_{out} is afforded while the degenerated stage retains its voltage headroom. We can see that the allowable voltage swing at the drain of M_2 is approximately the same for this structure and a simple degenerated transistor.

▶ Example 9.11

Determine the resistance seen at the source of M_2 in Fig. 9.28(a) if $\gamma = 0$.

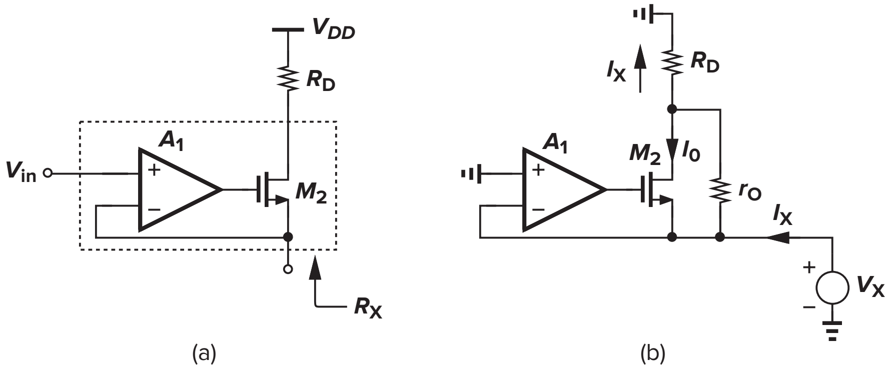


Figure 9.28

Solution

In the setup shown in Fig. 9.28(b), the small-signal gate voltage is equal to $-A_1 V_X$, and hence $I_0 = (-A_1 V_X - V_X) g_m$. Also, R_D carries a current of I_X, generating a voltage equal to $I_X R_D$ at the drain with respect to ground. Since the current flowing downward through r_O is given by $(I_X R_D - V_X)/r_O$, we have at the source node

$$\frac{I_X R_D - V_X}{r_O} + (-A_1 V_X - V_X) g_m + I_X = 0 \tag{9.22}$$

and

$$R_X = \frac{R_D + r_O}{1 + (A_1 + 1) g_m r_O} \tag{9.23}$$

Without A_1, this resistance would be equal to $(R_D + r_O)/(1 + g_m r_O)$. This example too suggests that the transconductance of M_2 is raised by a factor of $A_1 + 1$.

◀

In summary, the addition of the auxiliary amplifier in Fig. 9.26(b) raises the equivalent g_m of M_2 by a factor of $A_1 + 1$, thereby boosting the output impedance of the stage. We surmise from $A_v = -G_m R_{out}$ that the voltage gain can also be boosted, but where should the input be applied? As in a simple cascode stage, let us replace the degeneration resistor with a voltage-to-current converter (Fig. 9.29), obtaining an output impedance equal to $r_{O2} + (A_1 + 1)g_{m2}r_{O2}r_{O1} + r_{O1}$. The short-circuit transconductance is nearly equal to g_{m1} because the resistance seen looking into the source of M_2 is obtained from (9.23) with $R_D = 0$ and is given by $r_{O2}/[1 + (A_1 + 1)g_{m2}r_{O2}] \approx [(A_1 + 1)g_{m2}]^{-1}$, a value much less than r_{O1}. It follows that

$$|A_v| \approx g_{m1}[r_{O2} + (A_1 + 1)g_{m2}r_{O2}r_{O1} + r_{O1}] \tag{9.24}$$

$$\approx g_{m1}g_{m2}r_{O1}r_{O2}(A_1 + 1) \tag{9.25}$$

As explained later in this section, this "gain-boosting" technique can be applied to cascode differential pairs and op amps as well.

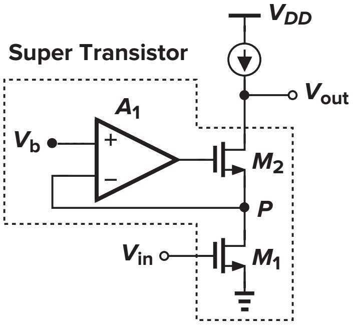


Figure 9.29 Basic gain-boosted stage.

Second Perspective Consider the degenerated stage shown in Fig. 9.30(a). We wish to increase the output resistance without stacking more cascode devices. Recall from Chapter 3 that if the drain voltage changes by ΔV, then the source voltage changes by $\Delta V_S = R_S/[r_O + (1 + g_m r_O)R_S]$ (with $\gamma = 0$), producing a change in the voltage across R_S and hence in the drain current. We can loosely view the effect as voltage division between R_S and $g_m r_O R_S$.

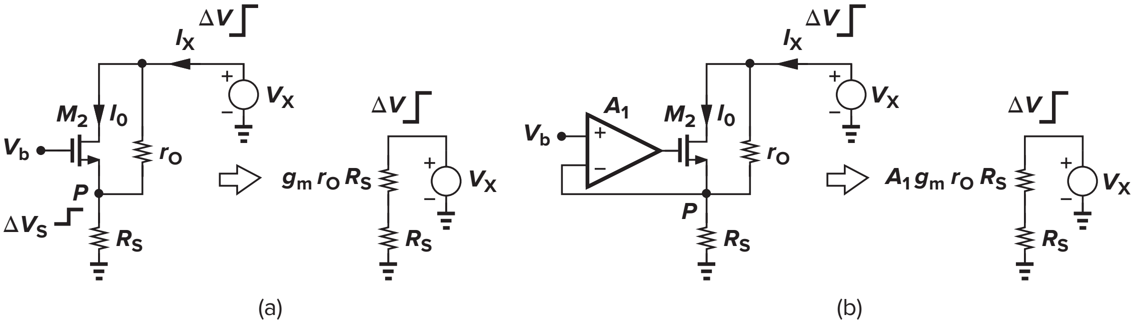


Figure 9.30 Response of (a) degenerated CS stage and (b) gain-boosted stage to a change in output voltage.

We now make an important observation. The change in the drain current in response to ΔV can be suppressed if two conditions hold: (a) the voltage across R_S remains constant, and (b) the current flowing through R_S remains equal to the drain current.[5] How should we keep V_P constant? We can compare V_P to

[5] A constant voltage source tied from P to ground allows the former condition but not the latter.

a "reference" voltage by means of an op amp and return the resulting error to a point in the circuit so as to ensure that V_P "tracks" the reference. Illustrated in Fig. 9.30(b), the idea is to apply the error, $A_1(V_b - V_P)$, to the gate of M_2, forcing V_P to be equal to V_b if the loop gain is large. The above two conditions are thus satisfied. For example, if the drain voltage rises, V_P also tends to rise, but, as a result, the gate voltage falls, reducing the current drawn by M_2. As derived below, this effect can be approximately viewed as voltage division between R_S and $A_1 g_m r_O R_S$. For $A_1 \to \infty$, V_P is "pinned" to V_b and the drain current is exactly equal to V_b/R_S regardless of the drain voltage. This topology is also called a "regulated cascode" as amplifier A_1 monitors and regulates the output current.

▶ Example 9.12

Figure 9.31 shows the regulated cascode subjected to an output impedance test. Determine the small-signal values of V_P, V_G, I_0, and I_{ro}. Assume that $(A_1 + 1)g_m r_O R_S$ is large.

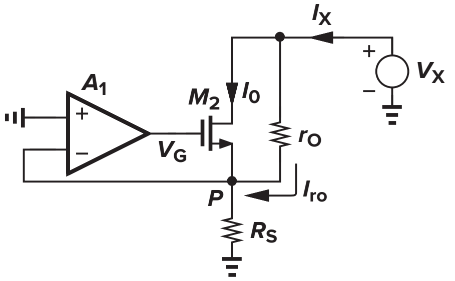


Figure 9.31

Solution

We know from our analysis of Fig. 9.27(b) that

$$V_X = [r_O + (A_1 + 1)g_m r_O R_S + R_S]I_X \tag{9.26}$$

and hence

$$V_P = I_X R_S \tag{9.27}$$

$$= \frac{R_S}{r_O + (A_1 + 1)g_m r_O R_S + R_S} V_X \tag{9.28}$$

If $(A_1 + 1)g_m r_O R_S$ is large, then $V_P \approx V_X/[(A_1 + 1)g_m r_O]$, implying that the amplifier suppresses the change in the voltage across R_S by another factor of $A_1 + 1$ compared to the case of a simple degenerated transistor. We also have

$$V_G = -A_1 V_P \tag{9.29}$$

$$= \frac{-A_1 R_S}{r_O + (A_1 + 1)g_m r_O R_S + R_S} V_X \tag{9.30}$$

The small-signal gate-source voltage is equal to $V_G - V_P \approx -V_X/(g_m r_O)$, yielding $I_0 \approx -V_X/r_O$. Moreover,

$$I_{ro} = \frac{V_X - V_P}{r_O} \tag{9.31}$$

$$= \frac{r_O + (A_1 + 1)g_m r_O R_S}{r_O + (A_1 + 1)g_m r_O R_S + R_S} \frac{V_X}{r_O} \tag{9.32}$$

$$\approx \frac{V_X}{r_O} \tag{9.33}$$

Interestingly, I_0 and I_{ro} are nearly equal and opposite. That is, the amplifier adjusts the gate voltage such that the change in the intrinsic drain current, I_0, almost cancels the current drawn by r_O. We say that the small-signal current of M_2 circulates through r_O.

In summary, the above two perspectives portray two principles behind the gain-boosting technique: the amplifier boosts the g_m of the cascode device, or the amplifier regulates the output current by monitoring and pinning the source voltage.

9.4.2 Circuit Implementation

In this section, we deal with the implementation of the auxiliary amplifier in the regulated cascode and extend the gain-boosting technique to op amps. The simplest realization of A_1 is a common-source stage, as shown in Fig. 9.32(a). If I_1 is ideal, then $|A_1| = g_{m3}r_{O3}$, yielding $|V_{out}/V_{in}| \approx g_{m1}r_{O1}g_{m2}r_{O2}(g_{m3}r_{O3}+1)$, as in a *triple* cascode. However, this topology limits the output voltage swing because the minimum voltage at node P is dictated by V_{GS3} rather than the overdrive of M_1. We note that V_{out} must remain above $V_{GS3} + (V_{GS2} - V_{TH2})$ here.

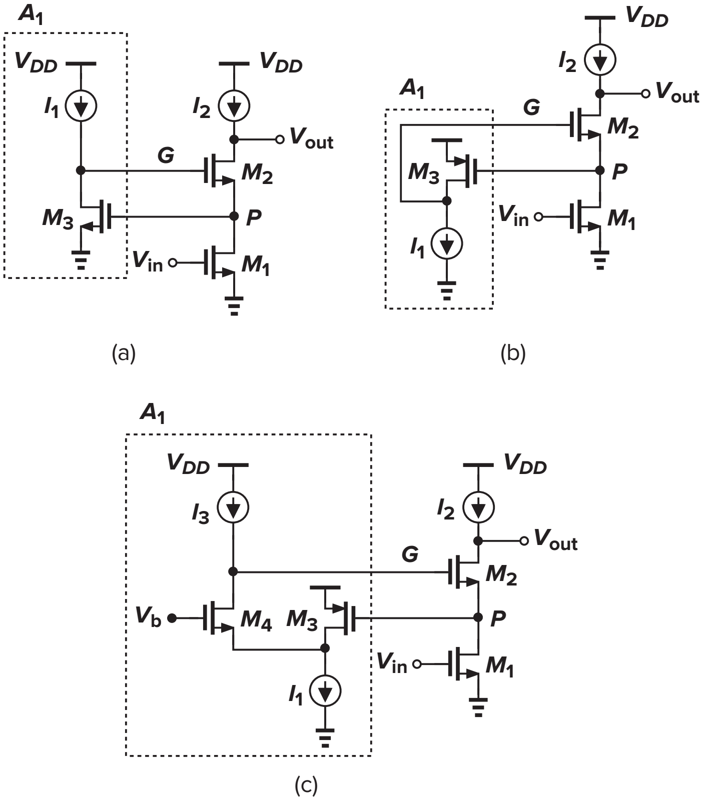


Figure 9.32 Gain-boosted amplifier using (a) an NMOS CS stage, (b) a PMOS CS stage, and (c) a folded-cascode stage.

To avoid this headroom limitation, we consider a PMOS common-source stage for A_1 [Fig. 9.32(b)]. The operation and gain-boosting properties remain the same, but V_P can now be as low as the overdrive of M_1. Unfortunately, M_3 may enter the triode region here because the gate voltage of M_2 tends to be too high for the drain of M_3. Specifically, if we target $V_P = V_{GS1} - V_{TH1}$, then $V_G = V_{GS2} + V_{GS1} - V_{TH1}$, revealing that the drain of M_3 is higher than its gate by V_{GS2}. If $V_{GS2} > |V_{TH3}|$, M_3 resides in the triode region.

The above analysis implies that we must insert one more stage in the feedback loop so as to reach compatible bias levels between consecutive stages. Let us interpose an NMOS common-gate stage between M_3 and the gate of M_2 [Fig. 9.32(c)]. The reader recognizes the resulting A_1 topology as a folded cascode, but we also observe that M_4 provides an upward level shift from its source to its drain, allowing V_G to be higher than the drain voltage of M_3.

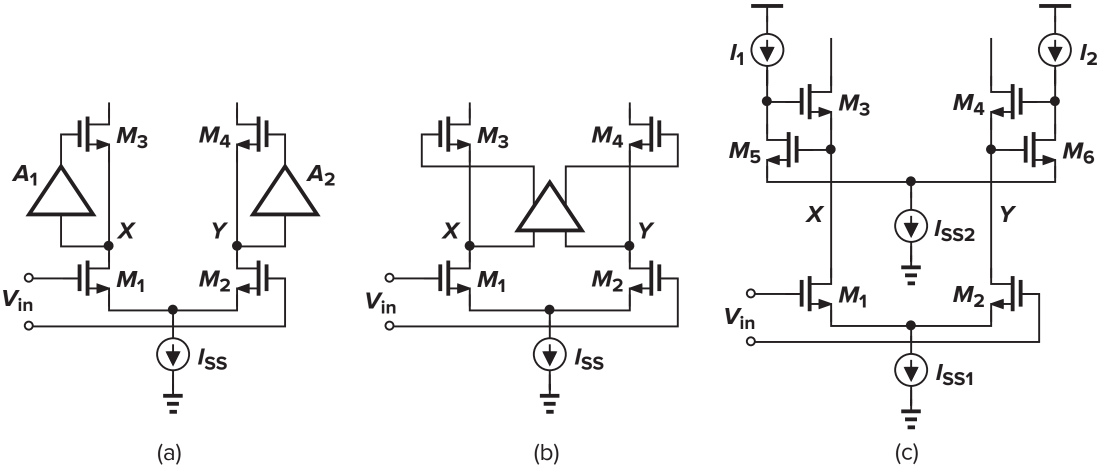
▶ **Example 9.13**

Determine the allowable range for V_b in Fig. 9.32(c).

Solution

The minimum value of V_b places I_1 at the edge of the triode region, i.e., $V_{b,min} = V_{GS4} + V_{I1}$. The maximum value biases M_4 at the edge of the triode region, i.e., $V_{b,max} = V_{GS2} + V_P + V_{TH4}$. Thus, V_b has a comfortably wide range and need not be precise. ◀

We now apply gain boosting to a differential cascode stage, as shown in Fig. 9.33(a). Since the signals at nodes X and Y are differential, we surmise that the two single-ended gain-boosting amplifiers A_1 and A_2 can be replaced by one differential amplifier [Fig. 9.33(b)]. Following the topology of Fig. 9.32(a), we implement the differential auxiliary amplifier as shown in Fig. 9.33(c), but noting that the minimum level at the drain of M_3 is equal to $V_{OD3} + V_{GS5} + V_{ISS2}$, where V_{ISS2} denotes the voltage required across I_{SS2}. In a simple differential cascode, on the other hand, the minimum would be approximately one threshold voltage lower.



(a) (b) (c)

Figure 9.33 Boosting the output impedance of a differential cascode stage.

The voltage swing limitation in Fig. 9.33(c) results from the fact that the gain-boosting amplifier incorporates an NMOS differential pair. If nodes X and Y are sensed by a PMOS pair, the minimum value of V_X and V_Y is not dictated by the gain-boosting amplifier. Now recall from Sec. 9.2 that the minimum input CM level of a folded-cascode stage using a PMOS input pair can be zero. Thus, we employ such a topology for the gain-boosting amplifier, arriving at the circuit shown in Fig. 9.34. Here, the minimum allowable level of V_X and V_Y is given by $V_{OD1,2} + V_{ISS1}$.

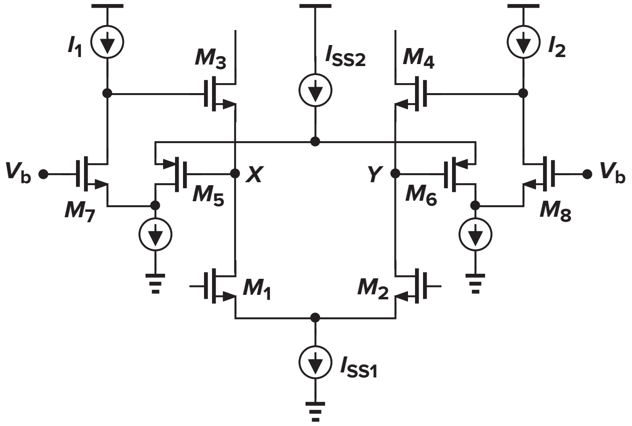


Figure 9.34 Folded-cascode circuit used as auxiliary amplifier.

▶ Example 9.14

Calculate the output impedance of the circuit shown in Fig. 9.34.

Solution

Using the half-circuit concept and replacing the ideal current sources with transistors, we obtain the equivalent depicted in Fig. 9.35. The voltage gain from X to P is approximately equal to $g_{m5}R_{out1}$, where $R_{out1} \approx [g_{m7}r_{O7}(r_{O9}\|r_{O5})]\|(g_{m11}r_{O11}r_{O13})$. Thus, $R_{out} \approx g_{m3}r_{O3}r_{O1}g_{m5}R_{out1}$. In essence, since the output impedance of a cascode is boosted by a folded-cascode stage, the overall output impedance is similar to that of a "quadruple" cascode.

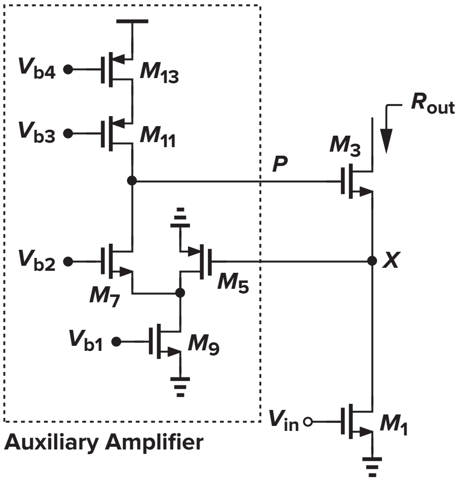


Figure 9.35

◀

Regulated cascodes can also be utilized in the load current sources of a cascode op amp. Shown in Fig. 9.36(a), such a topology boosts the output impedance of the PMOS current sources as well, thereby achieving a very high voltage gain. To allow maximum swings at the output, amplifier A_2 must employ an NMOS-input folded-cascode differential pair. Similar ideas apply to folded-cascode op amps [Fig. 9.36(b)].

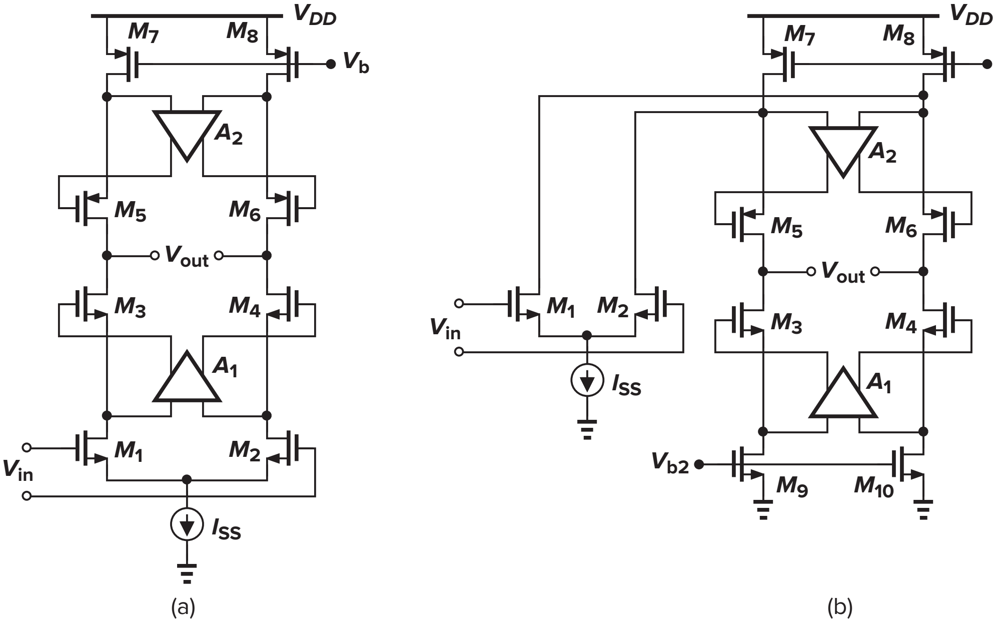


Figure 9.36 Gain boosting applied to both signal path and load devices.

9.4.3 Frequency Response

Recall that the premise behind gain boosting is to increase the gain without adding a second stage or more cascode devices. Does this mean that the op amps of Fig. 9.36 have a one-stage nature? After all, the gain-boosting amplifier introduces its own pole(s). In contrast to two-stage op amps, where the entire signal experiences the poles associated with each stage, in a gain-boosted op amp, most of the signal flows directly through the cascode devices to the output. Only a small "error" component is processed by the auxiliary amplifier and "slowed down."

In order to analyze the frequency response of the regulated cascode, we simplify the circuit to that shown in Fig. 9.37, where the auxiliary amplifier contains one pole at ω_0, $A_1(s) = A_0/(1+s/\omega_0)$, and only the load capacitance, C_L, is included. We wish to determine $V_{out}/V_{in} = -G_m Z_{out}$. To compute $G_m(s)$ (with the output node grounded), we note from Example 9.11 that the impedance seen looking into the source of M_2 is equal to $r_{O2}/[1+(A_1+1)g_{m2}r_{O2}]$, and divide the drain current of M_1 between this impedance and r_{O1}:

$$G_m(s) = g_{m1}\frac{r_{O1}}{r_{O1} + \dfrac{r_{O2}}{1+(A_1+1)g_{m2}r_{O2}}} \tag{9.34}$$

$$= \frac{g_{m1}r_{O1}[1+(A_1+1)g_{m2}r_{O2}]}{r_{O1}+(A_1+1)g_{m2}r_{O2}r_{O1}+r_{O2}} \tag{9.35}$$

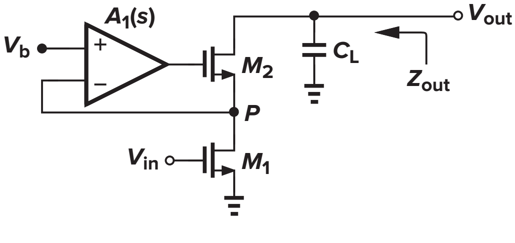


Figure 9.37 Circuit for analysis of frequency response.

Now, we calculate $Z_{out}(s)$ as the parallel combination of C_L and the impedance seen looking into the drain of M_2. From Eq. (9.21), we have

$$Z_{out} = [r_{O1} + (A_1 + 1)g_{m2}r_{O2}r_{O1} + r_{O2}]||\frac{1}{C_L s} \tag{9.36}$$

It follows that

$$\frac{V_{out}}{V_{in}}(s) = -G_m(s)Z_{out}(s) \tag{9.37}$$

$$= \frac{-g_{m1}r_{O1}[1 + (A_1 + 1)g_{m2}r_{O2}]}{(r_{O1} + r_{O2})C_L s + (A_1 + 1)g_{m2}r_{O2}r_{O1}C_L s + 1} \tag{9.38}$$

While it is tempting to assume that $A_1 \gg 1$ and hence neglect some terms, we must bear in mind that A_1 falls at high frequencies. Replacing A_1 with $A_0/(1 + s/\omega_0)$ yields

$$\frac{V_{out}}{V_{in}}(s) = \frac{-g_{m1}r_{O1}[(1 + g_{m2}r_{O2})\frac{s}{\omega_0} + (A_0 + 1)g_{m2}r_{O2} + 1]}{\frac{(r_{O1} + r_{O2})C_L}{\omega_0}[1 + g_{m2}(r_{O2}||r_{O1})]s^2 + [(r_{O1} + r_{O2})C_L + (A_0 + 1)g_{m2}r_{O2}r_{O1}C_L + \frac{1}{\omega_0}]s + 1} \tag{9.39}$$

It is interesting to note that, if we had assumed A_1 to be large for G_m and Z_{out} calculations, we would have obtained a *first-order* transfer function. The circuit exhibits a zero in the left half plane given by

$$|\omega_z| \approx (A_0 + 1)\omega_0 \tag{9.40}$$

if $g_{m2}r_{O2} \gg 1$. Produced by the path through A_1, this zero is on the order of the unity-gain bandwidth of the auxiliary amplifier.

To estimate pole frequencies, we assume that one is much greater than the other and apply the dominant-pole approximation (Chapter 6). The dominant pole is given by the inverse of the coefficient of s in the denominator of (9.39):

$$|\omega_{p1}| = \frac{1}{[r_{O1} + (A_0 + 1)g_{m2}r_{O2}r_{O1} + r_{O2}]C_L + \frac{1}{\omega_0}} \tag{9.41}$$

$$\approx \frac{1}{A_0 g_{m2}r_{O2}r_{O1}C_L} \tag{9.42}$$

The first time constant in the denominator of (9.41) corresponds to the output pole if A_1 were ideal, i.e., if $\omega_0 = \infty$. The nondominant pole is equal to the ratio of the coefficients of s and s^2:

$$|\omega_{p2}| = \frac{[r_{O1} + (A_0 + 1)g_{m2}r_{O2}r_{O1} + r_{O2}]C_L + \frac{1}{\omega_0}}{\frac{(r_{O1} + r_{O2})C_L}{\omega_0}[1 + g_{m2}(r_{O1}||r_{O2})]} \tag{9.43}$$

$$\approx (A_0 + 1)\omega_0 + \frac{1}{g_{m2}r_{O2}r_{O1}C_L} \tag{9.44}$$

if $g_{m2}(r_{O1}||r_{O2}) \gg 1$ (not necessarily a good approximation, but just to see trends). We observe that the second pole is somewhat *above* the unity-gain bandwidth of the original cascode, $(g_{m2}r_{O2}r_{O1}C_L)^{-1}$. Note that the term $1/(g_{m2}r_{O2}r_{O1}C_L)$ also represents the output pole in the absence of A_1.

▶ **Example 9.15**

Is the dominant-pole approximation valid here?

Solution

Assuming $(A_0 + 1)g_{m2}r_{O2}r_{O1} \gg r_{O1}, r_{O2}$, we find the ratio of (9.44) and (9.41):

$$\frac{\omega_{p2}}{\omega_{p1}} \approx \left[(A_0+1)\omega_0 + \frac{1}{g_{m2}r_{O2}r_{O1}C_L}\right]\left[(A_0+1)g_{m2}r_{O2}r_{O1}C_L + \frac{1}{\omega_0}\right] \tag{9.45}$$

$$\approx (A_0+1)^2 g_{m2}r_{O2}r_{O1}C_L\omega_0 + 2(A_0+1) + \frac{1}{g_{m2}r_{O2}r_{O1}C_L\omega_0} \tag{9.46}$$

The second term is typically much greater than unity, making the approximation valid. ◀

Figure 9.38 plots the approximate frequency response of the cascode structure before and after gain boosting. The key point here is that the auxiliary amplifier contributes a second pole located above the original −3-dB bandwidth by an amount equal to $A_0\omega_0$.

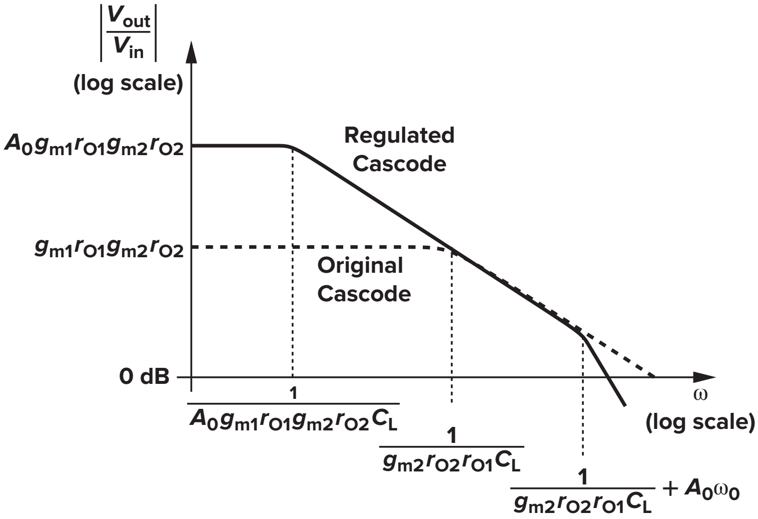


Figure 9.38 Frequency response of gain-boosted stage.

9.5 ■ Comparison

Our study of op amps in this chapter has introduced four principal topologies: telescopic cascode, folded cascode, two-stage op amp, and gain boosting. It is instructive to compare various performance aspects of these circuits to gain a better view of their applicability. Table 9.1 comparatively presents important attributes of each op amp topology. We study the speed differences in Chapter 10.

9.6 ■ Output Swing Calculations

In today's low-voltage op amp designs, the output voltage swing proves the most important factor. We have seen in previous sections how to assume a certain required output swing and accordingly allocate overdrive voltages to the transistors. But how do we verify that the final design indeed accommodates the

Table 9.1 Comparison of performance of various op amp topologies.

	Gain	Output Swing	Speed	Power Dissipation	Noise
Telescopic	Medium	Medium	Highest	Low	Low
Folded–Cascode	Medium	Medium	High	Medium	Medium
Two–Stage	High	Highest	Low	Medium	Low
Gain–Boosted	High	Medium	Medium	High	Medium

specified swing? To answer this question, we must first ask, what exactly happens if the circuit cannot sustain the swing? Since the border between the saturation and triode regions begins to diminish in nanometer devices, we cannot readily decide on the operation region of the transistors at the extremes of the output swing. A more rigorous approach is therefore necessary.

If the output voltage excursion pushes a transistor into the triode region, then the voltage gain drops. We can thus use simulations to examine the gain as the output swing grows. Illustrated in Fig. 9.39(a), the idea is to apply to the input a growing sinusoid (or different sinusoidal amplitudes in different simulations), monitor the resulting output, and calculate $|V_{out}/V_{in}|$ as V_{in} and V_{out} grow. The gain begins to drop as the output swing reaches its maximum "allowable" voltage, V_1. We may even choose V_1 to allow a small drop in the gain, say, 10% (about 1 dB). Beyond V_1, the gain falls further, causing significant nonlinearity.

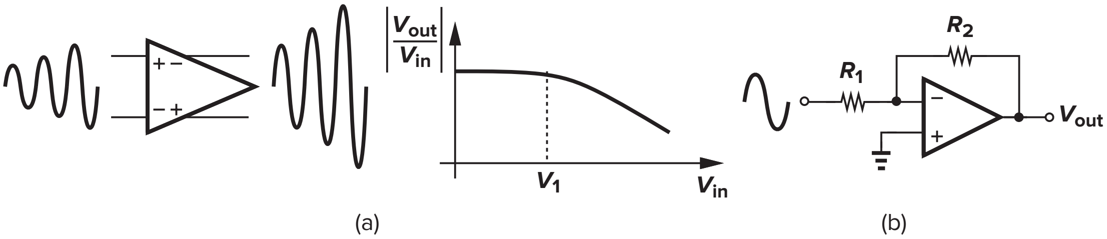


Figure 9.39 (a) Simulation of gain versus input amplitude, and (b) feedback amplifier.

The reader may wonder how much gain reduction is acceptable. In some applications, the reduction of the open-loop gain, and hence the gain error of the closed-loop system, are critical (Chapter 13). In other applications, we are concerned with the output distortion of the *closed-loop* circuit. In such a case, we place the op amp in the closed-loop environment of interest, e.g., the inverting configuration of Fig. 9.39(b), apply a sinusoid to the input, and measure the distortion (harmonics) at the output in simulations. The maximum output amplitude that yields an acceptable distortion is considered the maximum output swing.

9.7 ■ Common-Mode Feedback

9.7.1 Basic Concepts

In this and previous chapters, we have described many advantages of fully differential circuits over their single-ended counterparts. In addition to greater output swings, differential op amps avoid mirror poles, thus achieving a higher closed-loop speed. However, high-gain differential circuits require "common-mode feedback" (CMFB).

To understand the need for CMFB, let us begin with a simple realization of a differential amplifier [Fig. 9.40(a)]. In some applications, we short the inputs and outputs for part of the operation [Fig. 9.40(b)],

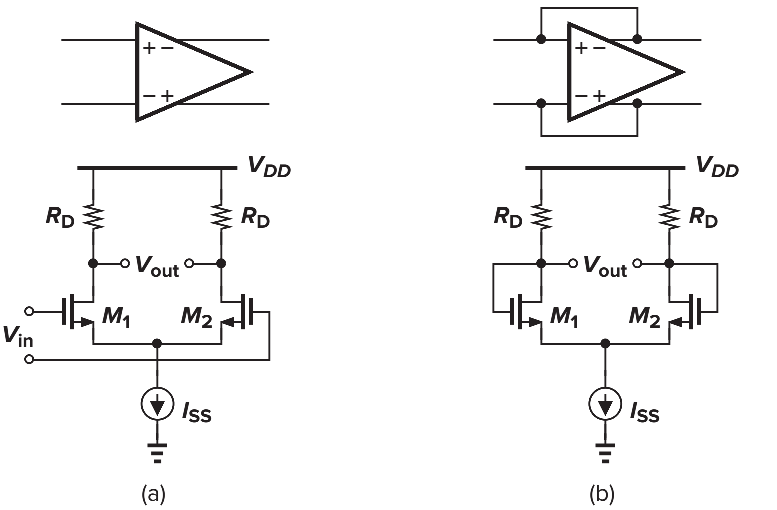


Figure 9.40 (a) Simple differential pair; (b) circuit with inputs shorted to outputs.

providing *differential* negative feedback. The input and output common-mode levels in this case are fairly well defined, equal to $V_{DD} - I_{SS}R_D/2$.

Now suppose the load resistors are replaced by PMOS current sources so as to increase the differential voltage gain [Fig. 9.41(a)]. What is the common-mode level at nodes X and Y? Since each of the input transistors carries a current of $I_{SS}/2$, the CM level depends on how close I_{D3} and I_{D4} are to this value. In practice, as exemplified by Fig. 9.41(b), mismatches in the PMOS and NMOS current mirrors defining I_{SS} and $I_{D3,4}$ create a finite error between $I_{D3,4}$ and $I_{SS}/2$. Suppose, for example, that the drain currents of M_3 and M_4 in the saturation region are slightly greater than $I_{SS}/2$. As a result, to satisfy Kirchhoff's current law at nodes X and Y, both M_3 and M_4 must enter the triode region so that their drain currents fall to $I_{SS}/2$. Conversely, if $I_{D3,4} < I_{SS}/2$, then both V_X and V_Y must drop so that M_5 enters the triode region, thereby producing only $2I_{D3,4}$.

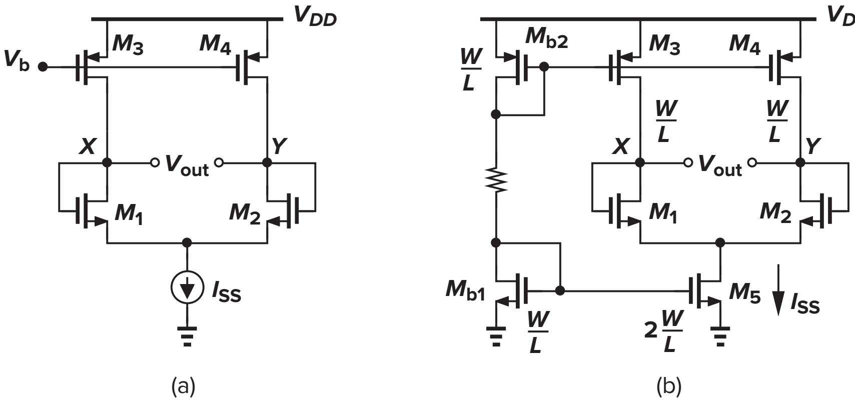


Figure 9.41 (a) High-gain differential pair with inputs shorted to outputs, and (b) effect of current mismatches.

The above difficulties fundamentally arise because in high-gain amplifiers, we wish a p-type current source [e.g., M_3 and M_4 in Fig. 9.41(b)] to balance an n-type current source (e.g., M_5). As illustrated in Fig. 9.42, the difference between I_P and I_N must flow through the intrinsic output impedance of the

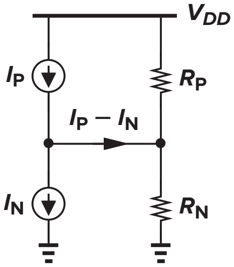


Figure 9.42 Simplified model of high-gain amplifier.

amplifier, creating an output voltage change of $(I_P - I_N)(R_P \| R_N)$. Since the current error depends on mismatches and $R_P \| R_N$ is quite high, the voltage error may be large, thus driving the p-type or n-type current source into the triode region. As a general rule, if the output CM level cannot be determined by "visual inspection" and requires calculations based on device properties, then it is poorly defined. This is the case in Fig. 9.41 but not in Fig. 9.40. We emphasize that differential feedback cannot define the CM level.

Students often make two mistakes here. First, they assume that differential feedback corrects the output common-mode level. As explained for the simple circuit of Fig. 9.41(a), differential feedback from X and Y to the inputs cannot prohibit the output CM level from taking off toward V_{DD} or ground. Second, they finely adjust V_b in simulations so as to bring V_X and V_Y to around $V_{DD}/2$ concluding that the circuit does not need CM feedback. We have recognized, however, that random mismatches between the top and bottom current sources cause the CM level to fall or rise considerably. Such mismatches are always present in actual circuits and cause the op amp to fail if no CMFB is used.

▶ Example 9.16

Consider the telescopic op amp designed in Example 9.5 and repeated in Fig. 9.43 with bias current mirrors. Suppose M_9 suffers from a 1% current mismatch with respect to M_{10}, producing $I_{SS} = 2.97$ mA rather than 3 mA. Assuming perfect matching for other transistors, explain what happens in the circuit.

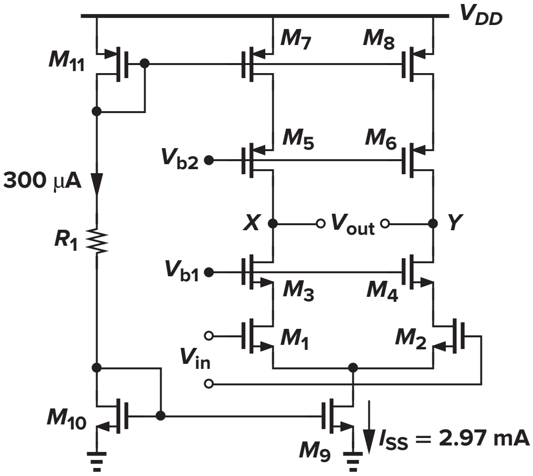


Figure 9.43

Solution

From Example 9.5, the single-ended output impedance of the circuit equals 266 kΩ. Since the difference between the drain currents of M_3 and M_5 (and M_4 and M_6) is 30 μA/2 = 15 μA, the output voltage error would be 266 kΩ × 15 μA= 3.99 V. Since this large error cannot be produced, V_X and V_Y must rise so much that M_5–M_6 and M_7–M_8 enter the triode region, yielding $I_{D7,8} = 1.485$ mA. We should also mention that another important source of CM error in the simple biasing scheme of Fig. 9.43 is the *deterministic* error between $I_{D7,8}$ and I_{D11} (and also between I_{D9} and I_{D10}) due to their different drain-source voltages. This error can nonetheless be reduced by means of the current mirror techniques of Chapter 5.

The foregoing study implies that in high-gain amplifiers, the output CM level is sensitive to device properties and mismatches and it cannot be stabilized by means of *differential* feedback. Thus, a common-mode feedback network must be added to sense the CM level of the two outputs and adjust one of the bias currents in the amplifier. Following our view of feedback systems in Chapter 8, we divide the task of CMFB into three operations: sensing the output CM level, comparison with a reference, and returning the error to the amplifier's bias network. Figure 9.44 conceptually illustrates the idea.

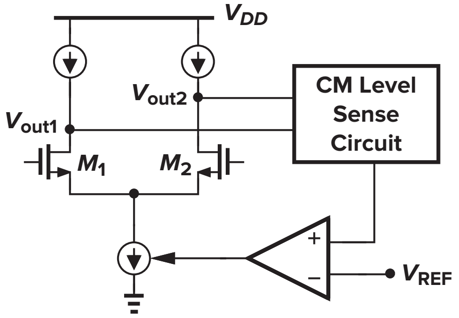


Figure 9.44 Conceptual topology for common-mode feedback.

9.7.2 CM Sensing Techniques

In order to sense the output CM level, we recall that $V_{out,CM} = (V_{out1} + V_{out2})/2$, where V_{out1} and V_{out2} are the single-ended outputs. It therefore seems plausible to employ a resistive divider as shown in Fig. 9.45, generating $V_{out,CM} = (R_1V_{out2} + R_2V_{out1})/(R_1 + R_2)$, which reduces to $(V_{out1} + V_{out2})/2$ if $R_1 = R_2$. The difficulty, however, is that R_1 and R_2 must be much greater than the output impedance of the op amp so as to avoid lowering the open-loop gain. For example, in the design of Fig. 9.43, the output impedance equals 266 kΩ, necessitating a value of several megaohms for R_1 and R_2. As explained in Chapter 18, such large resistors occupy a very large area and, more important, suffer from substantial parasitic capacitance to the substrate.

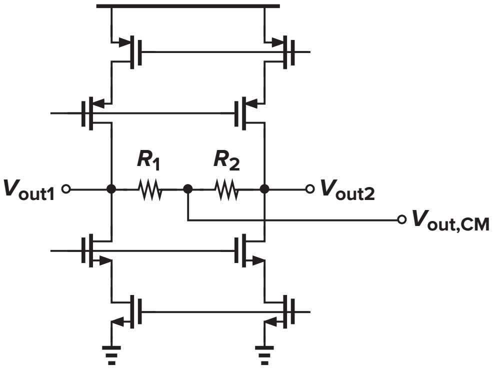


Figure 9.45 Common-mode feedback with resistive sensing.

To eliminate the resistive loading, we can interpose source followers between each output and its corresponding resistor. Illustrated in Fig. 9.46, this technique produces a CM level that is in fact lower than the output CM level by $V_{GS7,8}$, but this shift can be taken into account in the comparison operation. Note that R_1 and R_2 or I_1 and I_2 must be large enough to ensure that M_7 or M_8 is not "starved" when

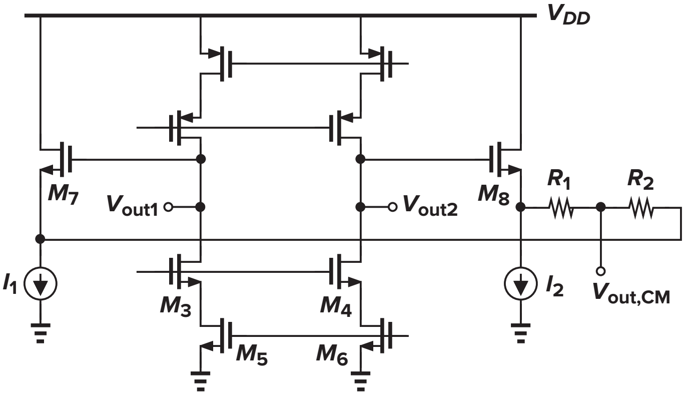


Figure 9.46 Common-mode feedback using source followers.

a large differential swing appears at the output. As conceptually depicted in Fig. 9.47, if, say, V_{out2} is quite higher than V_{out1}, then I_1 must sink both $I_X \approx (V_{out2} - V_{out1})/(R_1 + R_2)$ and I_{D7}. Consequently, if $R_1 + R_2$ or I_1 is not sufficiently large, I_{D7} drops to zero and $V_{out,CM}$ no longer represents the true output CM level.

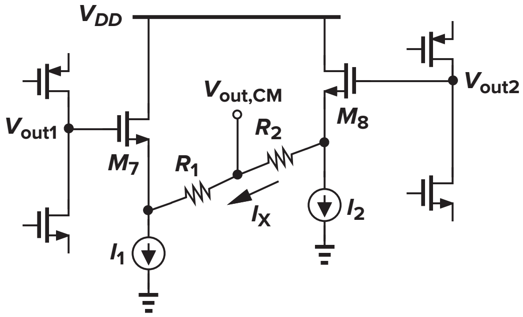


Figure 9.47 Current starvation of source followers for large swings.

The sensing method of Fig. 9.46 nevertheless suffers from an important drawback: it limits the differential output swings (even if $R_{1,2}$ and $I_{1,2}$ are large enough). To understand why, let us determine the minimum allowable level of V_{out1} (and V_{out2}), noting that without CMFB, it would be equal to $V_{OD3} + V_{OD5}$. With the source followers in place, $V_{out1,min} = V_{GS7} + V_{I1}$, where V_{I1} denotes the minimum voltage required across I_1. This is roughly equal to two overdrive voltages plus one threshold voltage. Thus, the swing at each output is reduced by approximately V_{TH}, a significant value in low-voltage design.

Looking at Fig. 9.45, the reader may wonder if the output CM level can be sensed by means of *capacitors*, rather than resistors, so as to avoid degrading the low-frequency open-loop gain of the op amp. This is indeed possible in some cases and will be studied in Chapter 13.

Another type of CM sensing is depicted in Fig. 9.48(a). Here, identical transistors M_7 and M_8 operate in the deep triode region, introducing a total resistance between P and ground equal to

$$R_{tot} = R_{on7} \| R_{on8} \tag{9.47}$$

$$= \frac{1}{\mu_n C_{ox}\frac{W}{L}(V_{out1} - V_{TH})} \Bigg\| \frac{1}{\mu_n C_{ox}\frac{W}{L}(V_{out2} - V_{TH})} \tag{9.48}$$

$$= \frac{1}{\mu_n C_{ox}\frac{W}{L}(V_{out2} + V_{out1} - 2V_{TH})} \tag{9.49}$$

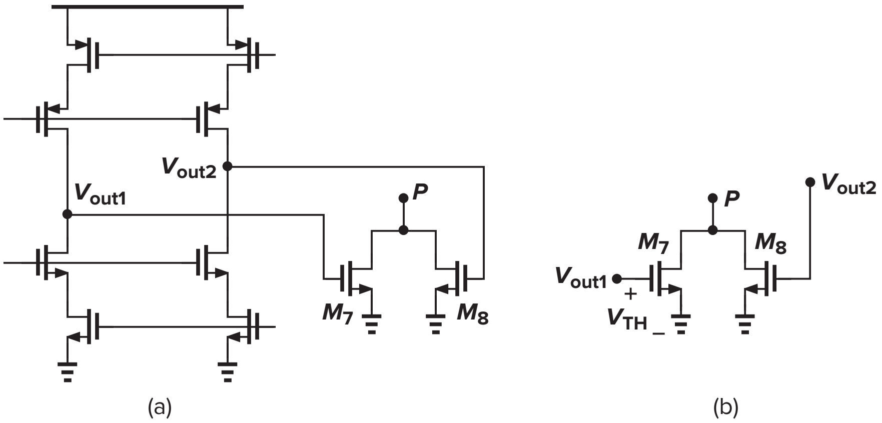


Figure 9.48 (a) Common-mode sensing using MOSFETs operating in the deep triode region, and (b) output levels placing M_7 at the edge of saturation.

where W/L denotes the aspect ratio of M_7 and M_8. Equation (9.49) indicates that R_{tot} is a function of $V_{out2} + V_{out1}$ but independent of $V_{out2} - V_{out1}$. From Fig. 9.48(a), we observe that if the outputs rise together, then R_{tot} drops, whereas if they change differentially, one R_{on} increases and the other decreases. This resistance can thus be utilized as a measure of the output CM level.

In the circuit of Fig. 9.48(a), the use of M_7 and M_8 limits the output voltage swings. Here, it may seem that $V_{out,min} = V_{TH7,8}$, which is relatively close to two overdrive voltages, but the difficulty arises from the assumption above that both M_7 and M_8 operate in the deep triode region. In fact, if, say, V_{out1} drops from the equilibrium CM level to about one threshold voltage above ground [Fig. 9.48(b)] and V_{out2} rises by the same amount, then M_7 enters the saturation region, thus exhibiting a variation in its on-resistance that is not counterbalanced by that of M_8.

It is important to bear in mind that CM sensing must produce a quantity *independent* of the differential signals. The following example illustrates this point.

▶ Example 9.17

A student simulates the step response of a closed-loop op amp circuit [e.g., that in Fig. 9.48(a)] and observes the output waveforms shown in Fig. 9.49. Explain why V_{out1} and V_{out2} do not change symmetrically.

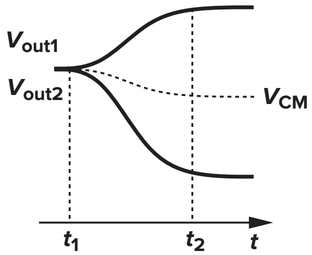


Figure 9.49

Solution

As evident from the waveforms, the output CM level *changes* from t_1 to t_2, indicating that the CM sensing mechanism is nonlinear and interprets the CM levels at t_1 and t_2 differently. For example, if M_7 or M_8 in Fig. 9.48 does not remain in the deep triode region at t_2, then Eq. (9.49) no longer holds and V_{CM} becomes a function of the *differential* signals.

◀

Another CM sensing method is illustrated in Fig. 9.50. Here, the differential pairs compare the inputs with V_{REF}, generating a current, I_{CM}, in proportion to the input CM level. To prove this point, we write the small-signal drain currents of M_2 and M_4 as $(g_m/2)V_{out1}$ and $(g_m/2)V_{out2}$, respectively, concluding that $I_{CM} \propto V_{out1} + V_{out2}$. This current can be copied to current sources within the op amp with negative feedback so as to keep $V_{out,CM}$ approximately equal to V_{REF}.

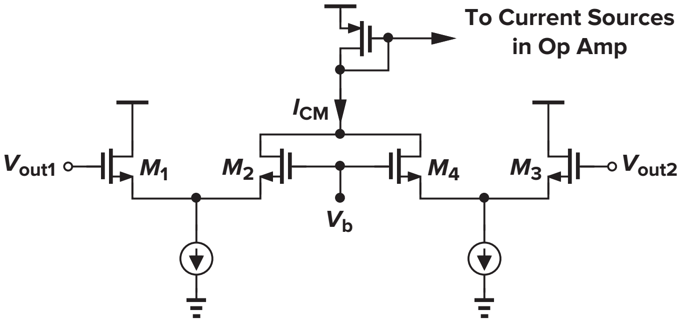


Figure 9.50 CM sensing circuit with high nonlinearity.

The foregoing topology faces serious issues. As V_{out1} and V_{out2} experience large swings, I_{out} no longer remains proportional to $V_{out1} + V_{out2}$ due to the nonlinearity of the differential pairs. In fact, if I_{D1} and I_{D2} are expressed as $f(V_{out1} - V_{REF})$ and $f(V_{out2} - V_{REF})$, respectively, we observe that $I_{D1} + I_{D2}$ depends on the individual values of V_{out1} and V_{out2} unless $f()$ is a linear function. As a result, the reconstructed CM level does not remain constant in the presence of large differential output swings.

9.7.3 CM Feedback Techniques

We now study techniques of comparing the measured CM level with a reference and returning the error to the op amp's bias network. In the circuit of Fig. 9.51, we employ a simple amplifier to detect the difference between $V_{out,CM}$ and a reference voltage, V_{REF}, applying the result to the NMOS current sources with negative feedback. If both V_{out1} and V_{out2} rise, so does V_E, thereby increasing the drain currents of M_3–M_4 and lowering the output CM level. In other words, if the loop gain is large, the feedback network forces the CM level of V_{out1} and V_{out2} to approach V_{REF}. Note that the feedback can be applied to the PMOS current sources as well. Also, the feedback may control only a fraction of the current to allow optimization of

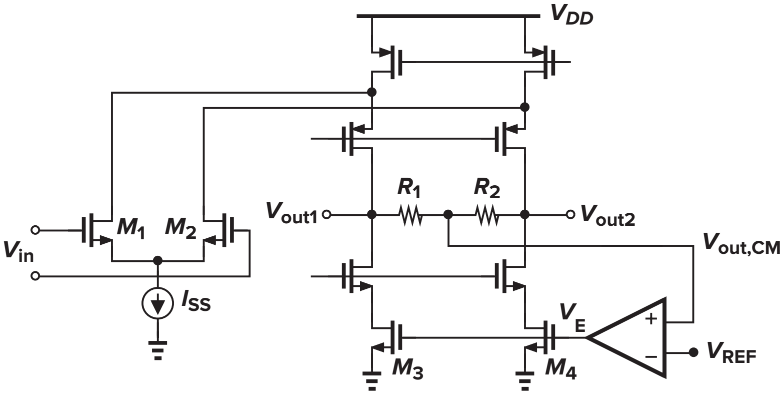


Figure 9.51 Sensing and controlling output CM level.

the settling behavior. For example, each of M_3 and M_4 can be decomposed into two parallel devices, one biased at a constant current and the other driven by the error amplifier.

In a folded-cascode op amp, the CM feedback may control the tail current of the input differential pair. Illustrated in Fig. 9.52, this method increases the tail current if V_{out1} and V_{out2} rise, lowering the drain currents of M_5–M_6 and restoring the output CM level.

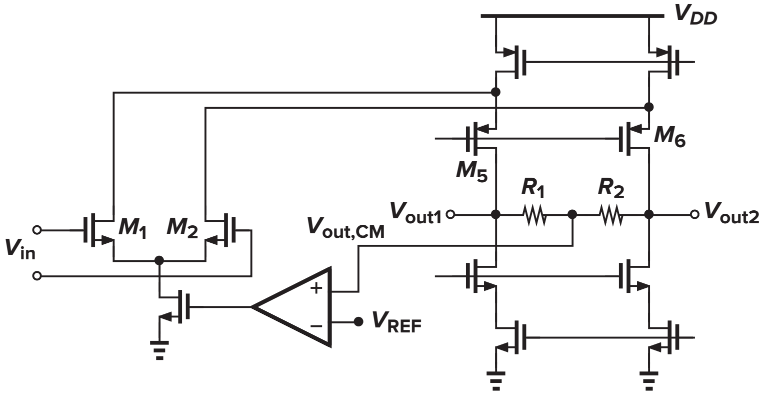


Figure 9.52 Alternative method of controlling output CM level.

How do we perform comparison and feedback with the sensing scheme of Fig. 9.48? Here, the output CM voltage is directly converted to a resistance or a current, prohibiting comparison with a reference voltage. A simple feedback topology utilizing this technique is depicted in Fig. 9.53, where $R_{on7}\|R_{on8}$ adjusts the bias current of M_5 and M_6. The output CM level sets $R_{on7}\|R_{on8}$ such that I_{D5} and I_{D6} exactly balance I_{D9} and I_{D10}, respectively. For example, if V_{out1} and V_{out2} rise, $R_{on7}||R_{on8}$ falls and the drain currents of M_5 and M_6 increase, pulling V_{out1} and V_{out2} down. Assuming that $I_{D9} = I_{D10} = I_D$, we must have $V_b - V_{GS5} = 2I_D(R_{on7}\|R_{on8})$, and hence $R_{on7}\|R_{on8} = (V_b - V_{GS5})/(2I_D)$. From (9.49),

$$\frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{7,8}(V_{out2} + V_{out1} - 2V_{TH})} = \frac{V_b - V_{GS5}}{2I_D} \tag{9.50}$$

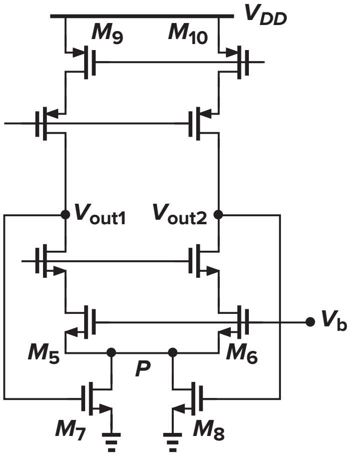


Figure 9.53 CMFB using triode devices.

that is,

$$V_{out1} + V_{out2} = \frac{2I_D}{\mu_n C_{ox}\left(\dfrac{W}{L}\right)_{7,8}} \frac{1}{V_b - V_{GS5}} + 2V_{TH} \tag{9.51}$$

The CM level can thus be obtained by noting that $V_{GS5} = \sqrt{2I_D/[\mu_n C_{ox}(W/L)_5]} + V_{TH5}$.

The CMFB network of Fig. 9.53 suffers from several drawbacks. First, the value of the output CM level is a function of device parameters. Second, the voltage drop across $R_{on7} \| R_{on8}$ limits the output voltage swings. Third, to minimize this drop, M_7 and M_8 are usually quite wide devices, introducing substantial capacitance at the output. The second issue can be alleviated by applying the feedback to the tail current of the input differential pair (Fig. 9.54), but the other two remain.

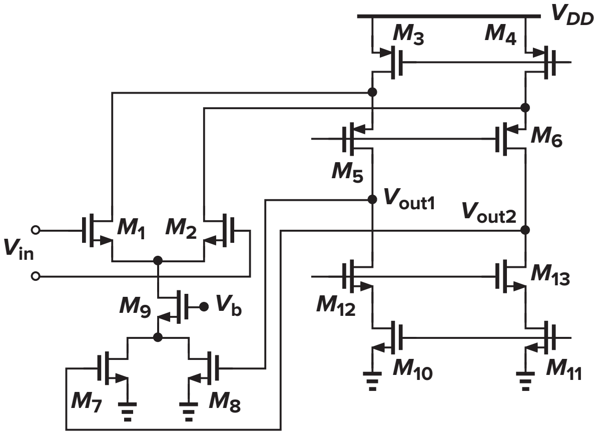


Figure 9.54 Alternative method of controlling output CM level.

How is V_b generated in Fig. 9.54? We note that $V_{out,CM}$ is somewhat sensitive to the value of V_b: if V_b is higher than expected, the tail current of M_1 and M_2 increases and the output CM level falls. Since the feedback through M_7 and M_8 attempts to correct this error, the overall change in $V_{out,CM}$ depends on the loop gain in the CMFB network. This is studied in the following example.

▶ **Example 9.18**

For the circuit of Fig. 9.54, determine the sensitivity of $V_{out,CM}$ to V_b, i.e., $dV_{out,CM}/dV_b$.

Solution

Setting V_{in} to zero and opening the loop at the gates of M_7 and M_8, we simplify the circuit as shown in Fig. 9.55. Note that g_{m7} and g_{m8} must be calculated in the triode region: $g_{m7} = g_{m8} = \mu_n C_{ox}(W/L)_{7,8} V_{DS7,8}$, where $V_{DS7,8}$ denotes the bias value of the drain-source voltage of M_7 and M_8. Since M_7 and M_8 operate in the deep triode region, $V_{DS7,8}$ is typically less than 100 mV.

In a well-designed circuit, the loop gain must be relatively high. We therefore surmise that the closed-loop gain is approximately equal to $1/\beta$, where β represents the feedback factor. We write from Chapter 8:

$$\beta = \frac{V_2}{V_1}|_{I2=0} \tag{9.52}$$

$$= -(g_{m7} + g_{m8})(R_{on7} \| R_{on8}) \tag{9.53}$$

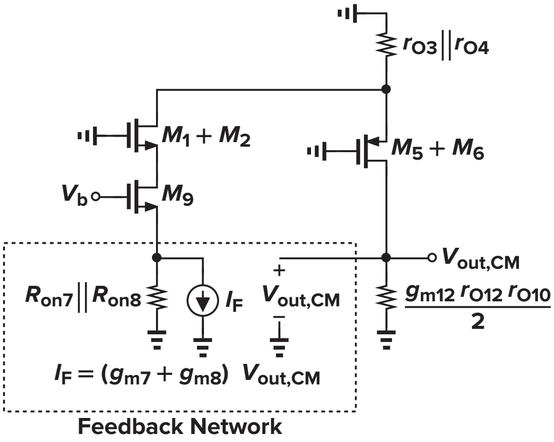


Figure 9.55

$$= -2\mu_n C_{ox}\left(\frac{W}{L}\right)_{7,8} V_{DS7,8} \cdot \frac{1}{2\mu_n C_{ox}(W/L)_{7,8}(V_{GS7,8} - V_{TH7,8})} \tag{9.54}$$

$$= -\frac{V_{DS7,8}}{V_{GS7,8} - V_{TH7,8}} \tag{9.55}$$

where $V_{GS7,8} - V_{TH7,8}$ denotes the overdrive voltage of M_7 and M_8. Thus,

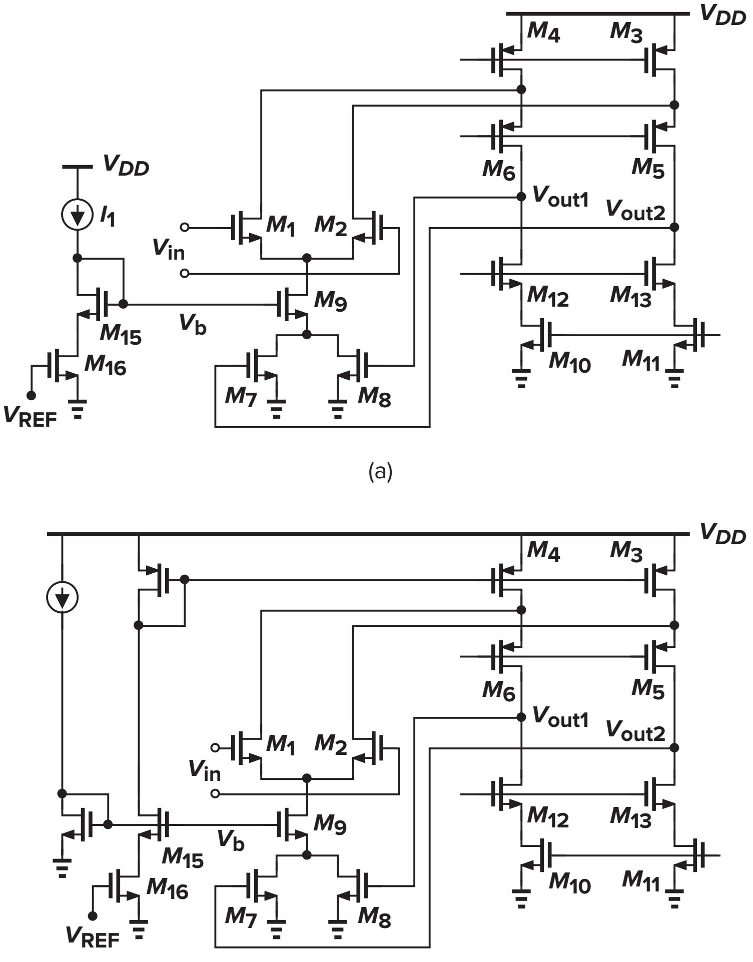
$$\left|\frac{dV_{out,CM}}{dV_b}\right|_{closed} \approx \frac{V_{GS7,8} - V_{TH7,8}}{V_{DS7,8}} \tag{9.56}$$

This is an important result. Since $V_{GS7,8}$ (i.e., the output CM level) is typically in the vicinity of $V_{DD}/2$, the above equation suggests that $V_{DS7,8}$ must be maximized to minimize this sensitivity, but at the cost of the loop gain.

We now introduce a modification to the circuit of Fig. 9.54 that both makes the output level relatively independent of device parameters and lowers the sensitivity to the value of V_b. Illustrated in Fig. 9.56(a), the idea is to define V_b by a current mirror arrangement such that I_{D9} "tracks" I_1 and V_{REF}. For simplicity, suppose $(W/L)_{15} = (W/L)_9$ and $(W/L)_{16} = (W/L)_7 + (W/L)_8$. Thus, $I_{D9} = I_1$ only if $V_{out,CM} = V_{REF}$. In other words, as with Fig. 9.52, the circuit produces an output CM level equal to a reference but it requires no resistors in sensing $V_{out,CM}$. The overall design can be simplified as shown in Fig. 9.56(b).

In practice, since $V_{DS15} \neq V_{DS9}$, channel-length modulation results in a finite error. Figure 9.57 depicts a modification that suppresses this error. Here, transistors M_{17} and M_{18} reproduce at the drain of M_{15} a voltage equal to the source voltage of M_1 and M_2, ensuring that $V_{DS15} = V_{DS9}$.

To arrive at another CM feedback topology, let us consider the simple differential pair shown in Fig. 9.58(a). Here, the output CM level, $V_{DD} - |V_{GS3,4}|$, is relatively well defined, but the voltage gain is quite low. To increase the differential gain, the PMOS devices must operate as current sources for *differential* signals. We therefore modify the circuit as depicted in Fig. 9.58(b), where for differential changes at V_{out1} and V_{out2}, node P is a virtual ground and the gain can be expressed as $g_{m1,2}(r_{O1,2}\|r_{O3,4}\|R_F)$. We preferably choose $R_F \gg r_{O1,2}||r_{O3,4}$. For common-mode levels, on the other hand, M_3 and M_4 operate as diode-connected devices. The circuit proves useful in low-gain applications.



(a)

(b)

Figure 9.56 Modification of CMFB for more accurate definition of output CM level.

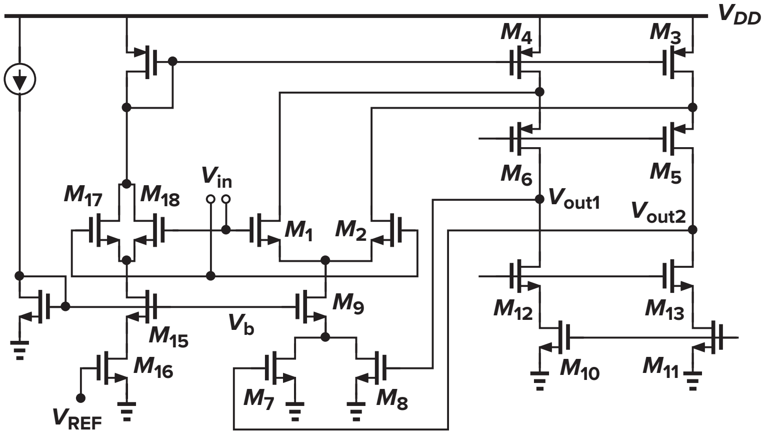


Figure 9.57 Modification to suppress error due to channel-length modulation.

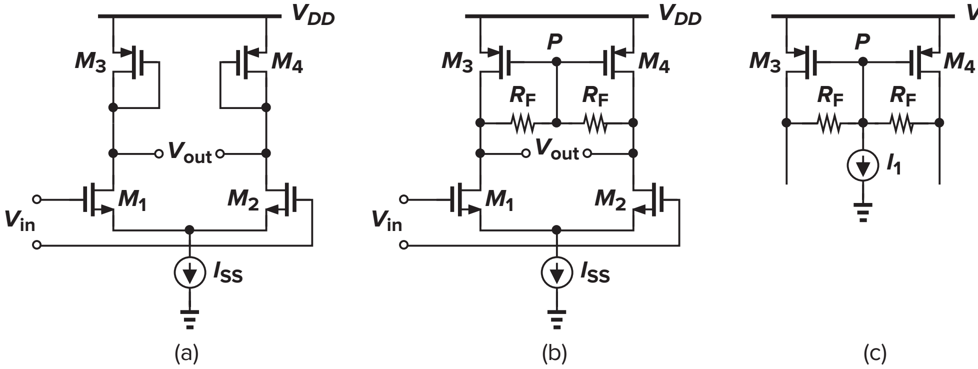


Figure 9.58 (a) Differential pair using diode-connected loads, (b) resistive CMFB, and (c) modification to allow low-voltage operation.

▶ Example 9.19

Determine the maximum allowable output swings in Fig. 9.58(b).

Solution

Each output can fall to two overdrive voltages above ground if $V_{in,CM}$ is chosen to place I_{SS} at the edge of the triode region. The highest level allowed at the output is equal to the output CM level plus $|V_{TH3,4}|$, i.e., $V_{DD} - |V_{GS3,4}| + |V_{TH3,4}| = V_{DD} - |V_{GS3,4} - V_{TH3,4}|$.

◀

In some applications, we wish to operate the circuit of Fig. 9.58(b) with a low supply voltage, but for small signals. This stage dictates a minimum V_{DD} of $|V_{GS3,4}|$ plus two overdrive voltages. We modify the circuit by drawing a small current from the two resistors and PMOS devices as illustrated in Fig. 9.58(c). Here, V_P is still equal to $V_{DD} - |V_{GS3,4}|$, but the drain voltages are *higher* than V_P by an amount equal to $I_1R_F/2$. For example, if $I_1R_F/2 = |V_{TH3,4}|$, then the PMOS devices operate at the edge of saturation, allowing a minimum V_{DD} of three overdrive voltages.

▶ Example 9.20

Facing voltage headroom limitations, a student constructs the circuit shown in Fig. 9.59(a), where the tail current source is replaced by two triode devices that sense the output CM level, $V_{out,CM}$. Determine the small-signal gain from the input CM level to the output CM level.

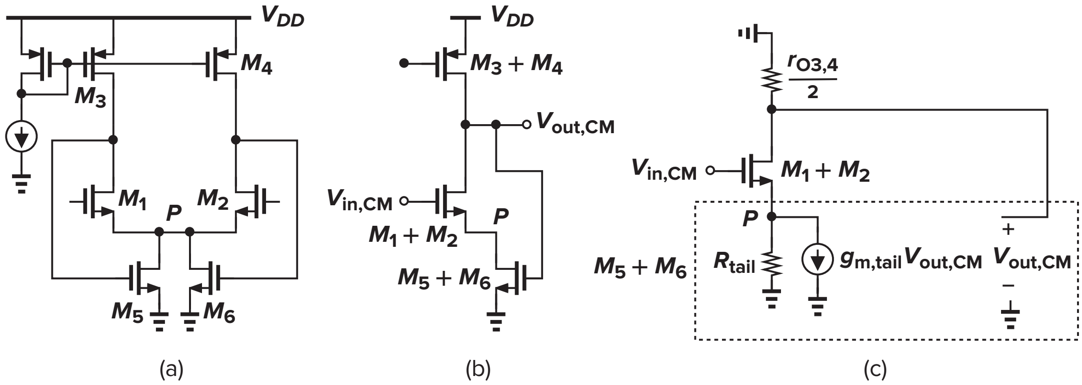


Figure 9.59

Solution

If the circuit is symmetric, the output nodes can be shorted, leading to the topology in Fig. 9.59(b).[6] To model the composite transistor $M_5 + M_6$, we define a transconductance $g_{m,tail} = g_{m5} + g_{m6} = 2\mu_n C_{ox}(W/L)_{5,6}V_P$, where V_P is the dc voltage at node P. We also approximate their total channel resistance by $R_{tail} = [2\mu_n C_{ox}(W/L)_{5,6}(V_{out,CM} - V_{TH5,6})]^{-1}$. The circuit therefore reduces to that shown in Fig. 9.59(c).

Assuming for simplicity that $\lambda = \gamma = 0$ for M_1 and M_2, we express the small-signal current drawn by $M_1 + M_2$ as $-V_{out,CM}/(r_{O3,4}/2)$. This current translates to a gate-source voltage of $-V_{out}/(2g_{m1,2}r_{O3,4}/2) = -V_{out}/(g_{m1,2}r_{O3,4})$, yielding a voltage of $V_{in,CM} + V_{out}/(g_{m1,2}r_{O3,4})$ at node P and hence a current of $[V_{in,CM} + V_{out}/(g_{m1,2}r_{O3,4})]/R_{tail}$ through R_{tail}. Since this current and $g_{m,tail}V_{out,CM}$ must add up to $-V_{out,CM}/(r_{O3,4}/2)$, we obtain

$$\frac{V_{out,CM}}{V_{in,CM}} = -\frac{1}{\dfrac{2R_{tail}}{r_{O3,4}} + g_{m,tail}R_{tail} + (g_{m1,2}r_{O3,4})^{-1}} \tag{9.57}$$

It is important to note that all of the three terms in the denominator are less than one (why?), revealing that $V_{out,CM}/V_{in,CM}$ is roughly around unity. That is, an error in the input CM level reaches the output without significant attenuation. This observation suggests a poor CMRR; the reader is encouraged to assume a g_m mismatch between M_1 and M_2 and compute the CMRR as outlined in Chapter 4.

9.7.4 CMFB in Two-Stage Op Amps

Offering nearly rail-to-rail output swings, two-stage op amps find wider application than other topologies in today's designs. However, such op amps require more complex common-mode feedback. To understand the issues, we consider three different CMFB methods in the context of the simple circuit shown in Fig. 9.60(a).

First, suppose the CM level of V_{out1} and V_{out2} is sensed and the result is used to control only V_{b2}; i.e., the second stage incorporates CMFB, but not the first stage [Fig. 9.60(b)]. In this case, no mechanism exists that controls the CM level at X and Y. For example, if I_{SS} happens to be less than the sum of the currents that M_3 and M_4 wish to draw, then V_X and V_Y rise, driving these transistors into the triode region so that $I_{D3} + I_{D4}$ eventually becomes equal to I_{SS}. This effect also reduces $|V_{GS5,6}|$, establishing in M_5–M_8 a current that may be well below the nominal value. This CMFB method is therefore not desired.

Second, we still sense the CM level V_{out1} and V_{out2} but return the result to the first stage, e.g., to I_{SS} [Fig. 9.60(c)]. Suppose, for example, that V_{out1} and V_{out2} begin too high. Then, the error amplifier, A_e, reduces I_{SS}, allowing V_X and V_Y to rise, $|I_{D5}|$ and $|I_{D6}|$ to fall, and V_{out1} and V_{out2} to go down. It is interesting to note that here M_5 and M_6 in fact sense the CM level at X and Y, helping the global loop control both stages' CM level. (If M_3 and M_4 had a tail current, as in a regular differential pair, this property would vanish and the CMFB loop would fail.)

While used in some designs, the second technique suffers from a critical drawback. Let us draw the equivalent circuit for common-mode levels (Fig. 9.61). How many poles does the CM feedback loop contain? We count one pole at X or Y, one at the main output, and at least one associated with the error amplifier. Moreover, since R_{CM} is so large as not to load the second stage, it forms with the input capacitance of A_e a pole that may not be negligible. Thus, even if the pole at the source of M_1 and M_2 is discounted, the CMFB loop still contains three or four poles. As explained in Chapter 10, this many poles make it difficult for the loop be stable.

In order to avoid stability issues, we can employ two separate CMFB loops for the first and second stages of the op amp. Figure 9.62 illustrates a simple example [7], where, in a manner similar to Fig. 9.58(b),

[6] We use the notation $M_j + M_{j+1}$ to denote the parallel combination of M_j and M_{j+1}.

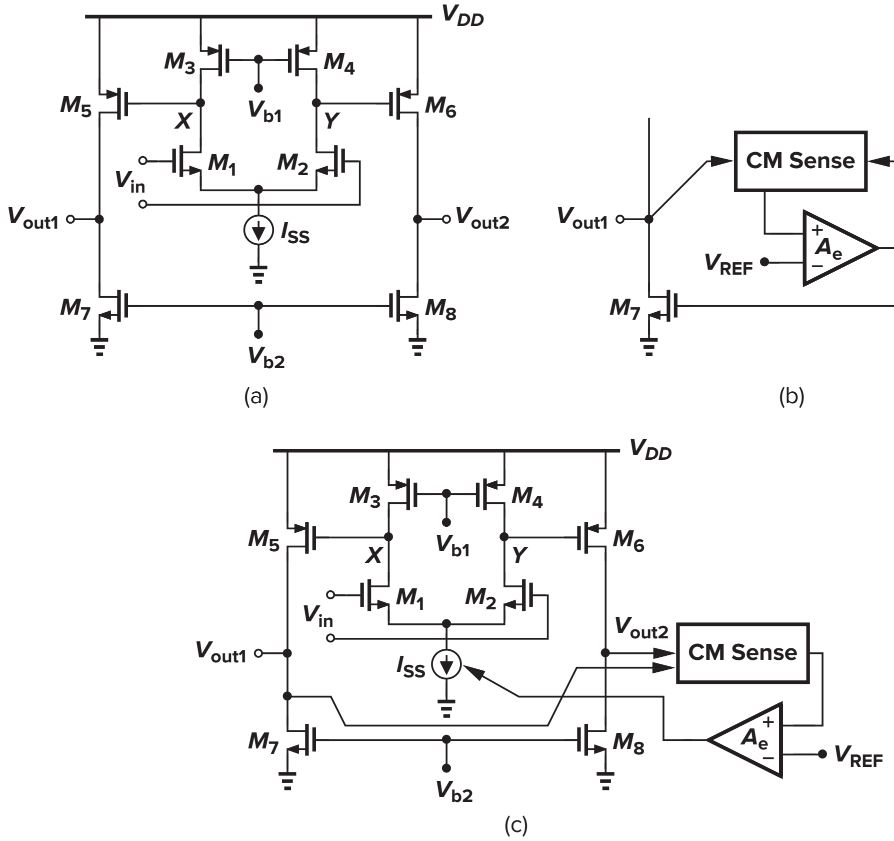


Figure 9.60 (a) Two-stage op amp, (b) CMFB around second stage, and (c) CMFB from second stage to first stage.

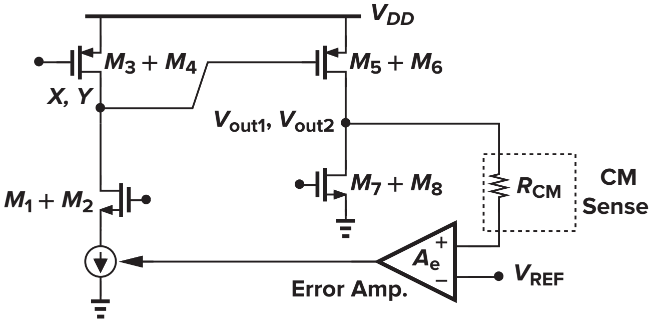


Figure 9.61 Equivalent CMFB loop to determine the number of poles.

R_1 and R_2 provide CMFB for the first stage and R_3 and R_4 for the second. Interestingly, all of the drain currents in this topology are copied from I_{SS}. Assuming a symmetric circuit, we recognize that (1) resistors R_1 and R_2 adjust $V_{GS3,4}$ until $|I_{D3}| = |I_{D4}| = I_{SS}/2$; (2) since $V_{GS3,4} = V_{GS5,6}$, M_5 and M_6 copy their currents from M_3 and M_4 as in a current mirror; and (3) resistors R_3 and R_4 adjust $V_{GS7,8}$ until $I_{D7} = I_{D8} = |I_{D5}| = |I_{D6}|$. The differential voltage gain is equal to $g_{m1}(r_{O1}||r_{O3}||R_1)g_{m5}(r_{O5}||r_{O7}||R_3)$.

Another CMFB technique for two-stage op amps is described in Chapter 11.

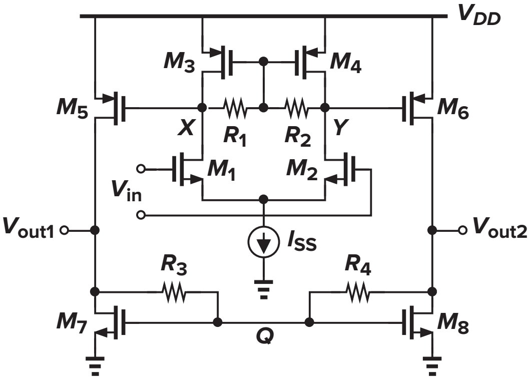


Figure 9.62 Simple CMFB loops around each stage.

▶ **Example 9.21**

A student delighted by the simplicity of the op amp in Fig. 9.62 designs the circuit for a given power budget, but realizes that the output CM level is inevitably well below $V_{DD}/2$, and hence the output swings are limited. Explain why and devise a solution.

Solution

The output CM level is equal to $V_{G7,8}$ (recall that R_3 and R_4 carry no current in the absence of signals). Since M_7 and M_8 are chosen wide enough for a small overdrive voltage, $V_{GS7,8}$ is only slightly greater than one threshold voltage and far from $V_{DD}/2$.

This issue can be resolved by drawing a small current from node Q (Fig. 9.63). Now, R_3 and R_4 sustain a drop of $R_3I_Q/2$ $(= R_4I_Q/2)$, producing an upward shift of the same amount in the output CM level [7]. Thus, I_Q can be chosen to create an output CM level around $V_{DD}/2$.

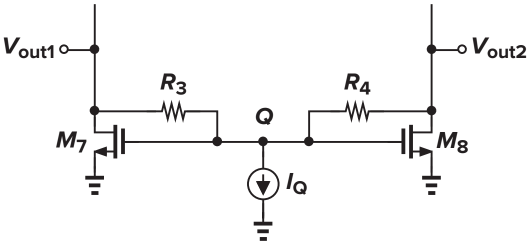


Figure 9.63

◀

If the first stage incorporates a telescopic cascode to achieve a high gain, then the CMFB loops can be realized as shown in Fig. 9.64. While not precise, the CM sensing of X and Y avoids loading the high impedances at these nodes, thereby maintaining a high voltage gain.

9.8 ■ Input Range Limitations

The op amp circuits studied thus far have evolved to achieve large differential output swings. While the differential input swings are usually much smaller (by a factor equal to the open-loop gain), the input *common-mode* level may need to vary over a wide range in some applications. For example, consider the simple unity-gain buffer shown in Fig. 9.65, where the input swing is nearly equal to the output swing. Interestingly, in this case the voltage swings are limited by the input differential pair rather than the output

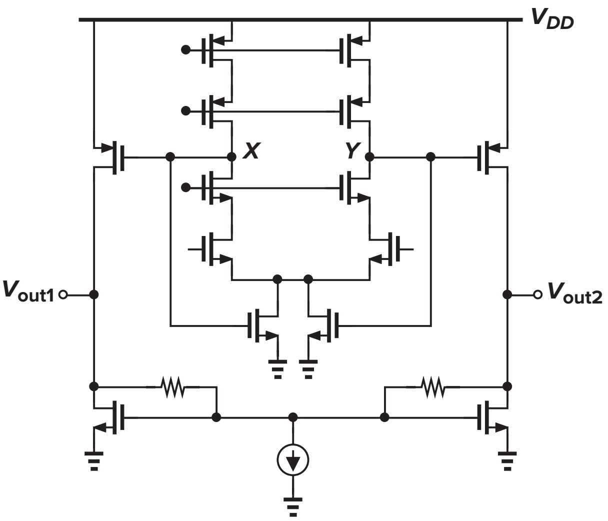


Figure 9.64 CMFB loops around cascode and output stages.

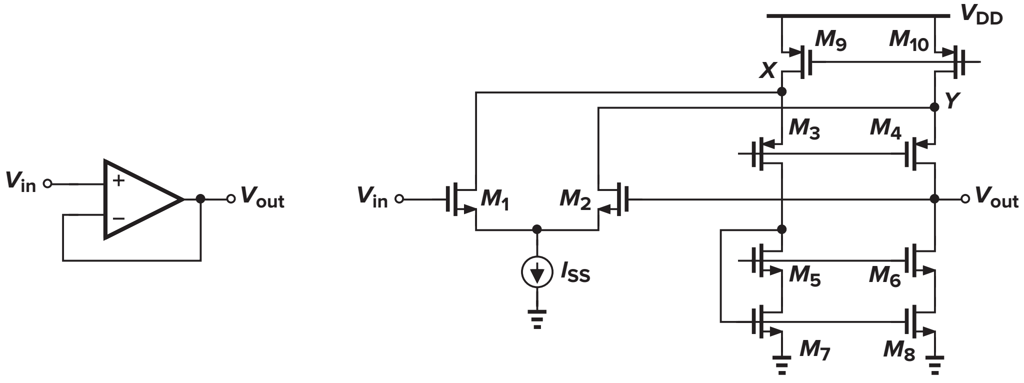


Figure 9.65 Unity-gain buffer.

cascode branch. Specifically, $V_{in,min} \approx V_{out,min} = V_{GS1,2} + V_{ISS}$, approximately one threshold voltage higher than the allowable minimum provided by M_5–M_8.

What happens if V_{in} falls below the minimum given above? The MOS transistor operating as I_{SS} enters the triode region, decreasing the bias current of the differential pair and hence lowering the transconductance. We then postulate that the limitation is overcome if the transconductance can somehow be restored.

A simple approach to extending the input CM range is to incorporate both NMOS and PMOS differential pairs such that when one is "dead," the other is "alive." Illustrated in Fig. 9.66, the idea is to combine two folded-cascode op amps with NMOS and PMOS input differential pairs. Here, as the input CM level approaches the ground potential, the NMOS pair's transconductance drops, eventually falling to zero. Nonetheless, the PMOS pair remains active, allowing normal operation. Conversely, if the input CM level approaches V_{DD}, M_{1P} and M_{2P} begin to turn off, but M_1 and M_2 function properly.

An important concern in the circuit of Fig. 9.66 is the *variation* of the overall transconductance of the two pairs as the input CM level changes. Considering the operation of each pair, we anticipate the behavior depicted in Fig. 9.67. Thus, many properties of the circuit, including gain, speed, and noise, vary. More sophisticated techniques of minimizing this variation are described in [8].

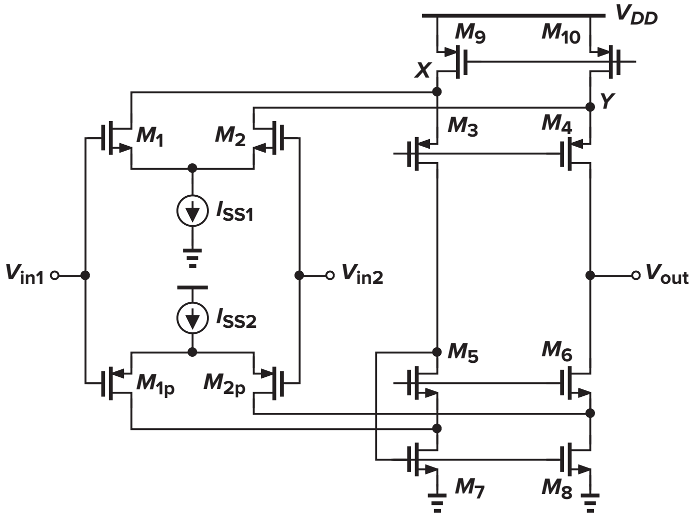


Figure 9.66 Extension of input CM range.

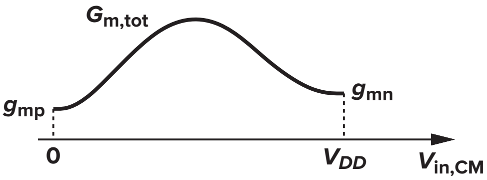


Figure 9.67 Variation of equivalent transconductance with the input CM level.

9.9 ■ Slew Rate

Op amps used in feedback circuits exhibit a large-signal behavior called "slewing." We first describe an interesting property of *linear* systems that vanishes during slewing. Consider the simple RC network shown in Fig. 9.68, where the input is an ideal voltage step of height V_0. Since $V_{out} = V_0[1 - \exp(-t/\tau)]$, where $\tau = RC$, we have

$$\frac{dV_{out}}{dt} = \frac{V_0}{\tau} \exp\frac{-t}{\tau} \tag{9.58}$$

That is, the slope of the step response is proportional to the final value of the output; if we apply a larger input step, the output rises more rapidly. This is a fundamental property of linear systems: if the input amplitude is, say, doubled while other parameters remain constant, the output signal level must double at *every* point, leading to a twofold increase in the slope.

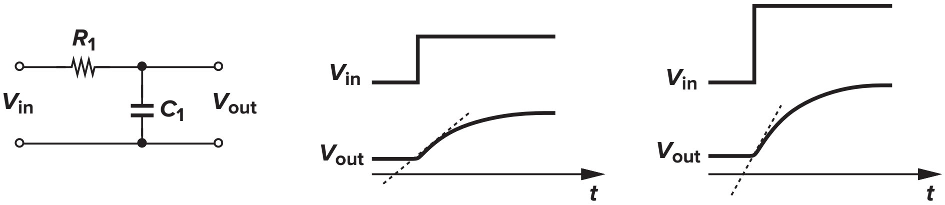


Figure 9.68 Response of a linear circuit to an input step.

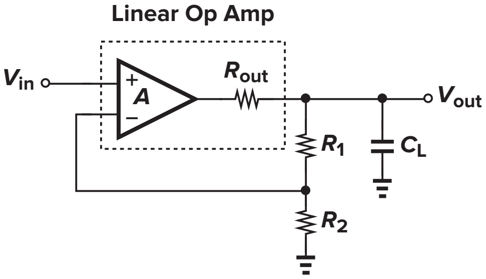


Figure 9.69 Response of linear op amp to step response.

The foregoing observation applies to linear feedback systems as well. Shown in Fig. 9.69 is an example, where the op amp is assumed linear. Here, we can write

$$\left[\left(V_{in} - V_{out}\frac{R_2}{R_1 + R_2}\right) A - V_{out}\right]\frac{1}{R_{out}} = \frac{V_{out}}{R_1 + R_2} + V_{out}C_L s \tag{9.59}$$

Assuming $R_1 + R_2 \gg R_{out}$, we have

$$\frac{V_{out}}{V_{in}}(s) \approx \frac{A}{\left(1 + A\frac{R_2}{R_1 + R_2}\right)\left[1 + \frac{R_{out}C_L}{1 + AR_2/(R_1 + R_2)}s\right]} \tag{9.60}$$

As expected, both the low-frequency gain and the time constant are divided by $1 + AR_2/(R_1 + R_2)$. The step response is therefore given by

$$V_{out} \approx V_0 \frac{A}{1 + A\frac{R_2}{R_1 + R_2}}\left[1 - \exp\frac{-t}{\frac{C_L R_{out}}{1 + AR_2/(R_1 + R_2)}}\right]u(t) \tag{9.61}$$

indicating that the slope is proportional to the final value. This type of response is called "linear settling."

With a realistic op amp, on the other hand, the step response of the circuit begins to deviate from (9.61) as the input amplitude increases. Illustrated in Fig. 9.70, the response to sufficiently small inputs follows the exponential of Eq. (9.61), but with large input steps, the output displays a linear *ramp* having a *constant slope.* Under this condition, we say that the op amp experiences slewing and call the slope of the ramp the "slew rate."

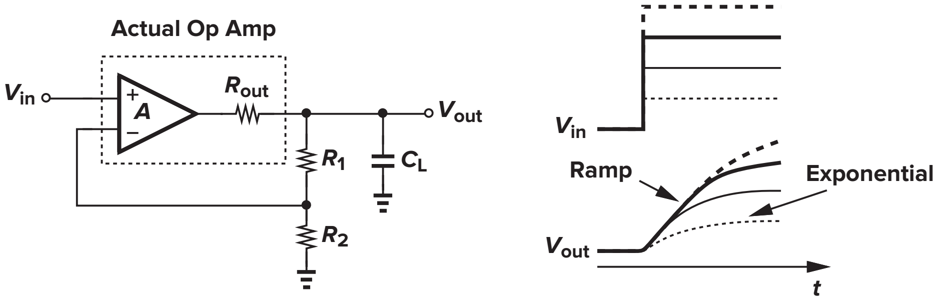


Figure 9.70 Slewing in an op amp circuit.

To understand the origin of slewing, let us replace the op amp of Fig. 9.70 by a simple CMOS implementation (Fig. 9.71), assuming for simplicity that $R_1 + R_2$ is very large. We first examine the circuit with a small input step. If V_{in} experiences a change of ΔV, I_{D1} increases by $g_m \Delta V/2$ and I_{D2} decreases by $g_m \Delta V/2$. Since the mirror action of M_3 and M_4 raises $|I_{D4}|$ by $g_m \Delta V/2$, the total small-signal current provided by the op amp equals $g_m \Delta V$. This current begins to charge C_L, but as V_{out} rises, so does V_X, reducing the difference between V_{G1} and V_{G2} and hence the output current of the op amp. As a result, V_{out} varies according to (9.61).

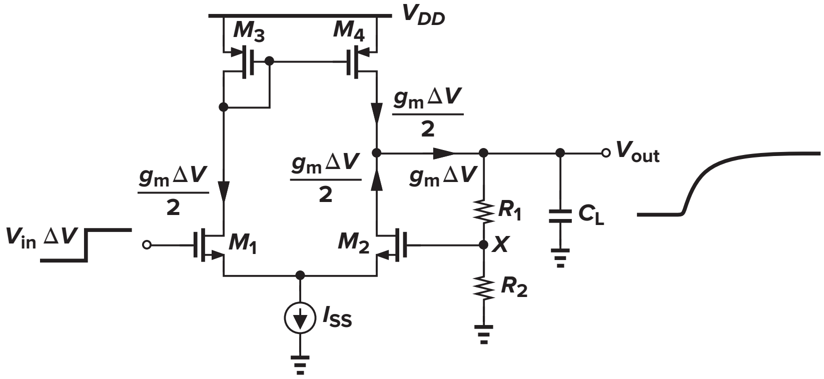


Figure 9.71 Small-signal operation of a simple op amp.

Now suppose ΔV is so large that M_1 absorbs all of I_{SS}, turning off M_2. The circuit then reduces to that shown in Fig. 9.72(a), generating a ramp output with a slope equal to I_{SS}/C_L (if the channel-length modulation of M_4 and the current drawn by $R_1 + R_2$ are neglected). Note that so long as M_2 remains off, the feedback loop is broken and the current charging C_L is constant and independent of the input level. As V_{out} rises, V_X eventually approaches V_{in}, M_2 turns on, and the circuit returns to linear operation.

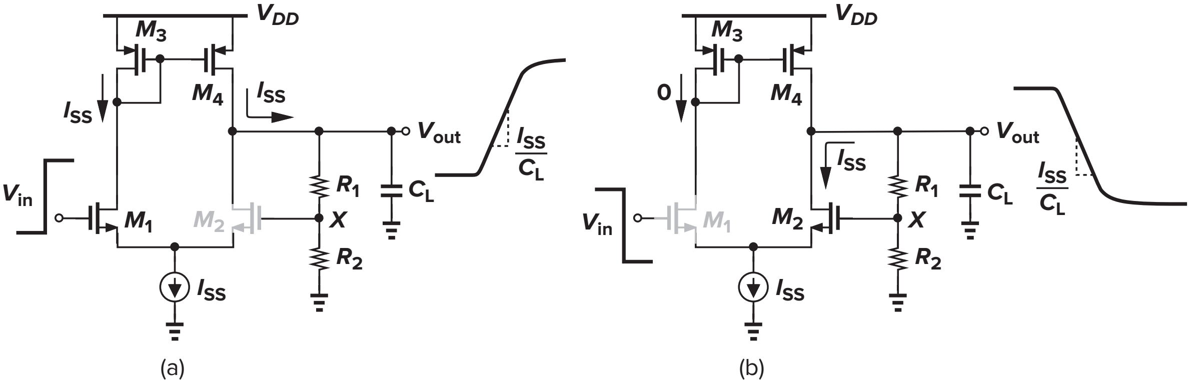


Figure 9.72 Slewing during (a) low-to-high and (b) high-to-low transitions.

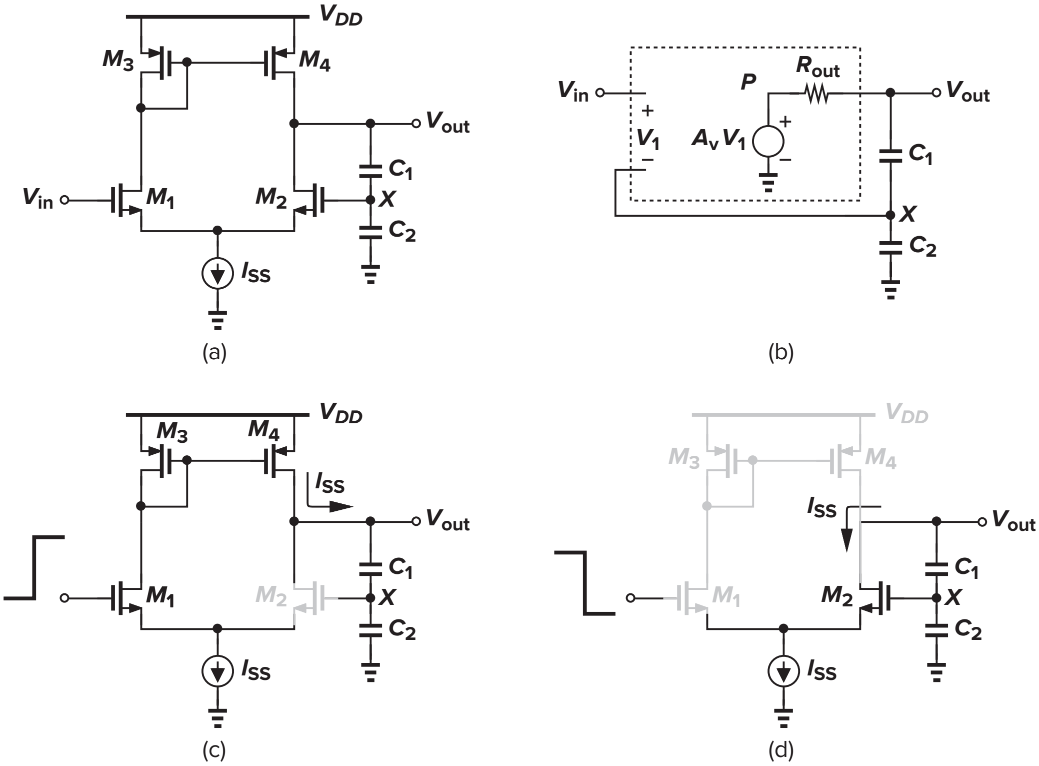
In Fig. 9.71, slewing occurs for falling edges at the input as well. If the input drops so much that M_1 turns off, then the circuit is simplified as in Fig. 9.72(b), discharging C_L by a current approximately equal to I_{SS}. After V_{out} decreases sufficiently, the difference between V_X and V_{in} is small enough to allow M_1 to turn on, leading to linear behavior thereafter.

The foregoing observations explain why slewing is a nonlinear phenomenon. If the input amplitude, say, doubles, the output level does not double at *all* points because the ramp exhibits a slope independent of the input.

Slewing is an undesirable effect in high-speed circuits that process large signals. While the small-signal bandwidth of a circuit may suggest a fast time-domain response, the large-signal speed may be limited by the slew rate simply because the current available to charge and discharge the dominant capacitor in the circuit is small. Moreover, since the input-output relationship during slewing is nonlinear, the output of a slewing amplifier exhibits substantial distortion. For example, if a circuit is to amplify a sinusoid $V_0 \sin \omega_0 t$ (in the steady state), then its slew rate must exceed $V_0\omega_0$.

▶ Example 9.22

Consider the feedback amplifier depicted in Fig. 9.73(a), where C_1 and C_2 set the closed-loop gain. (The bias network for the gate of M_2 is not shown.) (a) Determine the small-signal step response of the circuit. (b) Calculate the positive and negative slew rates.



(a) (b)

(c) (d)

Figure 9.73

Solution

(a) Modeling the op amp as in Fig. 9.73(b), where $A_v = g_{m1,2}(r_{O2}\|r_{O4})$ and $R_{out} = r_{O2}\|r_{O4}$, we have $V_X = C_1 V_{out}/(C_1 + C_2)$, and hence

$$V_P = \left(V_{in} - \frac{C_1}{C_1 + C_2} V_{out}\right) A_v \tag{9.62}$$

obtaining

$$\left[\left(V_{in} - \frac{C_1}{C_1 + C_2} V_{out}\right) A_v - V_{out}\right] \frac{1}{R_{out}} = V_{out} \frac{C_1 C_2}{C_1 + C_2} s \tag{9.63}$$

It follows that

$$\frac{V_{out}}{V_{in}}(s) = \frac{A_v}{1 + A_v \dfrac{C_1}{C_1 + C_2} + \dfrac{C_1 C_2}{C_1 + C_2} R_{out} s} \tag{9.64}$$

$$= \frac{A_v \Big/ \left(1 + A_v \dfrac{C_1}{C_1 + C_2}\right)}{1 + \dfrac{C_1 C_2}{C_1 + C_2} R_{out} s \Big/ \left(1 + A_v \dfrac{C_1}{C_1 + C_2}\right)} \tag{9.65}$$

revealing that both the low-frequency gain and the time constant of the circuit have decreased by a factor of $1 + A_v C_1/(C_1 + C_2)$. The response to a step of height V_0 is thus given by

$$V_{out}(t) = \frac{A_v}{1 + A_v \dfrac{C_1}{C_1 + C_2}} V_0 \left(1 - \exp \frac{-t}{\tau}\right) u(t) \tag{9.66}$$

where

$$\tau = \frac{C_1 C_2}{C_1 + C_2} R_{out} \Big/ \left(1 + A_v \frac{C_1}{C_1 + C_2}\right) \tag{9.67}$$

(b) Suppose a large positive step is applied to the gate of M_1 in Fig. 9.73(a) while the initial voltage across C_1 is zero. Then, M_2 turns off and, as shown in Fig. 9.73(c), V_{out} rises according to $V_{out}(t) = I_{SS}/[C_1 C_2/(C_1 + C_2)]t$. Similarly, for a large negative step at the input, Fig. 9.73(d) yields $V_{out} = -I_{SS}/[C_1 C_2/(C_1 + C_2)]t$.

As another example, let us find the slew rate of the telescopic op amp shown in Fig. 9.74(a). When a large differential input is applied, M_1 or M_2 turns off, reducing the circuit to that shown in Fig. 9.74(b). Thus, V_{out1} and V_{out2} appear as ramps with slopes equal to $\pm I_{SS}/(2C_L)$, and consequently $V_{out1} - V_{out2}$ exhibits a slew rate equal to I_{SS}/C_L. (Of course, the circuit is usually used in closed-loop form.)

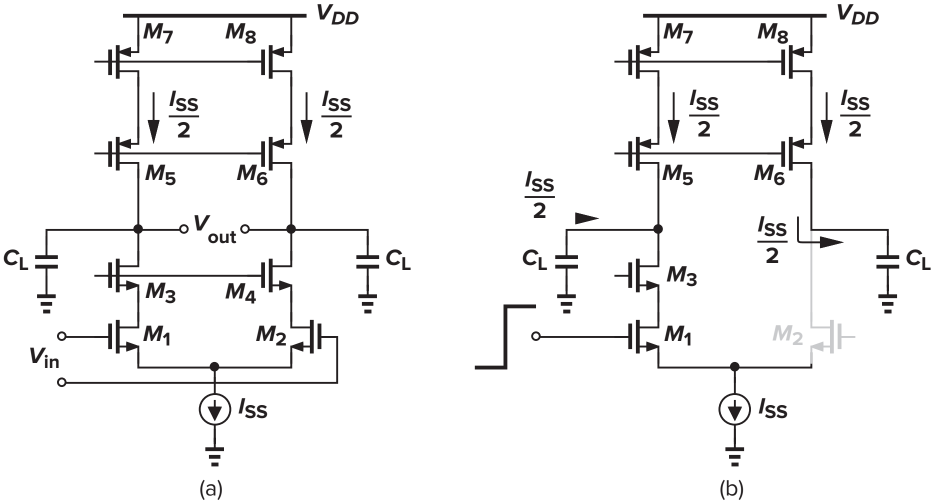
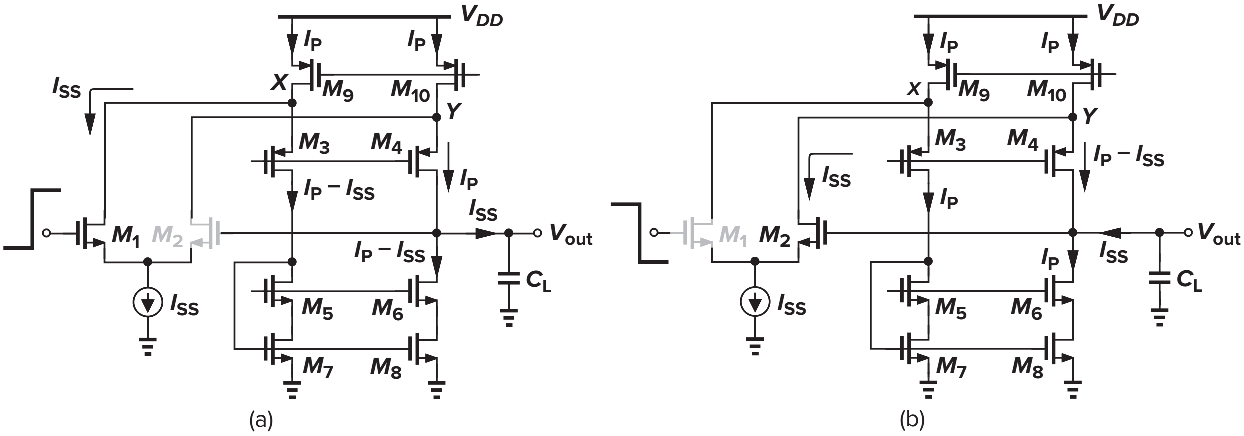


Figure 9.74 Slewing in telescopic op amp.

It is also instructive to study the slewing behavior of a folded-cascode op amp with single-ended output [Fig. 9.75(a)]. Figures 9.75(a) and (b) depict the equivalent circuit for positive and negative input steps,



(a) (b)

Figure 9.75 Slewing in folded-cascode op amp.

respectively. Here, the PMOS current sources provide a current of I_P, and the current that charges or discharges C_L is equal to I_{SS}, yielding a slew rate of I_{SS}/C_L. Note that the slew rate is independent of I_P if $I_P \geq I_{SS}$. In practice, we choose $I_P \approx I_{SS}$.

In Fig. 9.75(a), if $I_{SS} > I_P$, then during slewing, M_3 turns off and V_X falls to a low level such that M_1 and the tail current source enter the triode region. Thus, for the circuit to return to equilibrium after M_2 turns on, V_X must experience a large swing, slowing down the settling. This phenomenon is illustrated in Fig. 9.76.

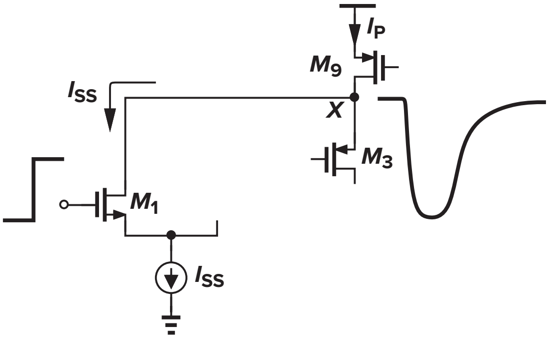


Figure 9.76 Long settling due to overdrive recovery after slewing.

To alleviate this issue, two "clamp" transistors can be added as shown in Fig. 9.77(a) [9]. The idea is that the difference between I_{SS} and I_P now flows through M_{11} or M_{12}, requiring only enough drop in V_X or V_Y to turn on one of these transistors. Figure 9.77(b) illustrates a more aggressive approach, where M_{11} and M_{12} clamp the two nodes directly to V_{DD}. Since the equilibrium value of V_X and V_Y is usually higher than $V_{DD} - V_{THN}$, M_{11} and M_{12} are off during small-signal operation.

What trade-offs are encountered in increasing the slew rate? In the examples of Figs. 9.74 and 9.75, for a given load capacitance, I_{SS} must be increased, and to maintain the same maximum output swing, all of the transistors must be made proportionally wider. As a result, the power dissipation and the input capacitance are increased. Note that if the device currents and widths scale together, $g_m r_O$ of each transistor, and hence the open-loop gain of the op amp, remain constant.

How does an op amp leave the slewing regime and enter the linear-settling regime? Since the point at which one of the input transistors "turns on" is ambiguous, the distinction between slewing and linear settling is somewhat arbitrary. The following example illustrates the point.

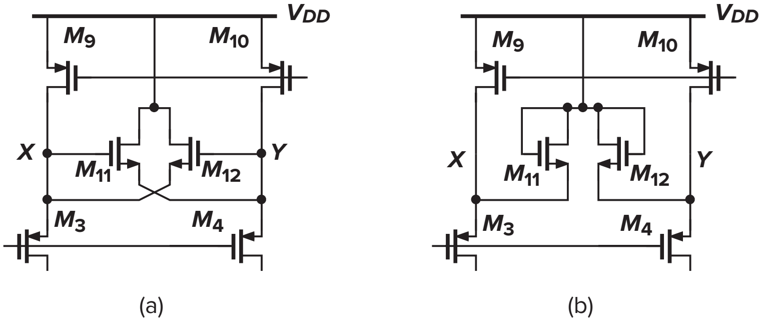


Figure 9.77 Clamp circuit to limit swings at X and Y.

▶ **Example 9.23**

Consider the circuit of Fig. 9.73(a) in the slewing regime [Fig. 9.73(c)]. As V_{out} rises, so does V_X, eventually turning M_2 on. As I_{D2} increases from zero, the differential pair becomes more linear. Considering M_1 and M_2 to operate linearly if the difference between their drain currents is less than αI_{SS} (e.g., $\alpha = 0.1$), determine how long the circuit takes to enter linear settling. Assume the input step has an amplitude of V_0.

Solution

The circuit displays a slew rate of $I_{SS}/[C_1C_2/(C_1+C_2)]$ until $|V_{in1} - V_{in2}|$ is sufficiently small. From Chapter 4, we can write

$$\alpha I_{SS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in1} - V_{in2})\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - (V_{in1} - V_{in2})^2} \tag{9.68}$$

obtaining

$$\Delta V_G^4 - \Delta V_G^2\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} + \left(\frac{2\alpha I_{SS}}{\mu_n C_{ox}\frac{W}{L}}\right)^2 = 0 \tag{9.69}$$

where $\Delta V_G = V_{in1} - V_{in2}$. Thus,

$$\Delta V_G \approx \alpha\sqrt{\frac{I_{SS}}{\mu_n C_{ox}\frac{W}{L}}} \tag{9.70}$$

(Recall that $\sqrt{I_{SS}/[\mu_n C_{ox}(W/L)]}$ is the equilibrium overdrive voltage of each transistor in the differential pair.) Alternatively, we recognize that for a small difference, αI_{SS}, between I_{D1} and I_{D2}, a small-signal approximation is valid: $\alpha I_{SS} = g_m \Delta V_G$. Thus, $\Delta V_G = \alpha I_{SS}/g_m \approx \alpha I_{SS}/\sqrt{\mu_n C_{ox}(W/L)I_{SS}}$. Note that this is a rough calculation because as M_2 turns on, the current charging the load capacitance is no longer constant.

Since V_X must rise to $V_0 - \Delta V_G$ for M_2 to carry the required current, V_{out} increases by $(V_0 - \Delta V_G)(1 + C_2/C_1)$, requiring a time given by

$$t = \frac{C_2}{I_{SS}}\left(V_0 - \alpha\sqrt{\frac{I_{SS}}{\mu_n C_{ox}\frac{W}{L}}}\right) \tag{9.71}$$

◀

In the earlier example, the value of α that determines the onset of linear settling depends, among other things, on the actual required linearity. In other words, for a nonlinearity of 1%, α can be quite a lot larger than for a nonlinearity of 0.1%.

The slewing behavior of two-stage op amps is somewhat different from that of the circuits studied earlier. This case is studied in Chapter 10.

9.10 ■ High-Slew-Rate Op Amps

Our formulation of the slew rate in various op amp topologies implies that, for a given capacitance, slew-limited settling can be improved only by raising the bias current and hence the power consumption. This trade-off can be mitigated if the current available to charge the capacitor of interest automatically rises during slewing and falls back to its original value thereafter. In this section, we study op amp topologies that exploit this idea.

9.10.1 One-Stage Op Amps

We begin with a simple common-source stage incorporating a current-source load biased at a value of I_0 [Fig. 9.78(a)]. In the absence of an input signal, $I_{D1} = I_0$, but if V_{in} jumps down to turn M_1 off, then I_0 flows through C_L, yielding a slew rate of I_0/C_L.[7] Can we automatically increase the drain current of M_2 during this transient? To this end, we must allow V_b to change and, in fact, follow the jump in V_{in}. For example, as shown in Fig. 9.78(b), we can simply apply V_{in} to both transistors so that a downward jump in V_{in} also raises $|I_{D2}|$. This complementary topology was studied in Chapter 3 and found to suffer from poor power supply rejection. We pursue other topologies here.

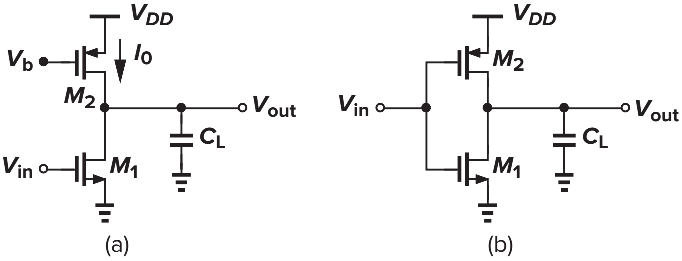


Figure 9.78 Slewing in (a) a simple CS stage and (b) a complementary CS stage.

Let us control M_2 in Fig. 9.78(a) by current mirror action, as depicted in Fig. 9.79(a), and ask how I_b must be controlled by V_{in}. Can I_b be derived from another common-source device [Fig. 9.79(b)]? No; as V_{in} jumps down in this circuit, I_b *decreases*. We must therefore include an additional signal inversion in the path controlling I_b. Alternatively, we can consider a differential topology, where both the input signal, V_{in}^+, and its inverted version, V_{in}^-, are available. Illustrated in Fig. 9.79(c), the idea is to control the bias current of M_2 by V_{in}^- and that of M_4 by V_{in}^+. For example, if V_{in}^+ jumps down and V_{in}^- jumps up, then (1) M_5 draws less current from M_8, lowering $|I_{D4}|$, (2) M_3 draws more current, discharging its load capacitance, (3) M_6 draws more current from M_7, raising $|I_{D2}|$, and (4) M_1 draws less current, allowing its drain capacitance to be charged by M_2.

The circuits of Figs. 9.78(b) and 9.79(c) are called "push-pull" stages as they turn the load current source into an "active" pull-up device. Loosely speaking, we also refer to them as "class-AB" amplifiers.[8]

[7] If V_{in} jumps *up*, M_1 must absorb both I_0 and the current flowing out of C_L.

[8] By contrast, topologies with a constant bias current are called "class-A" amplifiers.

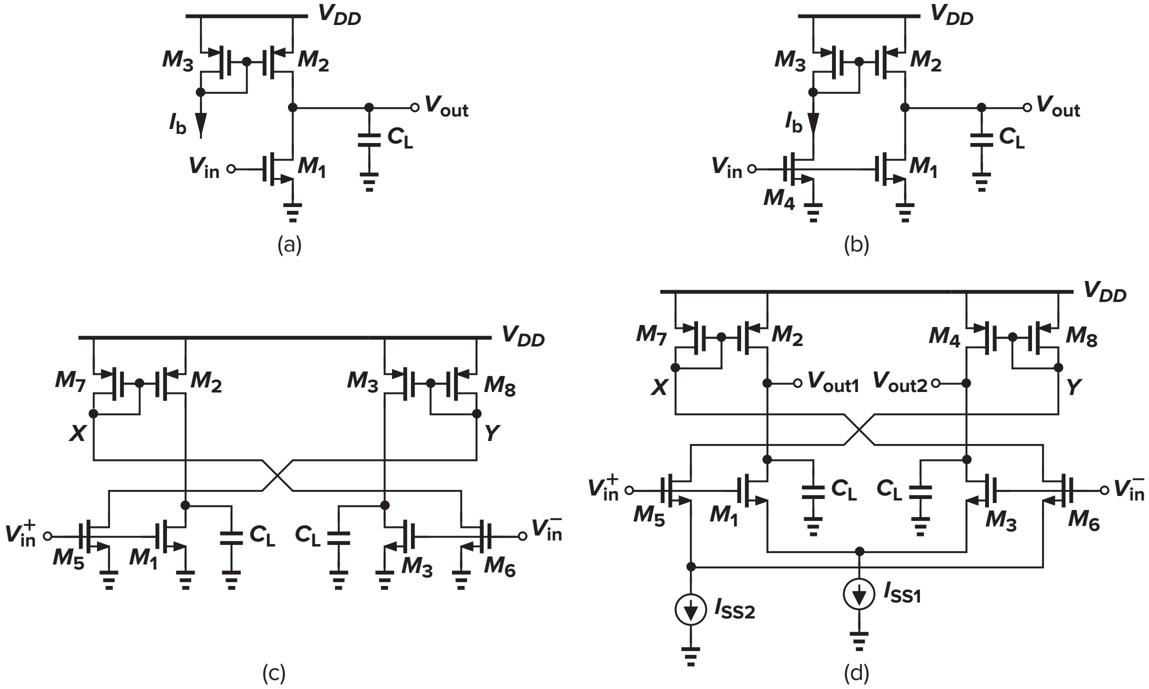


Figure 9.79 (a) CS stage with current mirror biasing, (b) injection of signal into the mirror with incorrect polarity, (c) injection of signal into the mirror with ccrrect polarity, and (d) addition of tail current sources.

By virtue of the temporary boost in the slew rate, such circuits alleviate the trade-off between the speed and the average power consumption.

In order to improve the input common-mode rejection, we add tail current sources to M_1 and M_3 and to M_5 and M_6 [Fig. 9.79(d)]. We now wish to calculate the circuit's slew rate with a large input step. If, for example, V_{in}^+ jumps up and M_1 and M_5 absorb all of their respective tail currents, then M_2 is off and V_{out1} falls at a rate of I_{SS1}/C_L while M_3 is off and V_{out2} rises at a rate of $I_{SS2}(W_4/W_8)/C_L$ (if $L_4 = L_8$). The differential slew rate is thus equal to $[I_{SS1} + I_{SS2}(W_4/W_8)]/C_L$. Without the push-pull action, on the other hand, this slew rate would be limited to I_{SS1}/C_L. If we choose W_4/W_8 equal to, say, 5 and I_{SS2} equal to I_{SS1}, then the SR increases by a factor of 6 with a twofold power penalty.[9]

▶ Example 9.24

Calculate the small-signal voltage gain of the class-AB op amp shown in Fig. 9.79(d).

Solution

In addition to the main path, the mirror path contributes gain as well. Since the mirror action amplifies the drain currents of M_5 and M_6 by a factor of W_4/W_8, we approximate the gain in this path as $(W_4/W_8)g_{m5}(r_{O3}||r_{O4})$ and add it to that of the main path:

$$|A_v| \approx g_{m1}(r_{O3}||r_{O4}) + (W_4/W_8)g_{m5}(r_{O3}||r_{O4}) \tag{9.72}$$

$$\approx [g_{m1} + (W_4/W_8)g_{m5}](r_{O3}||r_{O4}) \tag{9.73}$$

The mirror path thus raises the apparent transconductance from g_{m1} to $g_{m1} + (W_4/W_8)g_{m5}$. ◀

[9] One can argue that the fixed tail currents no longer allow class-AB operation, but we disregard this subtlety for now.

Let us now determine the transfer function of the above circuit and examine the effect of the mirror pole. We write the transfer function from the input through the mirror path to the output as

$$H_{mirr}(s) = \frac{W_4}{W_8} g_{m5}(r_{O3}||r_{O4}) \frac{1}{1+\dfrac{s}{\omega_{p,X}}} \frac{1}{1+\dfrac{s}{\omega_{out}}} \tag{9.74}$$

where $\omega_{p,X} \approx g_{m8}/C_Y$ and $\omega_{out} = [(r_{O3}||r_{O4})C_L]^{-1}$. For the main path, we simply have

$$H_{main}(s) = g_{m1}(r_{O3}||r_{O4}) \frac{1}{1+\dfrac{s}{\omega_{out}}} \tag{9.75}$$

It follows that

$$H_{tot}(s) = H_{main}(s) + H_{mirr}(s) \tag{9.76}$$

$$= \frac{r_{O3}||r_{O4}}{1+\dfrac{s}{\omega_{out}}} \left[\frac{W_4}{W_8} \frac{g_{m5}}{1+\dfrac{s}{\omega_{p,X}}} + g_{m1} \right] \tag{9.77}$$

$$= \frac{r_{O3}||r_{O4}}{1+\dfrac{s}{\omega_{out}}} \cdot \frac{(W_4/W_8)g_{m5} + g_{m1} + g_{m1}s/\omega_{p,X}}{1+\dfrac{s}{\omega_{p,X}}} \tag{9.78}$$

As seen in other examples in Chapter 6, the presence of the additional signal path leads to a zero in the transfer function. This zero frequency is given by

$$|\omega_z| = \left(\frac{W_4}{W_8} \frac{g_{m5}}{g_{m1}} + 1 \right) \omega_{p,X} \tag{9.79}$$

Unfortunately, it is not possible to equate ω_z to $\omega_{p,X}$ because $(W_4/W_8)(g_{m5}/g_{m1})$ is typically around unity or higher. Also, in practice, $\omega_{out} < \omega_{p,X}$.

It is tempting to raise the SR in Fig. 9.79(d) by increasing W_4/W_8, but we must note that, as a result, the pole frequency associated with the mirror nodes falls. Approximating this pole by g_{m8}/C_Y and writing $g_{m8} = \sqrt{I_{SS2}\mu_n C_{ox}(W/L)_8}$ and $C_Y \approx 2(W_4 + W_8)LC_{ox} + C_{DB8} + C_{DB5}$, we recognize that the mirror pole frequency is inversely proportional to W_4.

9.10.2 Two-Stage Op Amps

In order to achieve a high slew rate, we can apply push-pull operation to the second stage of a two-stage op amp. To this end, we view the arrangement shown in Fig. 9.79(c) as the second stage and precede it with a differential pair, arriving at the topology depicted in Fig. 9.80. This circuit provides a voltage gain of

$$|A_v| = g_{m9}(r_{O9}||r_{O11})[g_{m1} + (W_4/W_8)g_{m5}](r_{O1}||r_{O2}) \tag{9.80}$$

But how about the slew rate? Suppose, for example, V_{in1} and V_{in2} experience a large differential step such that the entire I_{SS} flows through node P. If this node is "agile" enough, i.e., if its capacitance is relatively small, V_P rises rapidly, applying a large overdrive to M_1 and M_5 and creating a high slew rate at the output. In other words, since V_P (or V_Q) can reach near V_{DD} when only M_9 (or M_{10}) is on, the available current is much larger than the bias current of the output stage. This behavior stands in contrast to that

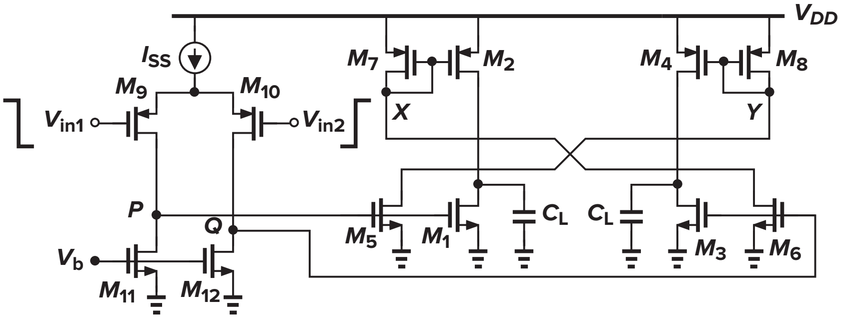


Figure 9.80 Two-stage op amp with slew enhancement.

of the circuit in Fig. 9.79(d), where the available current is a multiple of the tail currents and cannot be raised further "upon demand."

We return to this two-stage op amp in Chapter 10 and analyze its slew rate in the presence of frequency compensation.

9.11 ■ Power Supply Rejection

As with other analog circuits, op amps are often supplied from noisy lines and must therefore "reject" the noise adequately. For this reason, it is important to understand how noise on the supply manifests itself at the output of an op amp.

Let us consider the simple op amp shown in Fig. 9.81, assuming that the supply voltage varies slowly. If the circuit is perfectly symmetric, $V_{out} = V_X$. Since the diode-connected device "clamps" node X to V_{DD}, V_X and hence V_{out} experience approximately the same change as does V_{DD}. In other words, the gain from V_{DD} to V_{out} is close to unity. The power supply rejection ratio (PSRR) is defined as the gain from the input to the output divided by the gain from the supply to the output. At low frequencies:

$$\text{PSRR} \approx g_{mN}(r_{OP}\|r_{ON}) \tag{9.81}$$

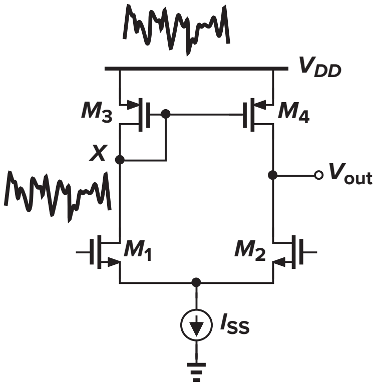


Figure 9.81 Supply rejection of differential pair with active current mirror.

▶ **Example 9.25**

Calculate the low-frequency PSRR of the feedback circuit shown in Fig. 9.82(a).

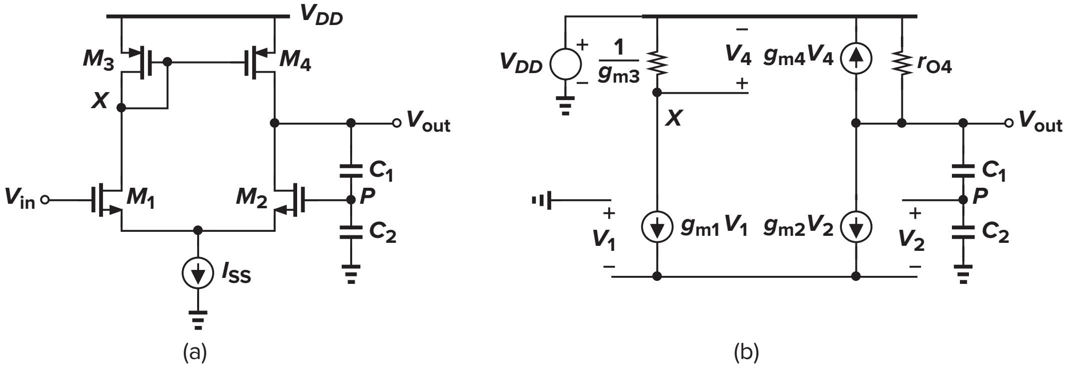


Figure 9.82

Solution

From the foregoing analysis, we may surmise that a change ΔV in V_{DD} appears unattenuated at the output. But, we should note that if V_{out} changes, so do V_P and I_{D2}, thereby opposing the change. Using Fig. 9.82(b) and neglecting channel-length modulation in M_1–M_3 for simplicity, we can write

$$V_{out}\frac{C_1}{C_1+C_2} - V_2 = -V_1 \tag{9.82}$$

and $g_{m1}V_1 + g_{m2}V_2 = 0$. Thus, if the circuit is symmetric,

$$V_2 = \frac{V_{out}}{2}\frac{C_1}{C_1+C_2} \tag{9.83}$$

We also have

$$-\frac{g_{m1}V_1}{g_{m3}}g_{m4} - \frac{V_{DD}-V_{out}}{r_{O4}} + g_{m2}V_2 = 0 \tag{9.84}$$

It follows that

$$\frac{V_{out}}{V_{DD}} = \frac{1}{g_{m2}r_{O4}\dfrac{C_1}{C_1+C_2}+1} \tag{9.85}$$

Thus,

$$\text{PSRR} \approx (1+\frac{C_2}{C_1})(g_{m2}r_{O4}\frac{C_1}{C_1+C_2}+1) \tag{9.86}$$

$$\approx g_{m2}r_{O4} \tag{9.87}$$

The denominator of Eq. (9.85) looks like one plus a loop gain. Is that true? Let us set the main input in Fig. 9.82(a) to zero and view the path from V_{DD} to V_{out} as an amplifier [Fig. 9.83(a)], omitting C_1 and C_2. In this case, the gain, $\partial V_{out}/\partial V_{DD}$, is equal to unity. Now, as depicted in Fig. 9.83(b), we sense V_{out} by means of a capacitive divider and return the result to some node within the amplifier. We expect the gain to drop by one plus the loop gain associated with the feedback loop. Indeed, this loop gain is equal to $[C_1/(C_1+C_2)]g_{m2}r_{O4}$ if channel-length modulation is neglected for M_1–M_3. We therefore recognize that feedback reduces $\partial V_{out}/\partial V_{DD}$ and $\partial V_{out}/\partial V_{in}$ by the same factor, leaving the PSRR relatively constant.

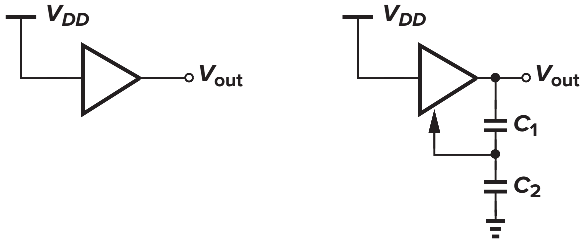


Figure 9.83 Equivalent circuits for path from V_{DD} to output.

9.12 ■ Noise in Op Amps

In low-noise applications, the input-referred noise of op amps becomes critical. We now extend the noise analysis of differential amplifiers in Chapter 7 to more sophisticated topologies. With many transistors in an op amp, it may seem difficult to intuitively identify the dominant sources of noise. A simple rule for inspection is to (mentally) change the gate voltage of each transistor by a small amount and predict the effect at the output.

Let us first consider the telescopic op amp shown in Fig. 9.84. At relatively low frequencies, the cascode devices contribute negligible noise, leaving M_1–M_2 and M_7–M_8 as the primary noise sources. The input-referred noise voltage per unit bandwidth is therefore similar to that in Fig. 7.59(a) and given by

$$\overline{V_n^2} = 4kT\left(2\frac{\gamma}{g_{m1,2}} + 2\frac{\gamma g_{m7,8}}{g_{m1,2}^2}\right) + 2\frac{K_N}{(WL)_{1,2}C_{ox}f} + 2\frac{K_P}{(WL)_{7,8}C_{ox}f}\frac{g_{m7,8}^2}{g_{m1,2}^2} \tag{9.88}$$

where K_N and K_P denote the $1/f$ noise coefficients of NMOS and PMOS devices, respectively.

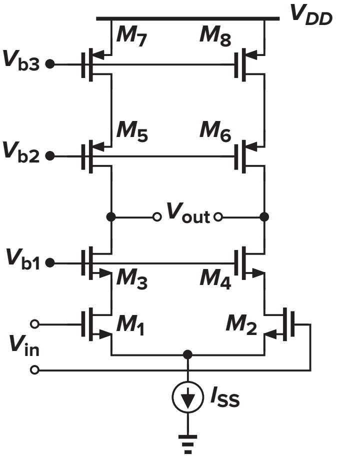
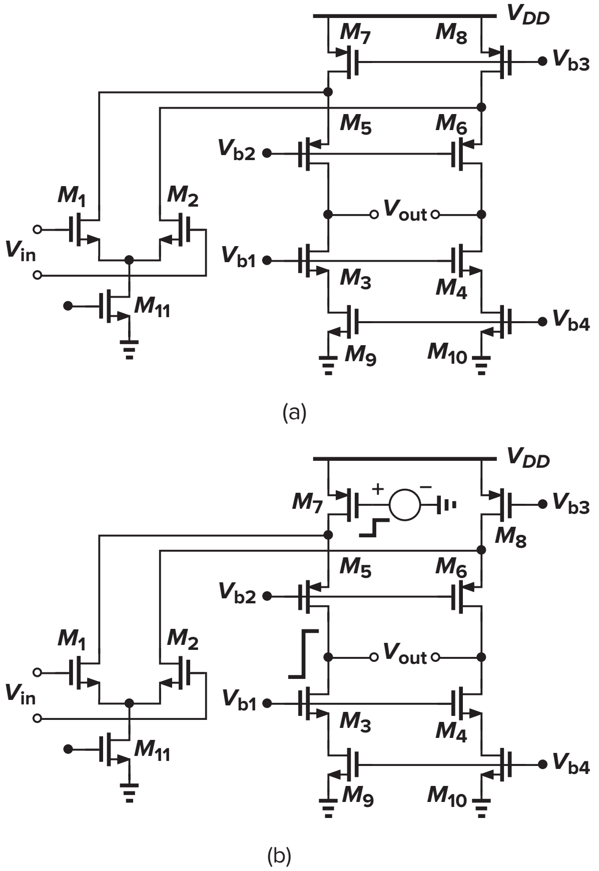


Figure 9.84 Noise in a telescopic op amp.

Next, we study the noise behavior of the folded-cascode op amp of Fig. 9.85(a), considering only thermal noise at this point. Again, the noise of the cascode devices is negligible at low frequencies, leaving M_1–M_2, M_7–M_8, and M_9–M_{10} as potentially significant sources. Do both pairs M_7–M_8 and M_9–M_{10} contribute noise? Using our simple rule, we change the gate voltage of M_7 by a small amount [Fig. 9.85(b)], noting that the output indeed changes considerably. The same observation applies to M_8–M_{10} as well. To determine the input-referred thermal noise, we first refer the noise of M_7–M_8 to the



(a)

(b)

Figure 9.85 Noise in a folded-cascode op amp.

output:

$$\overline{V_{n,out}^2}\Big|_{M7,8} = 2\left(4kT\frac{\gamma}{g_{m7,8}}g_{m7,8}^2R_{out}^2\right) \tag{9.89}$$

where the factor 2 accounts for the (uncorrelated) noise of M_7 and M_8 and R_{out} denotes the open-loop output resistance of the op amp. Similarly,

$$\overline{V_{n,out}^2}\Big|_{M9,10} = 2\left(4kT\frac{\gamma}{g_{m9,10}}g_{m9,10}^2R_{out}^2\right) \tag{9.90}$$

Dividing these quantities by $g_{m1,2}^2R_{out}^2$ and adding the contribution of M_1–M_2, we obtain the overall noise:

$$\overline{V_{n,int}^2} = 8kT\left(\frac{\gamma}{g_{m1,2}} + \gamma\frac{g_{m7,8}}{g_{m1,2}^2} + \gamma\frac{g_{m9,10}}{g_{m1,2}^2}\right) \tag{9.91}$$

The effect of flicker noise can be included in a similar manner (Problem 9.15). Note that the folded-cascode topology potentially suffers from greater noise than its telescopic counterpart. In applications where flicker noise is critical, we opt for a PMOS-input op amp as PMOS transistors typically exhibit less flicker noise than do NMOS devices.

As observed for the differential amplifiers in Chapter 7, the noise contribution of the PMOS and NMOS current sources *increases* in proportion to their transconductance. This trend results in a trade-off between output voltage swings and input-referred noise: for a given current, as implied by $g_m = 2I_D/(V_{GS} - V_{TH})$, if the overdrive voltage of the current sources is minimized to allow large swings, then their transconductance is maximized.

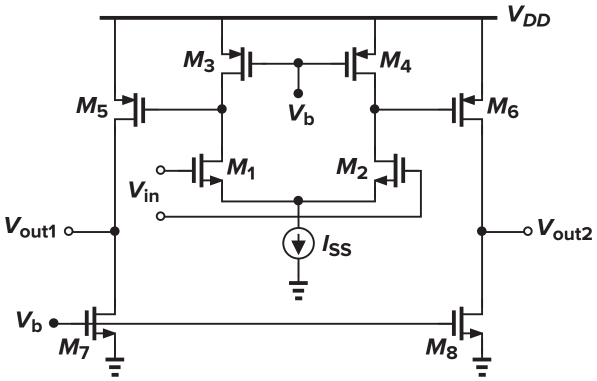


Figure 9.86 Noise in a two-stage op amp.

As another case, we calculate the input-referred thermal noise of the two-stage op amp shown in Fig. 9.86. Beginning with the second stage, we note that the noise current of M_5 and M_7 flows through $r_{O5}\|r_{O7}$. Dividing the resulting output noise voltage by the total gain, $g_{m1}(r_{O1}\|r_{O3}) \times g_{m5}(r_{O5}\|r_{O7})$, and doubling the power, we obtain the input-referred contribution of M_5–M_8:

$$\overline{V_n^2}\Big|_{M5-8} = 2 \times 4kT\gamma(g_{m5} + g_{m7})(r_{O5}\|r_{O7})^2 \frac{1}{g_{m1}^2(r_{O1}\|r_{O3})^2 g_{m5}^2(r_{O5}\|r_{O7})^2} \tag{9.92}$$

$$= 8kT\gamma \frac{g_{m5} + g_{m7}}{g_{m1}^2 g_{m5}^2 (r_{O1}\|r_{O3})^2} \tag{9.93}$$

The noise due to M_1–M_4 is simply equal to

$$\overline{V_n^2}\Big|_{M1-4} = 2 \times 4kT\gamma \frac{g_{m1} + g_{m3}}{g_{m1}^2} \tag{9.94}$$

It follows that

$$\overline{V_{n,tot}^2} = 8kT\gamma \frac{1}{g_{m1}^2}\left[g_{m1} + g_{m3} + \frac{g_{m5} + g_{m7}}{g_{m5}^2(r_{O1}\|r_{O3})^2}\right] \tag{9.95}$$

Note that the noise resulting from the second stage is usually negligible because it is divided by the gain of the first stage when referred to the main input.

▶ **Example 9.26**

A simple amplifier is constructed as shown in Fig. 9.87. Note that the first stage incorporates diode-connected—rather than current-source—loads. Assuming that all of the transistors are in saturation and $(W/L)_{1,2} = 50/0.6$, $(W/L)_{3,4} = 10/0.6$, $(W/L)_{5,6} = 20/0.6$, and $(W/L)_{7,8} = 56/0.6$, calculate the input-referred noise voltage if $\mu_n C_{ox} = 75\ \mu\text{A/V}^2$, $\mu_p C_{ox} = 30\ \mu\text{A/V}^2$, and $\gamma = 2/3$.

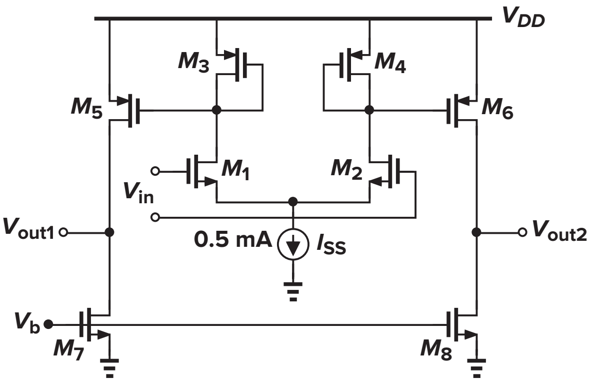


Figure 9.87

Solution

We first calculate the small-signal gain of the first stage:

$$A_{v1} \approx \frac{g_{m1}}{g_{m3}} \tag{9.96}$$

$$= \sqrt{\frac{50 \times 75}{10 \times 30}} \tag{9.97}$$

$$\approx 3.54 \tag{9.98}$$

The noise of M_5 and M_7 referred to the gate of M_5 is equal to $4kT(2/3)(g_{m5} + g_{m7})/g_{m5}^2 = 2.87 \times 10^{-17}$ V^2/Hz, which is divided by A_{v1}^2 when referred to the main input: $\overline{V_n^2}|_{M5,7} = 2.29 \times 10^{-18}$ V^2/Hz. Transistors M_1 and M_3 produce an input-referred noise of $\overline{V_n^2}|_{M1,3} = (8kT/3)(g_{m3} + g_{m1})/g_{m1}^2 = 1.10 \times 10^{-17}$ V^2/Hz. Thus, the total input-referred noise equals

$$\overline{V_{n,in}^2} = 2(2.29 \times 10^{-18} + 1.10 \times 10^{-17}) \tag{9.99}$$

$$= 2.66 \times 10^{-17} \text{ V}^2/\text{Hz} \tag{9.100}$$

where the factor of 2 accounts for the noise produced by both odd-numbered and even-numbered transistors in the circuit. This value corresponds to an input noise voltage of 5.16 nV/$\sqrt{\text{Hz}}$.

The noise-power trade-off described in Chapter 7 is present in op amps as well. Specifically, the devices and bias currents in an op amp can be linearly scaled so as to trade power consumption for noise. For example, if all of the transistor widths and I_{SS} in Fig. 9.87 are halved, then so is the power, while $\overline{V_{n,in}^2}$ is doubled and the voltage gain and swings remain unchanged. This simple scaling can be applied to all of the op amps studied in this chapter. We exploit this principle in the nanometer op amps designed in Chapter 11.

References

[1] R. G. Eschauzier, L. P. T. Kerklaan, and J. H. Huising, “A 100-MHz 100-dB Operational Amplifier with Multipath Nested Miller Compensation Structure,” *IEEE J. of Solid-State Circuits*, vol. 27, pp. 1709–1717, December 1992.

[2] R. M. Ziazadeh, H. T. Ng, and D. J. Allstot, “A Multistage Amplifier Topology with Embedded Tracking Compensation,” *CICC Proc.*, pp. 361–364, May 1998.

[3] F. You, S. H. Embabi, and E. Sanchez-Sincencio, “A Multistage Amplifier Topology with Nested G_m-C Compensation for Low-Voltage Application,” *ISSCC Dig. of Tech. Papers*, pp. 348–349, February 1997.

[4] B. J. Hosticka, “Improvement of the Gain of CMOS Amplifiers,” *IEEE J. of Solid-State Circuits*, vol. 14, pp. 1111–1114, December 1979.

[5] K. Bult and G. J. G. H. Geelen, “A Fast-Settling CMOS Operational Amplifier for SC Circuits with 90-dB DC Gain,” *IEEE J. of Solid-State Circuits*, vol. 25, pp. 1379–1384, December 1990.

[6] E. Sackinger and W. Guggenbuhl, “A High-Swing High-Impedance MOS Cascode Circuit,” *IEEE J. of Solid-State Circuits*, vol. 25, pp. 289–298, February 1990.

[7] A. Verma and B. Razavi, “A 10-Bit 500-MS/s 55-mW CMOS ADC,” *IEEE J. of Solid-State Circuits*, vol. 44, pp. 3039–3050, November 2009.

[8] R. Hogervost et al., “A Compact Power-Efficient 3-V CMOS Rail-to-Rail Input/Output Operational Amplifier for VLSI Cell Libraries,” *IEEE J. of Solid-State Circuits*, vol. 29, pp. 1505–1513, December 1994.

[9] D. A. Johns and K. Martin, *Analog Integrated Circuit Design* (New York; Wiley, 1997).

[10] P. E. Allen, B. J. Blalock, and G. A. Rincon, “A 1-V CMOS Op Amp Using Bulk-Driven MOSFETs,” *ISSCC Dig. of Tech. Papers*, pp.192–193, February 1995.

[11] S. Rabii and B. A. Wooley, “A 1.8-V Digital-Audio Sigma-Delta Modulator in 0.8-μm CMOS,” *IEEE J. of Solid-State Circuits*, vol. 32, pp. 783–796, June 1997.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

9.1. **(a)** Derive expressions for the transconductance and output resistance of a MOSFET in the triode region. Plot these quantities and $g_m r_O$ as a function of V_{DS}, covering both triode and saturation regions.

(b) Consider the amplifier of Fig. 9.6(b), with $(W/L)_{1-4} = 50/0.5$, $I_{SS} = 1$ mA, and input CM level of 1.3 V. Calculate the small-signal gain and the maximum output swing if all transistors remain in saturation.

(c) For the circuit of part (b), suppose we allow each PMOS device to enter the triode region by 50 mV so as to increase the allowable differential swing by 100 mV. What is the small-signal gain at the peaks of the output swing?

9.2. In the circuit of Fig. 9.9, assume that $(W/L)_{1-4} = 100/0.5$, $I_{SS} = 1$ mA, $V_b = 1.4$ V, and $\gamma = 0$.

(a) If M_5–M_8 are identical and have a length of 0.5 μm, calculate their minimum width such that M_3 operates in saturation.

(b) Calculate the maximum output voltage swing.

(c) What is the open-loop voltage gain?

(d) Calculate the input-referred thermal noise voltage.

9.3. Design the folded-cascode op amp of Fig. 9.15 for the following requirements: maximum differential swing $= 2.4$ V, total power dissipation $= 6$ mW. If all of the transistors have a channel length of 0.5 μm, what is the overall voltage gain? Can the input common-mode level be as low as zero?

9.4. In the op amp of Fig. 9.21(b), $(W/L)_{1-8} = 100/0.5$, $I_{SS} = 1$ mA, and $V_{b1} = 1.7$ V. Assume that $\gamma = 0$.

(a) What is the maximum allowable input CM level?

(b) What is V_X?

(c) What is the maximum allowable output swing if the gate of M_2 is connected to the output?

(d) What is the acceptable range of V_{b2}?

(e) What is the input-referred thermal noise voltage?

9.5. Design the op amp of Fig. 9.21(b) for the following requirements: maximum differential swing $= 2.4$ V, total power dissipation $= 6$ mW. (Assume that the gate of M_2 is never shorted to the output.)

9.6. If in Fig. 9.23, $(W/L)_{1-8} = 100/0.5$ and $I_{SS} = 1$ mA,

(a) What CM level must be established at the drains of M_3 and M_4 so that $I_{D5} = I_{D6} = 1$ mA? How does this constrain the maximum input CM level?

(b) With the choice made in part (a), calculate the overall voltage gain and the maximum output swing.

9.7. Design the op amp of Fig. 9.23 for the following requirements: maximum differential swing = 4 V, total power dissipation = 6 mW, $I_{SS} = 0.5$ mA.

9.8. Suppose the circuit of Fig. 9.24 is designed with I_{SS} equal to 1 mA, I_{D9}–I_{D12} equal to 0.5 mA, and $(W/L)_{9-12} = 100/0.5$.

(a) What CM level is required at X and Y?

(b) If I_{SS} requires a minimum voltage of 400 mV, choose the minimum dimensions of M_1–M_8 to allow a peak-to-peak swing of 200 mV at X and at Y.

(c) Calculate the overall voltage gain.

9.9. In Fig. 9.88, calculate the input-referred thermal noise if I_1 and I_2 are implemented by PMOS devices.

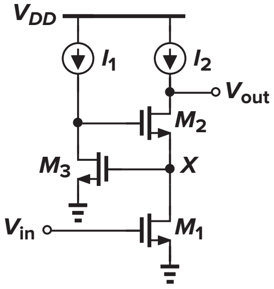


Figure 9.88

9.10. Suppose that in Fig. 9.88, $I_1 = 100\ \mu$A, $I_2 = 0.5$ mA, and $(W/L)_{1-3} = 100/0.5$. Assuming that I_1 and I_2 are implemented with PMOS devices having $(W/L)_P = 50/0.5$,

(a) Calculate the gate bias voltages of M_2 and M_3.

(b) Determine the maximum allowable output voltage swing.

(c) Calculate the overall voltage gain and the input-referred thermal noise voltage.

9.11. In the circuit of Fig. 9.53, each branch is biased at a current of 0.5 mA. Choose the dimensions of M_7 and M_8 such that the output CM level is equal to 1.5 V and $V_P = 100$ mV.

9.12. Consider the CMFB network in Fig. 9.51. The amplifier sensing $V_{out,CM}$ is to be implemented as a different pair with active current mirror load.

(a) Should the input pair of the amplifier use PMOS devices or NMOS devices?

(b) Calculate the loop gain for the CMFB network.

9.13. Repeat Problem 9.9.12**b** for the circuit of Fig. 9.52.

9.14. In the circuit of Fig. 9.73(a), assume that $(W/L)_{1-4} = 100/0.5$, $C_1 = C_2 = 0.5$ pF, and $I_{SS} = 1$ mA.

(a) Calculate the small-signal time constant of the circuit.

(b) With a 1-V step at the input [Fig. 9.73(c)], how long does it take for I_{D2} to reach $0.1 I_{SS}$?

9.15. It is possible to argue that the auxiliary amplifier in a gain-boosting stage *reduces* the output impedance. Consider the circuit as drawn in Fig. 9.89, where the drain voltage of M_2 is changed by ΔV to measure the output impedance. It seems that, since the feedback provided by A_1 attempts to hold V_X constant, the change in the current through r_{O2} is much *greater* than in the original circuit, suggesting that $R_{out} \approx r_{O2}$. Explain the flaw in this argument.

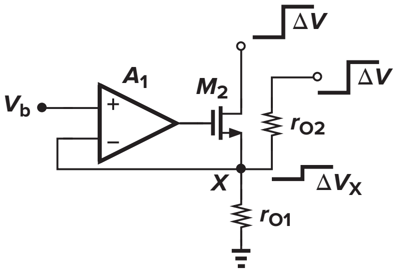


Figure 9.89

9.16. Calculate the CMRR of the circuit shown in Fig. 9.73(a).

9.17. Calculate the input-referred flicker noise of the op amp shown in Fig. 9.85(a).

9.18. In this problem, we design a two-stage op amp based on the topology shown in Fig. 9.90. Assume a power budget of 6 mW, a required output swing of 2.5 V, and $L_{eff} = 0.5\ \mu$m for all devices.

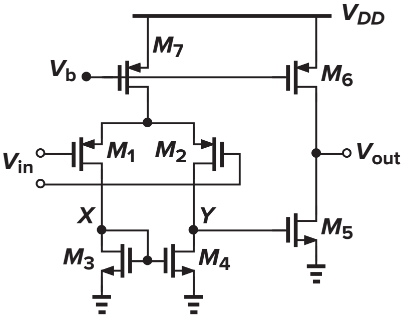


Figure 9.90

(a) Allocating a current of 1 mA to the output stage and roughly equal overdrive voltages to M_5 and M_6, determine $(W/L)_5$ and $(W/L)_6$. Note that the gate-source capacitance of M_5 is in the signal path, whereas that of M_6 is not. Thus, M_6 can be quite a lot larger than M_5.

(b) Calculate the small-signal gain of the output stage.

(c) With the remaining 1 mA flowing through M_7, determine the aspect ratio of M_3 (and M_4) such that $V_{GS3} = V_{GS5}$. This is to guarantee that if $V_{in} = 0$ and hence $V_X = V_Y$, then M_5 carries the expected current.

(d) Calculate the aspect ratios of M_1 and M_2 such that the overall voltage gain of the op amp is equal to 500.

9.19. Consider the op amp of Fig. 9.90, assuming that the second stage is to provide a voltage gain of 20 with a bias current of 1 mA.

(a) Determine $(W/L)_5$ and $(W/L)_6$ such that M_5 and M_6 have equal overdrive voltages.

(b) What is the small-signal gain of this stage if M_6 is driven into the triode region by 50 mV?

9.20. The op amp designed in Problem 9.9.18**d** is placed in unity-gain feedback. Assume that $|V_{GS7} - V_{TH7}| = 0.4$ V.

(a) What is the allowable input voltage range?

(b) At what input voltage are the input and output voltages *exactly* equal?

9.21. Calculate the input-referred noise of the op amp designed in Problem 9.9.18**d**.

9.22. It is possible to use the bulk terminal of PMOS devices as an input [10]. Consider the amplifier shown in Fig. 9.91 as an example.

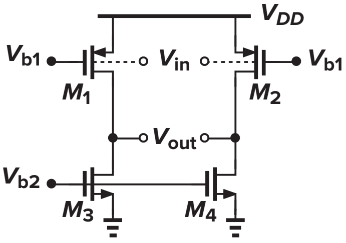


Figure 9.91

(a) Calculate the voltage gain.

(b) What is the acceptable input common-mode range?

(c) How does the small-signal gain vary with the input common-mode level?

(d) Calculate the input-referred thermal noise voltage and compare the result with that of a regular PMOS differential pair having NMOS current-source loads.

9.23. The idea of the active current mirror can be applied to the output stage of a two-stage op amp as well. That is, the load current source can become a function of the signal. Figure 9.92 shows an example [11]. Here, the first stage consists of M_1–M_4, and the output is produced by M_5–M_8. Transistors M_7 and M_8 operate as active current sources because their current varies with the signal voltage at nodes Y and X, respectively.

(a) Calculate the differential voltage gain of the op amp.

(b) Estimate the magnitude of the three major poles of the circuit.

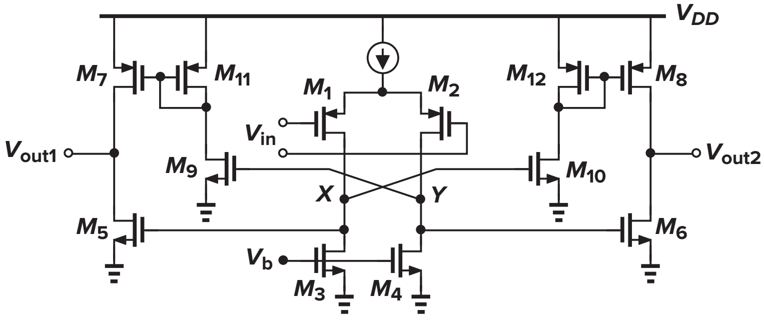


Figure 9.92

9.24. The circuit of Fig. 9.93 employs a fast path (M'_1 and M'_2) in parallel with the slow path. Calculate the differential voltage gain of the circuit. Which transistors typically limit the output swing?

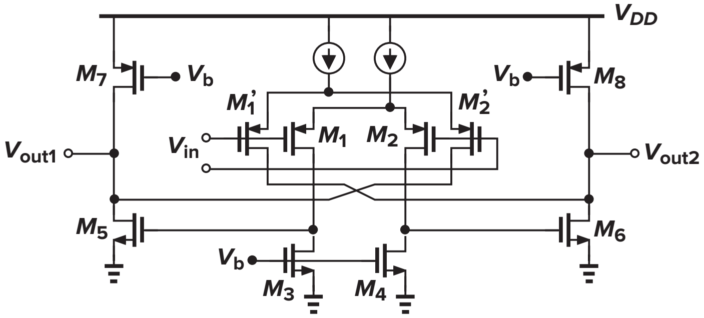


Figure 9.93

9.25. Calculate the input-referred thermal noise of the op amp in Fig. 9.93.

9.26. Determine the slew rate of a fully-differential folded-cascode op amp.

9.27. Calculate the slew rate in Fig. 9.75 if $I_{SS} > I_P$.

CHAPTER

10

Stability and Frequency Compensation

Negative feedback finds wide application in the processing of analog signals. As described in Chapter 8, feedback suppresses the effect of the variations of the open-loop characteristics. Feedback systems, however, suffer from potential instability; that is, they may oscillate.

In this chapter, we deal with the stability and frequency compensation of linear feedback systems to the extent necessary to understand the design issues of analog feedback circuits. Beginning with a review of stability criteria and the concept of phase margin, we study frequency compensation, introducing various techniques suited to different op amp topologies. We also analyze the impact of frequency compensation on the slew rate of two-stage op amps. The chapter ends with a study of Nyquist's stability criterion.

10.1 ■ General Considerations

Let us consider the negative-feedback system shown in Fig. 10.1(a), where β is assumed constant. Writing the closed-loop transfer function as

$$\frac{Y}{X}(s) = \frac{H(s)}{1+\beta H(s)} \tag{10.1}$$

we note that if $\beta H(s = j\omega_1) = -1$ at $\omega_1 \neq 0$, then the closed-loop "gain" goes to infinity, and the circuit can amplify its own noise until it eventually begins to oscillate. In other words, if the loop gain at ω_1, $\beta H(j\omega_1)$, is equal to -1, then the circuit may oscillate at frequency ω_1. This condition can be expressed as

$$|\beta H(j\omega_1)| = 1 \tag{10.2}$$

$$\angle \beta H(j\omega_1) = -180^\circ \tag{10.3}$$

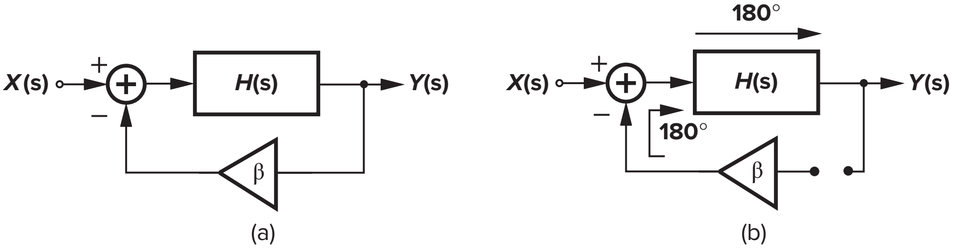


Figure 10.1 (a) Basic negative-feedback system, and (b) phase shift around the loop at ω_1.

which are called "Barkhausen's Criteria." Note that (1) these equations relate only to the loop gain (more precisely, the loop transmission)[1] and are independent of where the input and output are located, and (2) the total phase shift around the loop at ω_1 is 360° because *negative* feedback itself introduces 180° of phase shift [Fig. 10.1(b)]. The 360° phase shift is necessary for oscillation since the feedback signal must add *in phase* to the original noise to allow oscillation buildup. By the same token, a loop gain of unity (or greater) is also required to enable growth of the oscillation amplitude. These oscillation requirements are studied further in Chapter 15. The key point here is that the loop transmission, which can often be found from the open-loop system, reveals the stability of the closed-loop system.

In summary, a negative-feedback system may oscillate at ω_1 if (1) the phase shift around the loop at this frequency is so great that the feedback becomes *positive* and (2) the loop gain is still enough to allow signal buildup. Illustrated in Fig. 10.2(a), the situation can be viewed as excessive loop gain at the frequency at which the phase shift reaches $-180°$ or, equivalently, excessive phase at the frequency at which the loop gain drops to unity. Thus, to avoid instability, we must minimize the total phase shift so that if $|\beta H| = 1$, then $\angle \beta H$ is still more positive than $-180°$ [Fig. 10.2(b)]. In this chapter, we assume that β is less than or equal to unity and does not depend on the frequency.

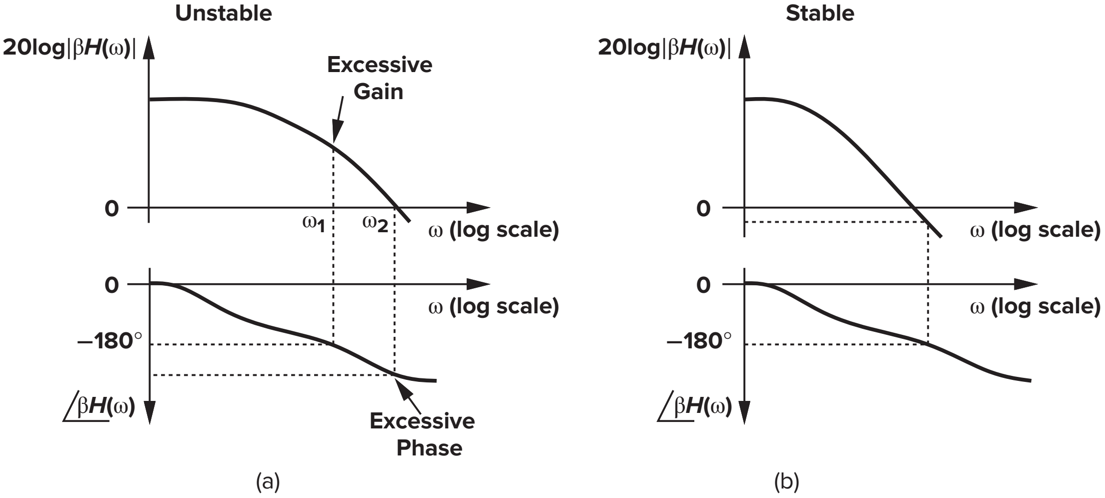


Figure 10.2 Bode plots of loop transmission for (a) unstable and (b) stable systems.

The frequencies at which the magnitude and phase of the loop gain are equal to unity and $-180°$, respectively, play a critical role in the stability and are called the "gain crossover frequency" and the "phase crossover frequency," respectively. In a stable system, the gain crossover must occur well before the phase crossover. For the sake of brevity, we denote the gain crossover by GX and the phase crossover by PX. It is helpful to note that the gain crossover frequency is the same as the unity-gain bandwidth of the loop transmission.

▶ Example 10.1

Explain whether the system depicted in Fig. 10.2(a) becomes more or less stable if the feedback is weakened, i.e., if β is reduced.

[1] The terms "loop gain" and "loop transmission" $[\beta H(s)]$, respectively, refer to the low-frequency value and the transfer function of the gain around the loop, but we sometimes use them interchangeably.

Solution

As illustrated in Fig. 10.3, a lower β shifts the plot of $20 \log |\beta H(\omega)|$ down and the GX to the left. Since $\angle \beta H(\omega)$ does not change, the system becomes *more* stable. After all, if we apply no feedback around an op amp, the circuit has no tendency to oscillate. Thus, the worst-case stability corresponds to $\beta = 1$, i.e, unity-gain feedback. For this reason, we often analyze the magnitude and phase plots for $\beta H = H$.

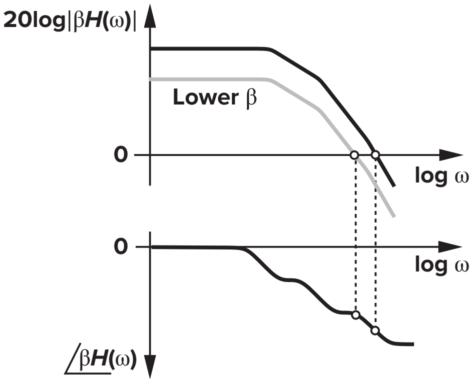


Figure 10.3

Before studying more specific cases, let us review a few basic rules for constructing Bode plots. A Bode plot illustrates the asymptotic behavior of the magnitude and phase of a complex function according to the magnitude of the poles and zeros. The following two rules are used. (1) The slope of the magnitude plot changes by +20 dB/dec at every zero frequency and by −20 dB/dec at every pole frequency. (2) For a pole (zero) frequency of ω_m, the phase begins to fall (rise) at approximately $0.1\omega_m$, experiences a change of -45° ($+45^\circ$) at ω_m, and approaches a change of -90° ($+90^\circ$) at approximately $10\omega_m$. The key point here is that the phase is much more significantly affected by high-frequency poles and zeros than the magnitude is.

It is also instructive to plot the location of the poles of a closed-loop system on a complex plane. Expressing each pole frequency as $s_p = j\omega_p + \sigma_p$ and noting that the impulse response of the system includes a term $\exp(j\omega_p + \sigma_p)t$, we observe that if s_p falls in the right half plane (RHP), i.e., if $\sigma_p > 0$, then the system oscillates because its time-domain response exhibits a growing exponential [Fig. 10.4(a)]. Even if $\sigma_p = 0$, the system sustains oscillations [Fig. 10.4(b)]. Conversely, if the poles lie in the left half plane (LHP), all time-domain exponential terms decay to zero [Fig. 10.4(c)].[2] In practice, we plot the location of the poles as the loop gain varies, thereby revealing how close to oscillation the system may come. Such a plot is called a "root locus."

We now study a feedback system incorporating a one-pole forward amplifier. Assuming $H(s) = A_0/(1 + s/\omega_0)$, we have from (10.1)

$$\frac{Y}{X}(s) = \frac{\dfrac{A_0}{1+\beta A_0}}{1 + \dfrac{s}{\omega_0(1+\beta A_0)}} \tag{10.4}$$

[2] We ignore the effect of zeros for now.

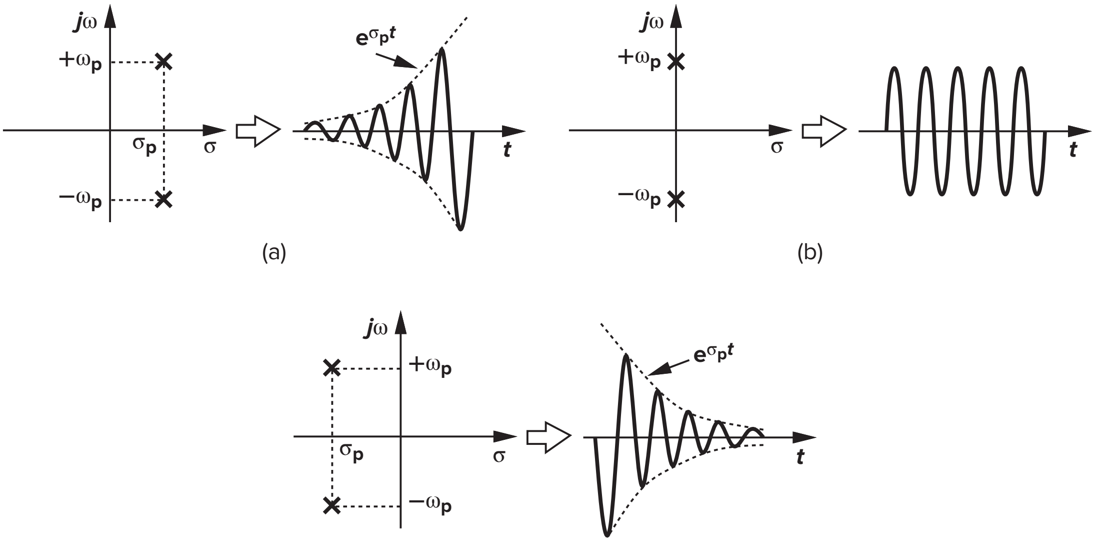


Figure 10.4 Time-domain response of a system versus the position of poles: (a) unstable with growing amplitude; (b) unstable with constant-amplitude oscillation; (c) stable.

In order to analyze the stability behavior, we plot $|\beta H(s = j\omega)|$ and $\angle\beta H(s = j\omega)$ (Fig. 10.5), observing that a single pole cannot contribute a phase shift greater than 90° and the system is unconditionally stable for all nonnegative values of β. Note that $\angle\beta H$ is independent of β.

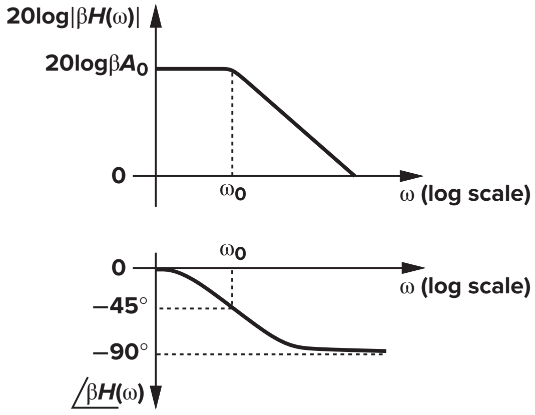


Figure 10.5 Bode plots of loop transmission for a one-pole system.

▶ Example 10.2

Construct the root locus for a one-pole system.

Solution

Equation (10.4) implies that the closed-loop system has a pole $s_p = -\omega_0(1 + \beta A_0)$, i.e., a real-valued pole in the left half plane that moves away from the origin as the loop gain increases (Fig. 10.6).

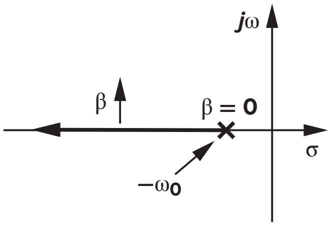


Figure 10.6

10.2 ■ Multipole Systems

Our study of op amps in Chapter 9 indicates that such circuits generally contain multiple poles. In two-stage op amps, for example, each gain stage introduces a "dominant" pole. It is therefore important to study a feedback system whose core amplifier exhibits more than one pole.

Let us consider a two-pole system first. For stability considerations, we plot $|\beta H|$ and $\angle\beta H$ as a function of the frequency. Shown in Fig. 10.7, the magnitude begins to drop at 20 dB/dec at $\omega = \omega_{p1}$ and at 40 dB/dec at $\omega = \omega_{p2}$. Also, the phase begins to change at $\omega = 0.1\omega_{p1}$, reaches $-45°$ at $\omega = \omega_{p1}$ and $-90°$ at $\omega = 10\omega_{p1}$, begins to change again at $\omega = 0.1\omega_{p2}$ (if $0.1\omega_{p2} > 10\omega_{p1}$), reaches $-135°$ at $\omega = \omega_{p2}$, and asymptotically approaches $-180°$. The system is therefore stable because $|\beta H|$ drops to below unity at a frequency where $\angle\beta H < -180°$.

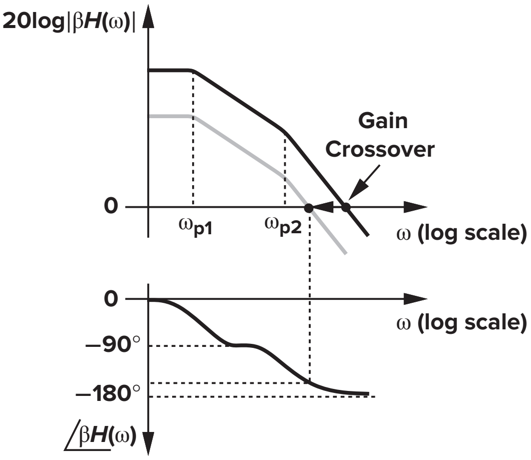


Figure 10.7 Bode plots of loop transmission for a two-pole system.

What happens if the feedback is made "weaker"? To reduce the amount of feedback, we decrease β, obtaining the gray magnitude plot in Fig. 10.7. As the feedback becomes weaker, the gain crossover point moves toward the origin while the phase crossover point remains constant, resulting in a more stable system. The stability is obtained at the cost of weaker feedback.

▶ **Example 10.3**

Construct the root locus for a two-pole system.

Solution

Writing the open-loop transfer function as

$$H(s) = \frac{A_0}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)} \tag{10.5}$$

we have

$$\frac{Y}{X}(s) = \frac{A_0}{\left(1+\frac{s}{\omega_{p1}}\right)\left(1+\frac{s}{\omega_{p2}}\right)+\beta A_0} \tag{10.6}$$

$$= \frac{A_0\omega_{p1}\omega_{p2}}{s^2+(\omega_{p1}+\omega_{p2})s+(1+\beta A_0)\omega_{p1}\omega_{p2}} \tag{10.7}$$

Thus, the closed-loop poles are given by

$$s_{1,2} = \frac{-(\omega_{p1}+\omega_{p2}) \pm \sqrt{(\omega_{p1}+\omega_{p2})^2 - 4(1+\beta A_0)\omega_{p1}\omega_{p2}}}{2} \tag{10.8}$$

As expected, for $\beta = 0$, $s_{1,2} = -\omega_{p1}, -\omega_{p2}$. As β increases, the term under the square root drops, taking on a value of zero for

$$\beta_1 = \frac{1}{A_0}\frac{(\omega_{p1}-\omega_{p2})^2}{4\omega_{p1}\omega_{p2}} \tag{10.9}$$

As shown in Fig. 10.8, the poles begin at $-\omega_{p1}$ and $-\omega_{p2}$, move toward each other, coincide for $\beta = \beta_1$, and become complex for $\beta > \beta_1$. The closed-loop system does not become unstable because the poles do not reach the $j\omega$ axis.

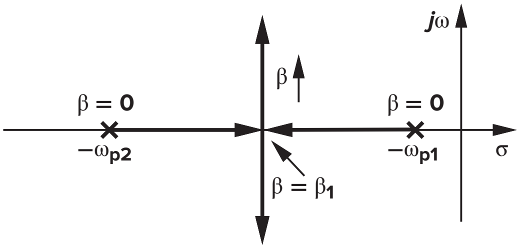


Figure 10.8

The foregoing calculations point to the complexity of the algebra required to construct a root locus for higher-order systems. For this reason, many root locus techniques have been devised so as to minimize such computations.

We now study a three-pole system. Shown in Fig. 10.9(a) are the Bode plots of the magnitude and phase of the loop gain. The third pole gives rise to additional phase shift, possibly moving the phase crossover to frequencies lower than the gain crossover and leading to oscillation.

Since the third pole also decreases the *magnitude* of the loop gain at a greater rate, the reader may wonder why the gain crossover does not move as much as the phase crossover does. As mentioned before, the phase begins to change at approximately one-tenth of the pole frequency, whereas the magnitude begins to drop only near the pole frequency. For this reason, additional poles (and zeros) affect the phase to a much greater extent than they do the magnitude.

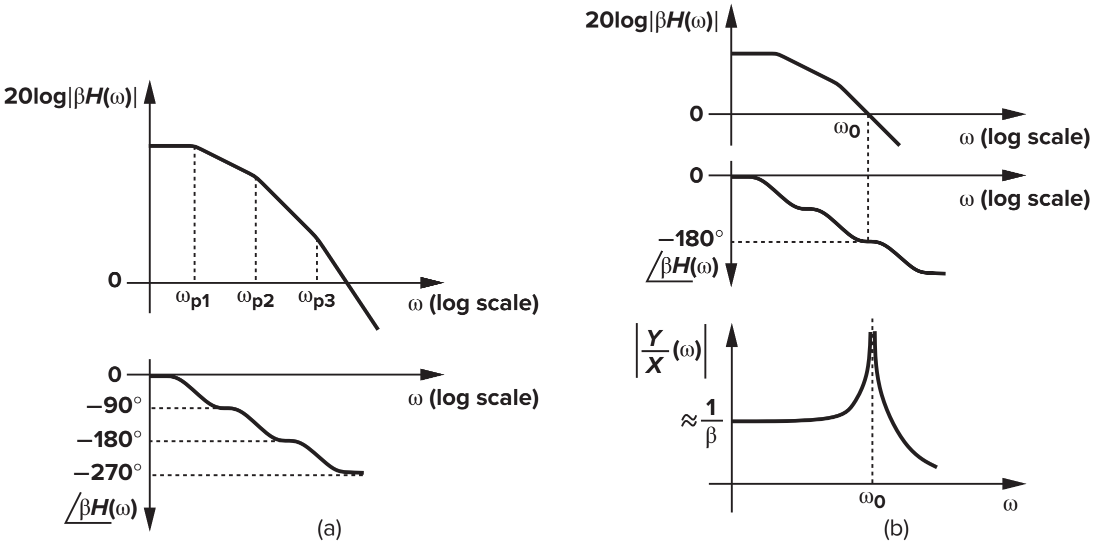


Figure 10.9 (a) Bode plots of loop transmission for a three-pole system and (b) closed-loop response.

As with a two-pole system, if the feedback factor in Fig. 10.9 decreases, the circuit becomes more stable because the gain crossover moves toward the origin while the phase crossover remains constant. For this reason, a feedback amplifier designed for a *higher* closed-loop gain tends to be more stable (why?).

It is important not to confuse the βH plots with the *closed-loop* frequency response, Y/X. As an example, consider a system with the loop response shown in Fig. 10.9(b), where the gain and phase crossover frequencies coincide. The closed-loop response, $|Y/X|$, exhibits infinite gain at ω_0, predicting oscillation at this frequency.

10.3 ■ Phase Margin

We have seen that to ensure stability, $|\beta H|$ must drop to unity before $\angle\beta H$ crosses $-180°$. We may naturally ask: How far should PX be from GX? Let us first consider a "marginal" case where, as depicted in Fig. 10.10(a), GX is only slightly below PX; for example, at GX, the phase equals $-175°$. How does the closed-loop system respond in this case? Noting that at GX, $\beta H(j\omega_1) = 1 \times \exp(-j175°)$, we have for the closed-loop system

$$\frac{Y}{X}(j\omega_1) = \frac{H(j\omega_1)}{1+\beta H(j\omega_1)} \tag{10.10}$$

$$= \frac{\dfrac{1}{\beta}\exp(-j175°)}{1+\exp(-j175°)} \tag{10.11}$$

$$= \frac{1}{\beta}\cdot\frac{-0.9962 - j0.0872}{0.0038 - j0.0872} \tag{10.12}$$

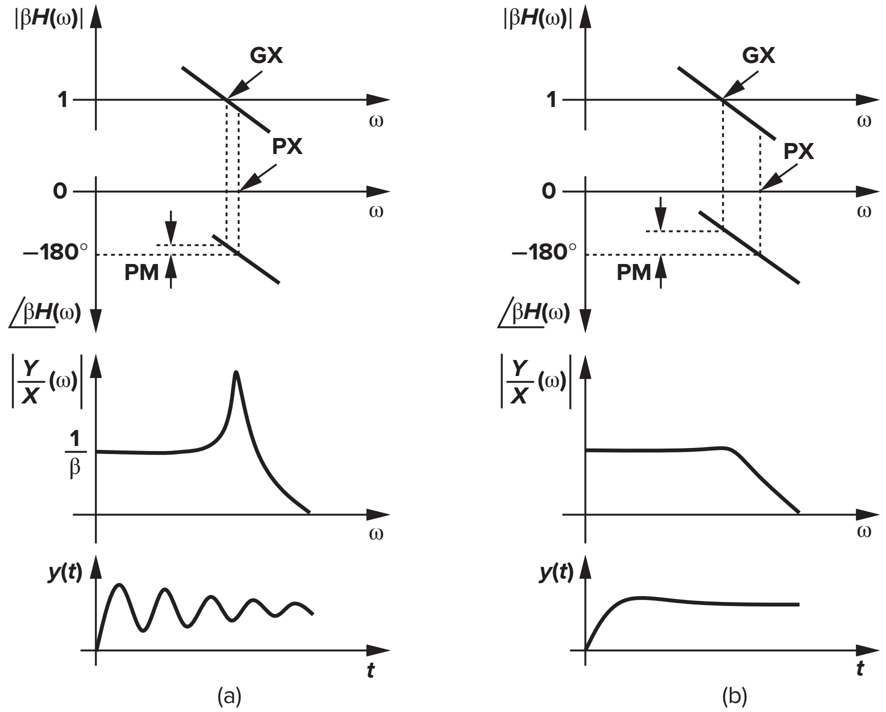


Figure 10.10 Closed-loop frequency and time response for (a) small and (b) large margin between gain and phase crossover points.

and hence

$$\left|\frac{Y}{X}(j\omega_1)\right| = \frac{1}{\beta}\cdot\frac{1}{0.0872} \tag{10.13}$$

$$\approx \frac{11.5}{\beta} \tag{10.14}$$

Since at low frequencies, $|Y/X| \approx 1/\beta$, the closed-loop frequency response exhibits a sharp peak in the vicinity of $\omega = \omega_1$. In other words, the closed-loop system is near oscillation, and its step response, $y(t)$, exhibits a very underdamped behavior. This point also reveals that a second-order system may suffer from ringing although it is stable.

Now suppose, as shown in Fig. 10.10(b), GX precedes PX by a greater margin. Then, we expect a relatively "well-behaved" closed-loop response in both the frequency domain and the time domain. It is therefore plausible to conclude that the greater the spacing between GX and PX (while GX remains below PX), the more stable the feedback system. Alternatively, the phase of βH at the gain crossover frequency can serve as a measure of stability: the smaller $|\angle\beta H|$ at this point, the more stable the system.

This observation leads us to the concept of "phase margin" (PM), defined as PM $= 180^\circ + \angle\beta H(\omega = \omega_1)$, where ω_1 is the gain crossover frequency. We see that stability calls for a positive and large PM.

▶ **Example 10.4**

A two-pole feedback system is designed such that $|\beta H(\omega_{p2})| = 1$ and $|\omega_{p1}| \ll |\omega_{p2}|$ (Fig. 10.11). How much is the phase margin?

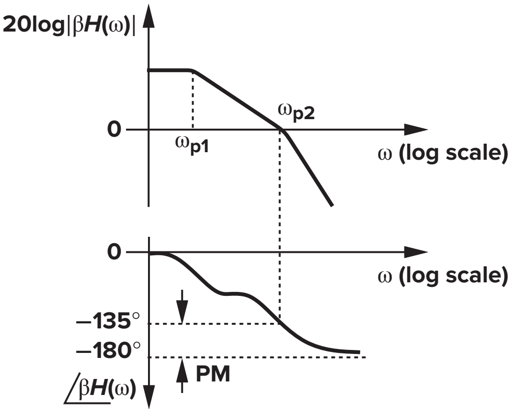


Figure 10.11

Solution

Since $\angle \beta H$ reaches -135° at $\omega = \omega_{p2}$, the phase margin is equal to 45°. The key point to remember is that, if the loop gain drops to unity at a frequency *above* the second pole, the phase margin is *less than* 45°. As explained below, since PM $= 45^\circ$ is typically inadequate, we say that the ultimate unity-gain bandwidth cannot exceed the second pole of the open-loop op amp if a well-behaved time response is desired.

The above example suggests that for a phase margin greater than 45°, the gain crossover frequency must lie between the first pole and the second (in the absence of zeros). That is, the unity-gain bandwidth cannot exceed the second pole frequency.

How much phase margin is adequate? It is instructive to examine the closed-loop frequency response for different phase margins [1]. For PM $= 45^\circ$, at the gain crossover frequency $\angle \beta H(\omega_1) = -135^\circ$ and $|\beta H(\omega_1)| = 1$ (Fig. 10.12), yielding

$$\frac{Y}{X} = \frac{H(j\omega_1)}{1 + 1 \times \exp(-j135^\circ)} \tag{10.15}$$

$$= \frac{H(j\omega_1)}{0.29 - 0.71j} \tag{10.16}$$

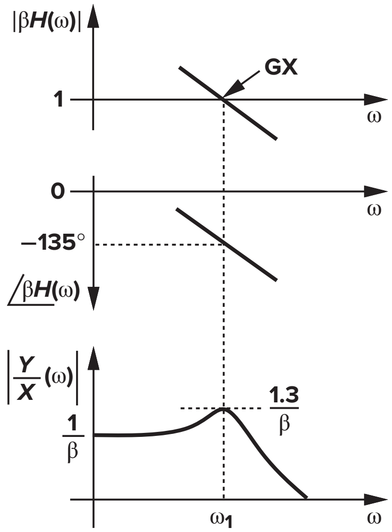


Figure 10.12 Closed-loop frequency response for 45° phase margin.

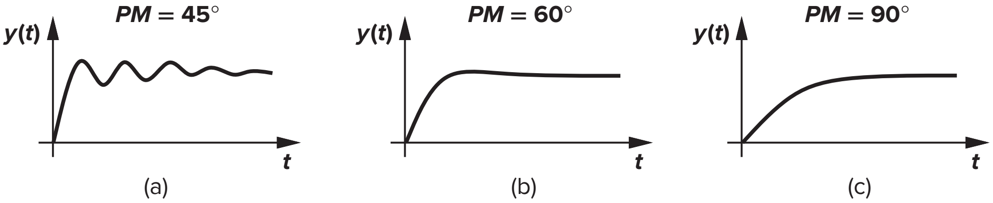
It follows that

$$\left|\frac{Y}{X}\right| = \frac{1}{\beta} \cdot \frac{1}{|0.29 - 0.71j|} \tag{10.17}$$

$$\approx \frac{1.3}{\beta}. \tag{10.18}$$

Consequently, the frequency response of the feedback system suffers from a 30% peak at $\omega = \omega_1$.

It can be shown that for PM $= 60^\circ$, $Y(j\omega_1)/X(j\omega_1) = 1/\beta$, suggesting a negligible frequency peaking. This typically means that the step response of the feedback system exhibits little ringing, providing a fast settling. For greater phase margins, the system is more stable, but the time response slows down (Fig. 10.13). Thus, PM $= 60^\circ$ is typically considered the optimum value.



(a) (b) (c)

Figure 10.13 Closed-loop time response for 45°, 60°, and 90° phase margins.

The concept of phase margin is well suited to the design of circuits that process *small* signals. In practice, the large-signal step response of feedback amplifiers does not follow the illustration of Fig. 10.13. This is not only due to slewing but also because of the nonlinear behavior resulting from large excursions in the bias voltages and currents of the amplifier. Such excursions in fact cause the pole and zero frequencies to *vary* during the transient, leading to a complicated time response. Thus, for large-signal applications, time-domain simulations of the closed-loop system prove more relevant and useful than small-signal ac computations of the open-loop amplifier.

As an example of a feedback circuit exhibiting a reasonable phase margin but poor settling behavior, consider the unity-gain amplifier of Fig. 10.14, where the aspect ratio of all transistors is equal to 50 μm / 0.6 μm. With the choice of the device dimensions, bias currents, and capacitor values shown here, SPICE yields a phase margin of approximately 65° and a unity-gain frequency of 150 MHz. The large-signal step response, however, suffers from significant ringing.

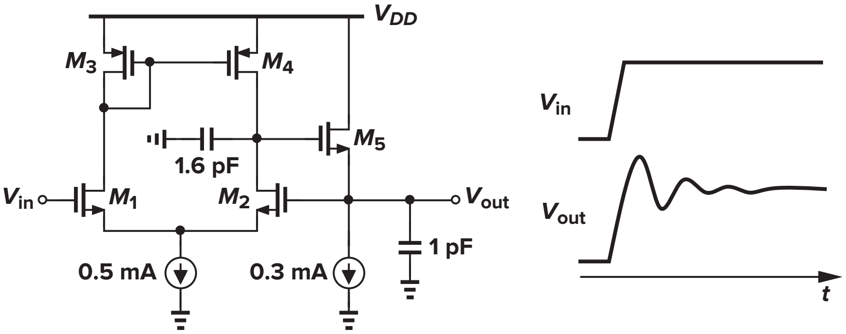


Figure 10.14 Unity-gain buffer.

10.4 ■ Basic Frequency Compensation

Typical op amp circuits contain many poles. In a folded-cascode topology, for example, both the folding node and the output node contribute poles. For this reason, op amps must usually be "compensated," that is, their open-loop transfer function must be modified such that the closed-loop circuit is stable and the time response is well behaved.

The need for compensation arises because $|\beta H|$ does not drop to unity well before $\angle\beta H$ reaches $-180°$. We then postulate that stability can be achieved by (1) minimizing the overall phase shift, thus pushing the phase crossover *out* [Fig. 10.15(a)]; or (2) dropping the gain with frequency, thereby pushing the gain crossover *in* [Fig. 10.15(b)]. The first approach requires that we attempt to minimize the number of poles in the signal path by proper design. Since each additional stage contributes at least one pole, this means that the number of stages must be minimized, a remedy that yields low voltage gain and/or limited output swings (Chapter 9). The second approach, on the other hand, retains the low-frequency gain and the output swings, but it reduces the bandwidth by forcing the gain to fall at lower frequencies.

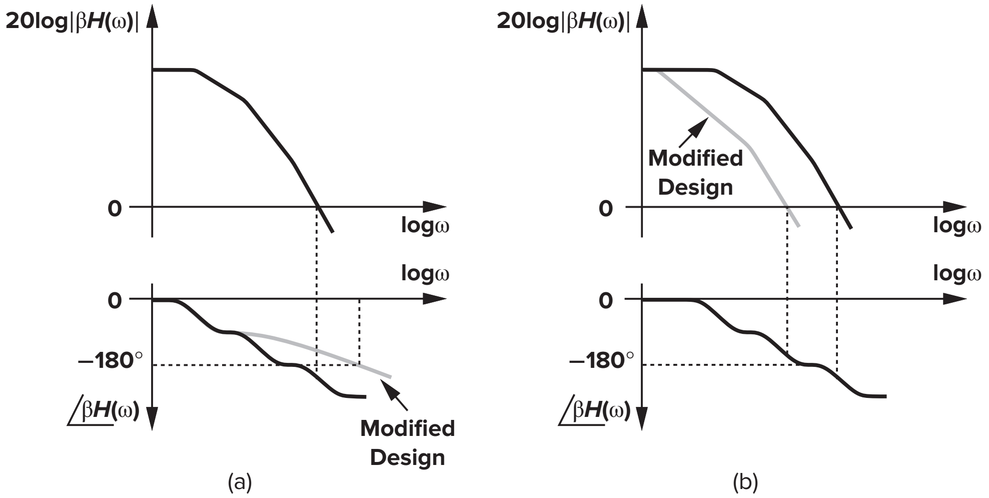


Figure 10.15 Frequency compensation by (a) moving PX out and (b) pushing GX in.

In practice, we first try to design an op amp so as to minimize the number of poles while meeting other requirements. Since the resulting circuit may still suffer from insufficient phase margin, we then compensate the op amp, i.e., modify the design so as to move the gain crossover toward the origin. These efforts proceed with the β value chosen according to the final design requirements. For example, a closed-loop gain of 4 in some cases translates to $\beta \approx 0.25$ if the loop gain is large.[3] In other words, we need not compensate the circuit for $\beta = 1$ if the closed-loop gain is always higher.

Let us apply the above concepts to the telescopic-cascode op amp shown in Fig. 10.16, where a PMOS current mirror performs differential to single-ended conversion. We identify a number of poles in the signal paths: path 1 contains a high-frequency pole at the source of M_3, a mirror pole at node A, and another high-frequency pole at the source of M_7, whereas path 2 contains a high-frequency pole at the source of M_4. The two paths share a pole at the output.

It is instructive to estimate the relative position of these poles. Since the output resistance of the op amp is much higher than the small-signal resistances seen at the other nodes in the circuit, we expect

[3] But in "switched-capacitor" circuits, the closed-loop gain changes from one mode to another (Chapter 13).

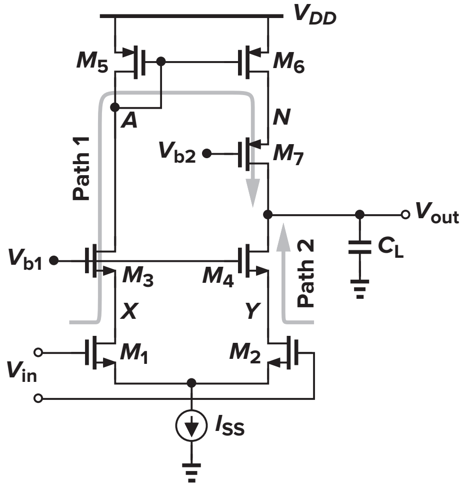


Figure 10.16 Telescopic op amp with single-ended output.

that, even with a moderate load capacitance, the output pole, $\omega_{p,out}$, is the closest to the origin. Called the "dominant pole," $\omega_{p,out}$ usually sets the open-loop 3-dB bandwidth.

We also surmise that the first "nondominant pole," i.e., the closest pole to the origin after the dominant pole, arises at node A. This is because the total capacitance at this node, roughly equal to $C_{GS5} + C_{GS6} + C_{DB5} + 2C_{GD6} + C_{DB3} + C_{GD3}$, is typically quite a lot larger than that at nodes X, Y, and N, and the small-signal resistance of M_5, approximately $1/g_{m5}$, is also relatively large.

Which node yields the next nondominant pole: N or X (and Y)? Recall from Chapter 9 that, to obtain a low overdrive and consume a reasonable voltage headroom, the PMOS devices in the op amp are typically wider than the NMOS transistors. Comparing M_4 and M_7 and neglecting body effect, we note that since $g_m = 2I_D/|V_{GS} - V_{TH}|$, if the two transistors are designed to have the same overdrive, they also exhibit the same transconductance. However, from square-law characteristics, we have $W_4/W_7 = \mu_p/\mu_n$, which is about 1/2 to 1/3. Thus, nodes N and X (or Y) see roughly equal small-signal resistances to ground, but node N suffers from much more capacitance. It is therefore plausible to assume that node N contributes the next nondominant pole. Figure 10.17 illustrates the results, denoting the capacitance at nodes A, N, and X by C_A, C_N, and C_X, respectively. The poles at nodes X and Y are nearly equal, and their corresponding terms in the transfer functions of path 1 and path 2 can be factored out. Thus, they count as one pole rather than two.

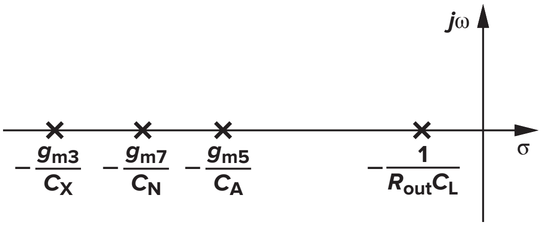


Figure 10.17 Pole locations for the op amp of Fig. 10.16.

With the position of the poles roughly determined, we can construct the magnitude and phase plots for βH, using $\beta = 1$ for the worst case. Shown in Fig. 10.18, such characteristics indicate that the mirror pole typically limits the phase margin because its phase contribution occurs at lower frequencies than that of other nondominant poles.

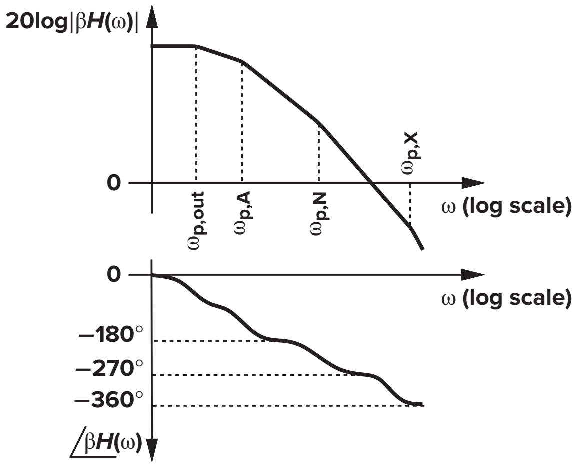


Figure 10.18 Bode plots of loop transmission for op amp of Fig. 10.16.

Recall from Chapter 6 that differential pairs using active current mirrors exhibit a left-half-plane zero located at twice the mirror pole frequency. The circuit of Fig. 10.16 contains such a zero as well. Located at $2\omega_{p,A}$, the zero impacts both the magnitude and phase characteristics. The analysis is left to the reader.

Compensation Procedure How should we compensate the telescopic-cascode op amp? Recall that our ultimate goal is to ensure a loop gain sufficiently less than unity at the phase crossover frequency. Let us assume that the number and location of the nondominant poles and hence the phase plot at frequencies higher than roughly $10\omega_{p,out}$, remain constant. We begin with the original response shown in Fig. 10.19, which has a negative phase margin. We must force the loop gain to drop such that the gain crossover point moves toward the origin. To accomplish this, we simply lower the frequency of the dominant pole, ω_{p1}, by increasing the load capacitance. The key point is that the phase contribution of the dominant pole in the vicinity of the gain or phase crossover point is close to 90° and relatively independent of the location of the pole. That is, as illustrated in Fig. 10.19, translating the dominant pole toward the origin affects the magnitude plot, but not the critical part of the phase plot. If ω_{p1} is lowered sufficiently, the PM reaches an acceptable value, but at the cost of bandwidth.

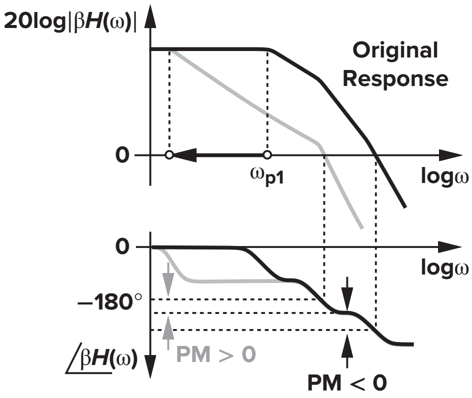


Figure 10.19 Translating the dominant pole toward the origin.

In order to determine how much the dominant pole must be shifted down as well as arrive at an important conclusion, let us assume that (1) the second nondominant pole ($\omega_{p,N}$) in Fig. 10.16 is much higher than the mirror pole so that the phase shift at $\omega = \omega_{p,A}$ is equal to $-135°$, and (2) a phase margin of 45° (which is usually inadequate) is necessary. To compensate the circuit, we begin from $\angle\beta H(\omega) = -180^\circ + \text{PM} = -135^\circ$ and identify the corresponding gain crossover frequency, in this case, $\omega_{p,A}$ (Fig. 10.20). Since the dominant pole must drop the gain to unity at $\omega_{p,A}$ with a slope of 20 dB/dec, we draw a straight line from $\omega_{p,A}$ toward the origin with such a slope, thus obtaining the new magnitude of the dominant pole, $\omega'_{p,out}$. Therefore, the load capacitance must be increased by a factor of $\omega_{p,out}/\omega'_{p,out}$.

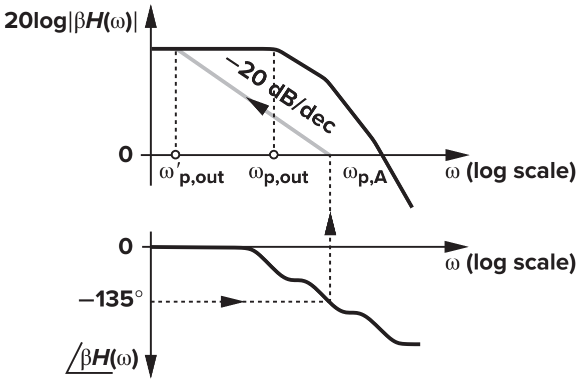


Figure 10.20 Translating the dominant pole toward the origin for 45° phase margin.

From the new magnitude plot, we note that the unity-gain bandwidth of the compensated (open-loop) op amp is equal to the *frequency of the first nondominant pole* (of course with a phase margin of 45°). This is a fundamental result, indicating that to achieve a wide bandwidth in a feedback system employing a multipole op amp, the first *nondominant* pole must be as far as possible. For this reason, the mirror pole proves undesirable.

We should also mention that although $\omega_{p,out} = (R_{out}C_L)^{-1}$, increasing R_{out} does *not* compensate the op amp. As shown in Fig. 10.21, a higher R_{out} results in a greater low-frequency loop gain, only affecting the low-frequency portion of the characteristics. Also, moving one of the nondominant poles toward the origin does not improve the phase margin. (Why?)

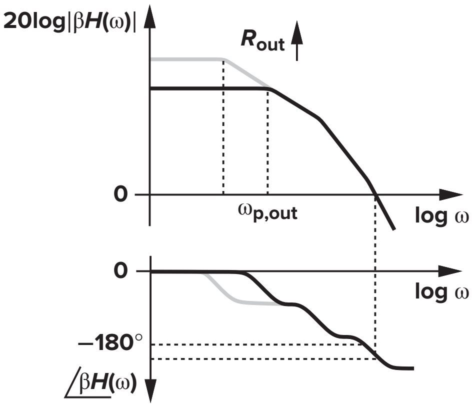


Figure 10.21 Bode plots of loop gain for higher output resistance.

In summary, frequency compensation moves the dominant pole of the open-loop amplifier to sufficiently low values so that the unity-gain bandwidth is well below the phase crossover frequency. Also, the compensated bandwidth cannot exceed the first nondominant pole frequency since a phase margin of greater than 45° is typically required.

▶ **Example 10.5**

An op amp is compensated to have a phase margin of 60° with unity-gain feedback. By what factor can the compensation be relaxed if the circuit is to operate with a feedback factor of $\beta < 1$ [Fig. 10.22(a)]?

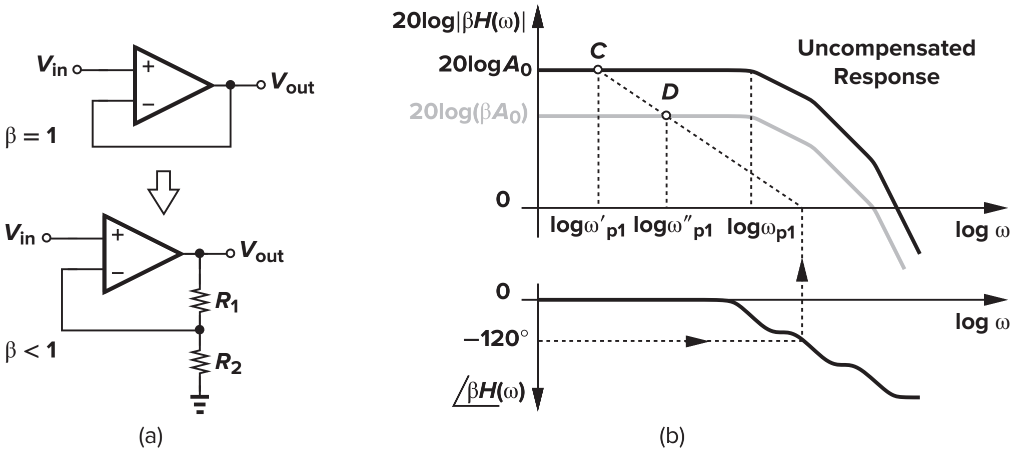


Figure 10.22

Solution

As illustrated in Fig. 10.22(b), the original compensation identifies the frequency at which $\angle\beta H = -120°$, draws a line from this frequency at a slope of 20 dB/dec toward the vertical axis, and hence moves the dominant pole from ω_{p1} to ω'_{p1}. With a feedback factor of β, the uncompensated magnitude response is shifted down by $-20\log\beta$, requiring a dominant pole at ω''_{p1}. To obtain this value, we equate the slope of the line CD to 20 dB/dec:

$$\frac{-20\log\beta}{\log\omega''_{p1} - \log\omega'_{p1}} = 20 \tag{10.19}$$

and hence $\omega''_{p1} = \omega'_{p1}/\beta$. That is, the compensation capacitor can be reduced by approximately a factor of $1/\beta$. This, of course, does *not* mean that the new feedback circuit settles faster; the weaker feedback translates to a proportionally smaller extension of the bandwidth. In fact, we can write the closed-loop −3-dB bandwidths as $(1+A_0)\omega'_{p1} \approx A_0\omega'_{p1}$ for the original op amp and $(1+\beta A_0)\omega''_{p1} \approx \beta A_0\omega''_{p1} \approx A_0\omega'_{p1}$ for the newly-compensated counterpart, concluding that the closed-loop speed remains roughly the same.

A related question that we address in Problem 10.23 is the following: If an op amp is compensated to have PM = 60° with unity-gain feedback, by how much does its PM increase if the feedback factor is reduced to $\beta < 1$?

◀

Now consider the fully differential telescopic cascode depicted in Fig. 10.23. In addition to achieving various useful properties of differential operation, this topology avoids the mirror pole, thereby exhibiting stable behavior for a greater bandwidth. In fact, we can identify one dominant pole at each output node and only *one* nondominant pole arising from node X (or Y). This suggests that fully differential telescopic-cascode circuits are stable and do not need compensation.

But how about the pole at node N (or K) in Fig. 10.23? Considering one of the PMOS cascodes as shown in Fig. 10.24(a), we may think that the capacitance at node N, $C_N = C_{GS5} + C_{SB5} + C_{GD7} + C_{DB7}$, shunts the output resistance of M_7 at high frequencies, thereby dropping the output impedance of the cascode. To quantify this effect, we first determine Z_{out} in Fig. 10.24(a):

$$Z_{out} = (1 + g_{m5}r_{O5})Z_N + r_{O5} \tag{10.20}$$

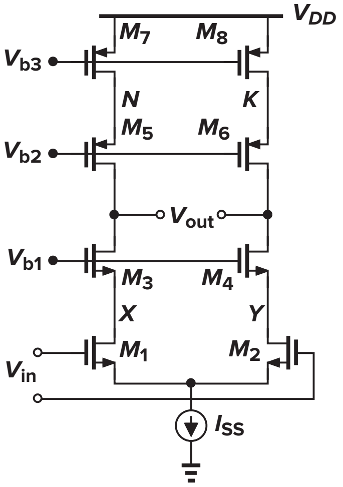


Figure 10.23 Fully differential telescopic op amp.

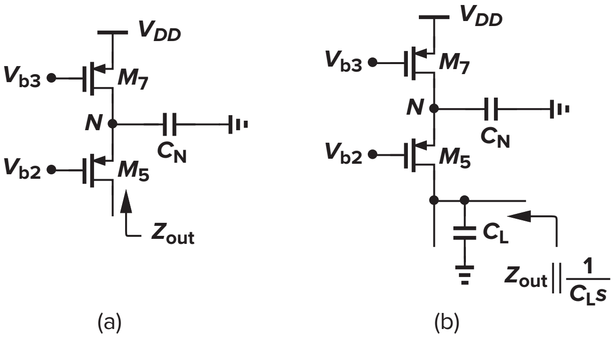


Figure 10.24 Effect of device capacitance at internal node of a cascode current source.

where body effect is neglected and $Z_N = r_{O7}||(C_N s)^{-1}$. Assuming the first term is much greater than the second, we have

$$Z_{out} \approx (1 + g_{m5}r_{O5})\frac{r_{O7}}{r_{O7}C_N s + 1} \tag{10.21}$$

Now, as illustrated in Fig. 10.24(b), we take the output load capacitance into account:

$$Z_{out}||\frac{1}{C_L s} = \frac{(1 + g_{m5}r_{O5})\dfrac{r_{O7}}{r_{O7}C_N s + 1}\cdot\dfrac{1}{C_L s}}{(1 + g_{m5}r_{O5})\dfrac{r_{O7}}{r_{O7}C_N s + 1} + \dfrac{1}{C_L s}} \tag{10.22}$$

$$= \frac{(1 + g_{m5}r_{O5})r_{O7}}{[(1 + g_{m5}r_{O5})r_{O7}C_L + r_{O7}C_N]s + 1} \tag{10.23}$$

Thus, the parallel combination of Z_{out} and the load capacitance still contains a single pole corresponding to a time constant $(1 + g_{m5}r_{O5})r_{O7}C_L + r_{O7}C_N$. Note that $(1 + g_{m5}r_{O5})r_{O7}C_L$ is simply due to the low-frequency output resistance of the cascode. In other words, the overall time constant equals the "output" time constant plus $r_{O7}C_N$. The key point in this calculation is that the pole in the PMOS cascode (at node N) is *merged* with the output pole, thus creating no *additional* pole. It merely lowers the dominant pole by a slight amount. For this reason, we loosely say that the signal does not "see" the pole in the cascode current sources.[4]

Comparison of the circuits shown in Figs. 10.16 and 10.23 now reveals that the fully differential configuration avoids both the mirror pole *and* the pole at node N. With the approximation made in (10.23), the circuit of Fig. 10.23 contains only one nondominant pole located at relatively high frequencies owing to the high transconductance of the NMOS transistors. This is a remarkable advantage of fully differential cascode op amps.

We have thus far observed that nondominant poles give rise to instability, requiring frequency compensation. Is it possible to cancel one or more of these poles by introducing *zeros* in the transfer function? For example, following the analysis of Fig. 6.41, we surmise that if a low-gain but fast path is placed in parallel with the main amplifier, a zero is created that can be positioned atop the first nondominant pole. However, cancellation of a pole by a zero in the presence of mismatches leads to long settling components in the step response of the closed-loop circuit. This effect is studied in Problem 10.19.

10.5 ■ Compensation of Two-Stage Op Amps

Our study of op amps in Chapter 9 indicates that two-stage topologies may prove inevitable if the output voltage swing must be maximized. This is especially true in today's low-voltage op amps. Thus, the stability and compensation of such op amps is of interest.

Consider the circuit shown in Fig. 10.25. We identify three poles: a pole at X (or Y), another at E (or F), and a third at A (or B). From our foregoing discussions, we know that the pole at X lies at relatively high frequencies. But how about the other two? Since node E exhibits a high small-signal resistance, even the capacitances of M_3, M_5, and M_9 can create a pole relatively close to the origin. At node A, the small-signal resistance is lower, but C_L may be large. Consequently, we say that the circuit contains *two* dominant poles.

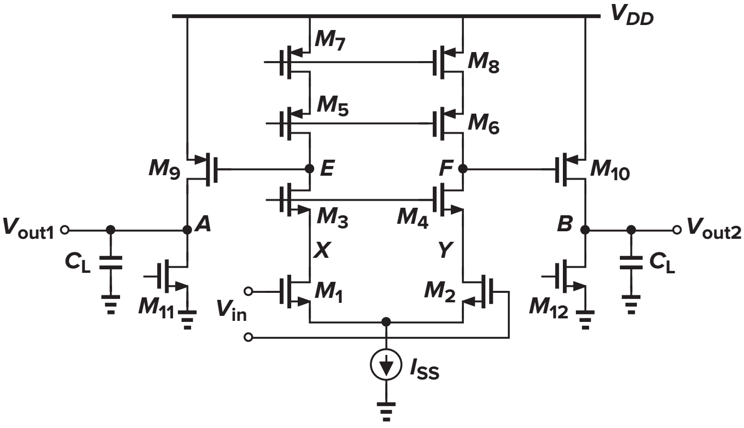


Figure 10.25 Two-stage op amp.

From these observations, we can construct the magnitude and phase plots shown in Fig. 10.26. Here, $\omega_{p,E}$ is assumed more dominant, but the relative positions of $\omega_{p,E}$ and $\omega_{p,A}$ depend on the design and the load capacitance. Note that, since the poles at E and A are relatively close to the origin, the phase

[4] If the second term in Eq. (10.20) is included in subsequent derivations, a pole and a zero that are nearly equal appear in the overall output impedance. Nonetheless, for $g_m r_O \gg 1$ and $C_L > C_N$, their effect is negligible.

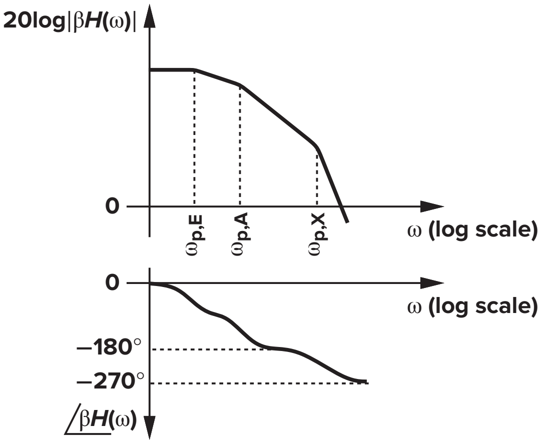


Figure 10.26 Bode plots of loop gain of two-stage op amp.

approaches −180° well below the third pole. In other words, the phase margin may be close to zero even before the third pole contributes significant phase shift.

Let us now investigate the frequency compensation of two-stage op amps. In Fig. 10.26, one of the dominant poles must be moved toward the origin so as to place the gain crossover well below the phase crossover. However, recall from Sec. 10.4 that the unity-gain bandwidth after compensation cannot exceed the frequency of the second pole of the open-loop system for PM > 45°. Thus, if in Fig. 10.26 the magnitude of $\omega_{p,E}$ is to be reduced, the available bandwidth is limited to approximately $\omega_{p,A}$, a low value. Furthermore, the very small magnitude of the new dominant pole translates to a large compensation capacitor.

Fortunately, a more efficient method of compensation can be applied to the circuit of Fig. 10.25. To arrive at this method, we note that, as illustrated in Fig. 10.27(a), the first stage exhibits a high output impedance, R_{out1}, and the second stage provides a moderate gain, A_{v2}, thereby creating a suitable environment for Miller multiplication of capacitors. Shown in Fig. 10.27(b), the idea is to create a large capacitance at node E, equal to $(1+A_{v2})C_C$, moving the corresponding pole to $R_{out1}^{-1}[C_E+(1+A_{v2})C_C]^{-1}$, where C_E denotes the capacitance at node E before C_C is added. As a result, a low-frequency pole can be established with a moderate capacitor value, saving considerable chip area. This technique is called “Miller compensation.”

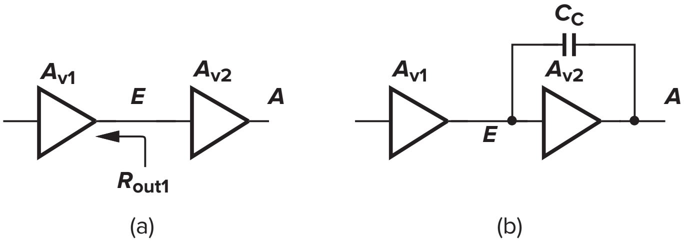


Figure 10.27 Miller compensation of a two-stage op amp.

In addition to lowering the required capacitor value, Miller compensation entails a very important property: it moves the *output* pole *away* from the origin. Illustrated in Fig. 10.28, this effect is called “pole splitting.” To understand the underlying principle, we simplify the output stage of Fig. 10.25 as in Fig. 10.29, where R_S denotes the output resistance of the first stage and $R_L = r_{O9}||r_{O11}$. From our

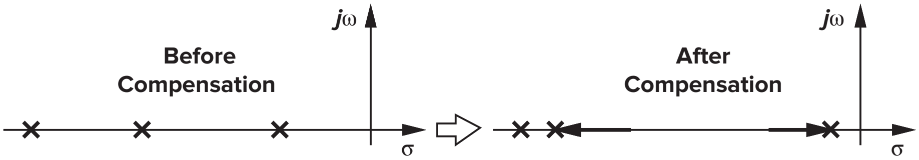


Figure 10.28 Pole splitting as a result of Miller compensation.

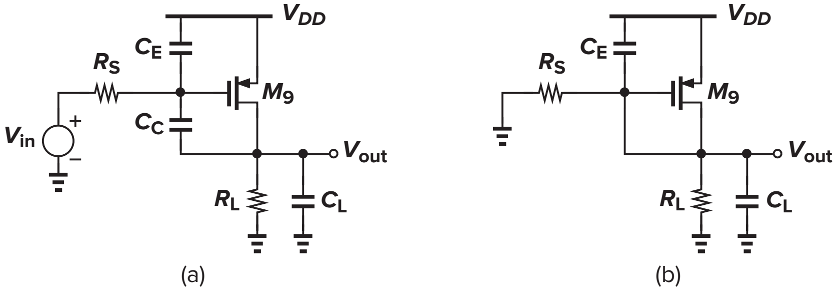


Figure 10.29 (a) Simplified circuit of a two-stage op amp, and (b) a rough model at high frequencies.

analysis in Chapter 6, we note that this compensated circuit contains two poles:

$$\omega'_{p1} \approx \frac{1}{R_S[(1+g_{m9}R_L)(C_C+C_{GD9})+C_E]+R_L(C_C+C_{GD9}+C_L)} \tag{10.24}$$

$$\omega'_{p2} \approx \frac{R_S[(1+g_{m9}R_L)(C_C+C_{GD9})+C_E]+R_L(C_C+C_{GD9}+C_L)}{R_SR_L[(C_C+C_{GD9})C_E+(C_C+C_{GD9})C_L+C_EC_L)]} \tag{10.25}$$

These expressions are based on the assumption that $|\omega'_{p1}| \ll |\omega'_{p2}|$. Before compensation, however, ω_{p1} and ω_{p2} are of the same order of magnitude. For $C_C = 0$ and relatively large C_L, we may approximate the magnitude of the output pole as $\omega_{p2} \approx 1/(R_LC_L)$.

To compare the magnitudes of ω'_{p2} before and after compensation, we consider a typical case: $C_C + C_{GD9} \gg C_E$, reducing (10.25) to $\omega'_{p2} \approx g_{m9}/(C_E + C_L)$. Noting that typically $C_E \ll C_L$, we conclude that Miller compensation increases the magnitude of the output pole by roughly a factor of $g_{m9}R_L$, a relatively large value. Intuitively, this is because at high frequencies, C_C provides a low impedance between the gate and drain of M_9, reducing the resistance seen by C_L from R_L to roughly $R_S||g_{m9}^{-1}||R_L \approx g_{m9}^{-1}$ [Fig. 10.29(b)]. From another perspective, C_C provides feedback around the second stage by sensing the output voltage; as a result, the output resistance falls and the second pole moves to higher frequencies.[5]

In summary, Miller compensation moves the interstage pole toward the origin and the output pole away from the origin, allowing a much greater bandwidth than that obtained by merely connecting the compensation capacitor from one node to ground. In practice, the choice of the compensation capacitor for proper phase margin requires some iteration because both poles move. The following example gives a rough estimate.

▶ **Example 10.6**

The two-stage op amp of Fig. 10.25 incorporates Miller compensation to reach a phase margin of 45°. Estimate the compensation capacitor value.

[5] This capacitor returns a current to the input of the second stage, thus lowering its input impedance as well.

Solution

After frequency compensation, the dominant pole moves down to approximately $(g_{m9}R_LC_CR_S)^{-1}$, where R_S denotes the output resistance of the first stage, and the second pole moves up to roughly g_{m9}/C_L. For a phase margin of 45°, the loop gain must drop to unity at the second pole. With a low-frequency loop gain of $\beta g_{m1}R_Sg_{m9}R_L$, we consider the postcompensation plot in Fig. 10.30 (on linear axes) and write

$$|\beta H(\omega)| \approx \frac{\beta g_{m1}R_Sg_{m9}R_L}{\sqrt{1+\omega^2/\omega'^2_{p1}}} \tag{10.26}$$

where the effect of ω'_{p2} on the magnitude is neglected. At $\omega = \omega'_{p2}$, the second term under the square root dominates, and

$$\frac{\beta g_{m1}R_Sg_{m9}R_L}{\omega'_{p2}/\omega'_{p1}} = 1 \tag{10.27}$$

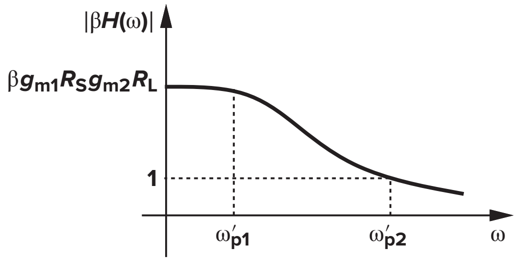


Figure 10.30

Substituting for the pole frequencies and assuming that $\beta = 1$, we obtain

$$C_C = \frac{g_{m1}}{g_{m9}}C_L \tag{10.28}$$

Note that g_{m1} and g_{m9} are the transconductances of the two stages. The reader can prove that, if the effect of ω'_{p2} is included, then $C_C = [g_{m1}/(\sqrt{2}g_{m9})]C_L$. Of course, C_C must generally be greater than this value so as to establish a higher phase margin, but this estimate serves as a reasonable starting point in the design.

This result assumes that $\beta = 1$; in practice, most op amps are configured for a closed-loop gain of 2 or higher, thus requiring a smaller C_C.

Our study of stability and compensation has thus far neglected the effect of *zeros* of the transfer function. While in cascode topologies, the zeros are far from the origin, in two-stage op amps incorporating Miller compensation, a nearby zero appears in the circuit. Recall from Chapter 6 that the circuit of Fig. 10.29 contains a right-half-plane zero at $\omega_z = g_{m9}/(C_C + C_{GD9})$. This is because $C_C + C_{GD9}$ forms a "feedforward" signal path from the input to the output. What is the effect of such a zero? The numerator of the transfer function reads $(1 - s/\omega_z)$, yielding a phase of $-\tan^{-1}(\omega/\omega_z)$, a negative value because ω_z is positive. In other words, as with poles in the left half plane, a zero in the right half plane contributes additional negative phase shift, thus moving the phase crossover toward the origin. Furthermore, from Bode approximations, the zero slows down the drop of the magnitude, thereby pushing the gain crossover away from the origin. As a result, the stability degrades considerably.

To better understand the foregoing discussion, let us construct the Bode plots for a third-order system containing a dominant pole ω_{p1}, two nondominant poles ω_{p2} and ω_{p3}, and a zero in the right half plane ω_z. For two-stage op amps, typically $|\omega_{p1}| < |\omega_z| < |\omega_{p2}|$. As shown in Fig. 10.31, the zero introduces significant phase shift while preventing the gain from falling sufficiently.

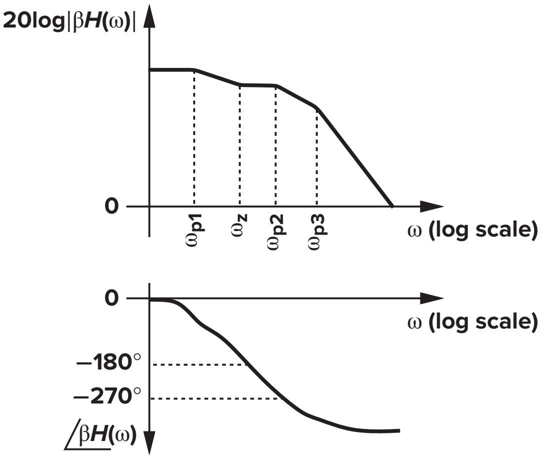


Figure 10.31 Effect of right-half-plane zero.

▶ **Example 10.7**

Noting that the Miller compensation in Fig. 10.29(a) yields $\omega_{p2} \approx g_{m9}/C_L$ and $\omega_z \approx g_{m9}/C_C$, a student decides to choose $C_C = C_L$, aiming to cancel the second pole by the zero. Explain what happens.

Solution

Recall that the zero is located in the *right half* plane and the poles in the left half plane. The compensated loop transmission can therefore be expressed as

$$\beta H(s) = \frac{\beta A_0(1 - \dfrac{s}{\omega_z})}{(1 + \dfrac{s}{\omega_{p1}})(1 + \dfrac{s}{\omega_{p2}})} \tag{10.29}$$

We recognize that the zero does *not* cancel the pole and still affects $|\beta H|$ and $\angle\beta H$. ◀

The right-half-plane zero in two-stage CMOS op amps, given by $g_m/(C_C + C_{GD})$, is a serious issue because g_m is relatively small and C_C is chosen large enough to position the dominant pole properly. Various techniques for eliminating or moving the zero have been invented. Illustrated in Fig. 10.32, places a resistor in series with the compensation capacitor, thereby modifying the zero frequency. The output stage now exhibits *three* poles, but for moderate values of R_z, the third pole is located at high frequencies and the first two poles are close to the values calculated with $R_z = 0$. Moreover, it can be

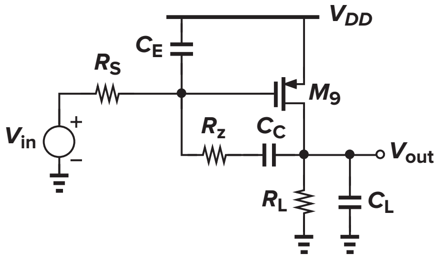


Figure 10.32 Addition of R_z to move the right-half-plane zero.

shown (Problem 10.8) that the zero frequency is given by

$$\omega_z \approx \frac{1}{C_C\left(g_{m9}^{-1} - R_z\right)} \tag{10.30}$$

Thus, if $R_z \geq g_{m9}^{-1}$, then $\omega_z \leq 0$. While $R_z = g_{m9}^{-1}$ seems a natural choice, in practice we may even move the zero well into the left half plane so as to cancel the first nondominant pole. This occurs if

$$\frac{1}{C_C\left(g_{m9}^{-1} - R_z\right)} = \frac{-g_{m9}}{C_L + C_E} \tag{10.31}$$

that is

$$R_z = \frac{C_L + C_E + C_C}{g_{m9}C_C} \tag{10.32}$$

$$\approx \frac{C_L + C_C}{g_{m9}C_C} \tag{10.33}$$

because C_E is typically much less than $C_L + C_C$.

The possibility of canceling the nondominant pole makes this technique attractive, but in reality two important drawbacks must be considered. First, it is difficult to guarantee the relationship given by (10.33), especially if C_L is unknown or variable. Mismatch between the pole and zero frequencies leads to the "doublet problem" (Problem 10.19). For example, as explained in Chapter 13, the load capacitance seen by an op amp may vary from one part of the period to another in switched-capacitor circuits, necessitating a corresponding change in R_z and complicating the design. The second drawback relates to the actual implementation of R_z. Typically realized by a MOS transistor in the triode region (Fig. 10.33), R_z changes substantially as output voltage excursions are coupled through C_C to node X, thereby degrading the large-signal settling response.

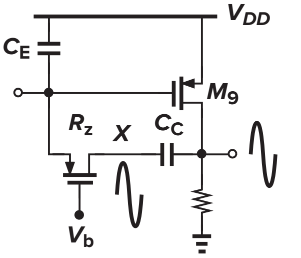


Figure 10.33 Effect of large output swings on R_z.

Generating V_b in Fig. 10.33 is not straightforward because R_Z must remain equal to $(1+C_L/C_C)/g_{m9}$ despite process and temperature variations. A common approach is illustrated in Fig. 10.34 [2], where diode-connected devices M_{13} and M_{14} are placed in series. If I_1 is chosen with respect to I_{D9} such that $V_{GS13} = V_{GS9}$, then $V_{GS15} = V_{GS14}$. Since $g_{m14} = \mu_p C_{ox}(W/L)_{14}(V_{GS14} - V_{TH14})$ and $R_{on15} = [\mu_p C_{ox}(W/L)_{15}(V_{GS15} - V_{TH15})]^{-1}$, we have $R_{on15} = g_{m14}^{-1}(W/L)_{14}/(W/L)_{15}$. For pole-zero cancellation to occur,

$$g_{m14}^{-1}\frac{(W/L)_{14}}{(W/L)_{15}} = g_{m9}^{-1}\left(1 + \frac{C_L}{C_C}\right) \tag{10.34}$$

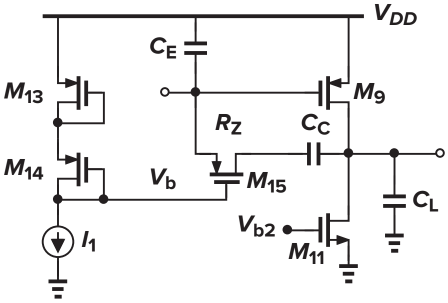


Figure 10.34 Generation of V_b for proper temperature and process tracking.

and hence

$$(W/L)_{15} = \sqrt{(W/L)_{14}(W/L)_9}\sqrt{\frac{I_{D9}}{I_{D14}}\frac{C_C}{C_C + C_L}} \tag{10.35}$$

If C_L is constant, (10.35) can be established with reasonable accuracy because it contains only the *ratio* of quantities.

Another method of guaranteeing Eq. (10.33) is to use a simple resistor for R_Z and define g_{m9} with respect to a resistor that closely matches R_Z [3]. Depicted in Fig. 10.35, this technique incorporates M_{b1}–M_{b4} along with R_S to generate $I_b \propto R_S^{-2}$. (This circuit is studied in detail in Chapter 12.) Thus, $g_{m9} \propto \sqrt{I_{D9}} \propto \sqrt{I_{D11}} \propto R_S^{-1}$. Proper ratioing of R_Z and R_S therefore ensures that (10.33) is valid even with temperature and process variations.

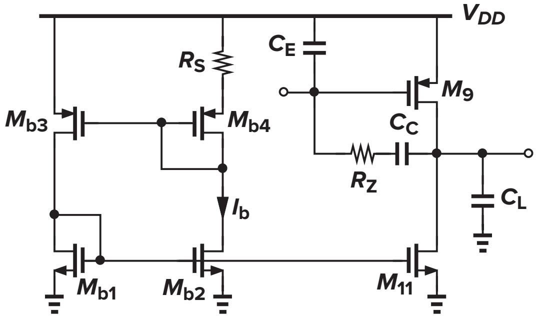


Figure 10.35 Method of defining g_{m9} with respect to R_S.

The principal drawback of the two methods described above is that they assume square-law characteristics for all of the transistors. As described in Chapter 17, short-channel MOSFETs may substantially deviate from the square-law regime, creating errors in the foregoing calculations. In particular, transistor M_9 is typically a short-channel device because it appears in the signal path and its raw speed is critical.

An attribute of two-stage op amps that makes them inferior to "one-stage" op amps is the susceptibility to the load capacitance. Since Miller compensation establishes the dominant pole at the output of the first stage, a higher load capacitance presented to the second stage moves the second pole toward the origin, degrading the phase margin. By contrast, in one-stage op amps, a higher load capacitance brings the *dominant* pole closer to the origin, *improving* the phase margin (albeit making the feedback system more overdamped). Illustrated in Fig. 10.36 is the step response of a unity-gain feedback amplifier employing a one-stage or a two-stage op amp, suggesting that the response approaches an oscillatory behavior if the load capacitance seen by the two-stage op amp increases.

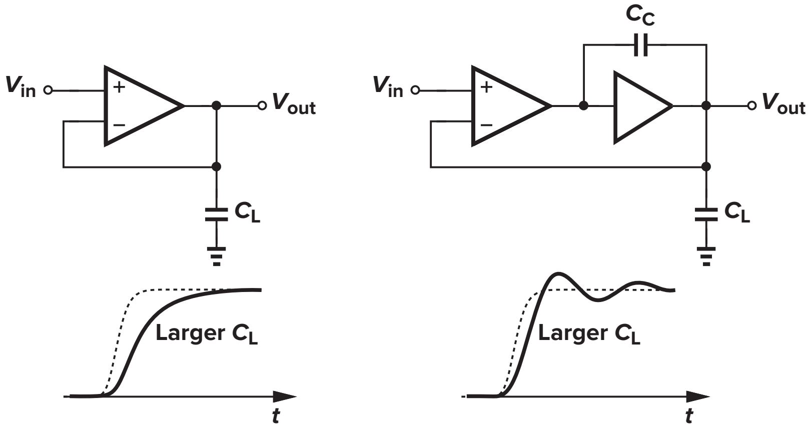


Figure 10.36 Effect of increased load capacitance on step response of one- and two-stage op amps.

10.6 ■ Slewing in Two-Stage Op Amps

It is instructive to study the slewing characteristics of two-stage op amps. Before delving into the details, let us consider the simple circuit shown in Fig. 10.37(a), where I_{in} is a current step given by $I_{SS}u(t)$ and C_F has a zero initial condition. If A is large, node X is a virtual ground and the voltage across C_F is approximately equal to V_{out}. Receiving a constant current equal to I_{SS}, C_F generates an output voltage given by

$$V_{out}(t) \approx \frac{I_{SS}}{C_F}t \tag{10.36}$$

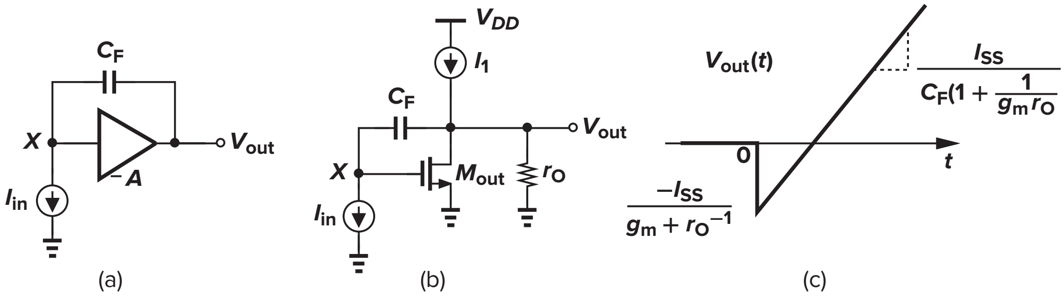


Figure 10.37 (a) Simplified circuit for slew study, (b) realization of (a), and (c) output waveform during slewing.

We now consider the implementation depicted in Fig. 10.37(b)[6] and write $V_{out}/r_O + g_m V_X + I_{in} = I_1$ and $C_F d(V_{out} - V_X)/dt = I_{in}$. Substituting for V_X from the former equation in the latter, we have

$$C_F\left(1+\frac{1}{g_m r_O}\right)\frac{dV_{out}}{dt} = I_{in} - \frac{C_F}{g_m}\frac{dI_{in}}{dt} \tag{10.37}$$

[6] The bias network for M_{out} is not shown.

We consider the terms on the right-hand side as two inputs and apply superposition, obtaining

$$V_{out}(t) = \frac{I_{SS}}{C_F(1+\dfrac{1}{g_m r_O})} tu(t) - \frac{I_{SS}}{g_m + \dfrac{1}{r_O}} u(t) \tag{10.38}$$

(This voltage, of course, rides on top of a bias value.) As illustrated in Fig. 10.37(c), V_{out} initially jumps to $-I_{SS}/(g_m + r_O^{-1})$ and subsequently ramps up with a slope equal to $I_{SS}/[C_F(1+g_m^{-1}r_O^{-1})]$. It is interesting to note that (1) at $t = 0^+$, C_F acts as a short circuit, allowing I_{in} to flow through $(1/g_m)||r_O$ and creating a downward step at the output; (2) the slope of the ramp suggests an equivalent capacitance of $C_F(1+g_m^{-1}r_O^{-1})$, revealing the Miller effect of C_F at the *output*; and (3) Eq. (10.38) does not depend on I_1 because this current simply serves as the bias current of M_{out}. We approximate the output voltage as $V_{out}(t) \approx (I_{SS}/C_F)tu(t)$.

Let us return to a two-stage op amp and suppose that in Fig. 10.38(a), V_{in} experiences a large positive step at $t = 0$, turning off M_2, M_4, and M_3. The circuit can then be simplified to that in Fig. 10.38(b), revealing that C_C is charged by a constant current I_{SS} if parasitic capacitances at node X are negligible. Recognizing that the gain of the output stage makes node X a virtual ground, we write $V_{out}(t) \approx (I_{SS}t/C_C)u(t)$. Thus, the positive slew rate[7] equals I_{SS}/C_C. Note that during slewing, M_5 must provide *two* currents: I_{SS} and I_1. If M_5 is not wide enough to sustain $I_{SS} + I_1$ in saturation, then V_X drops significantly, possibly driving M_1 into the triode region.

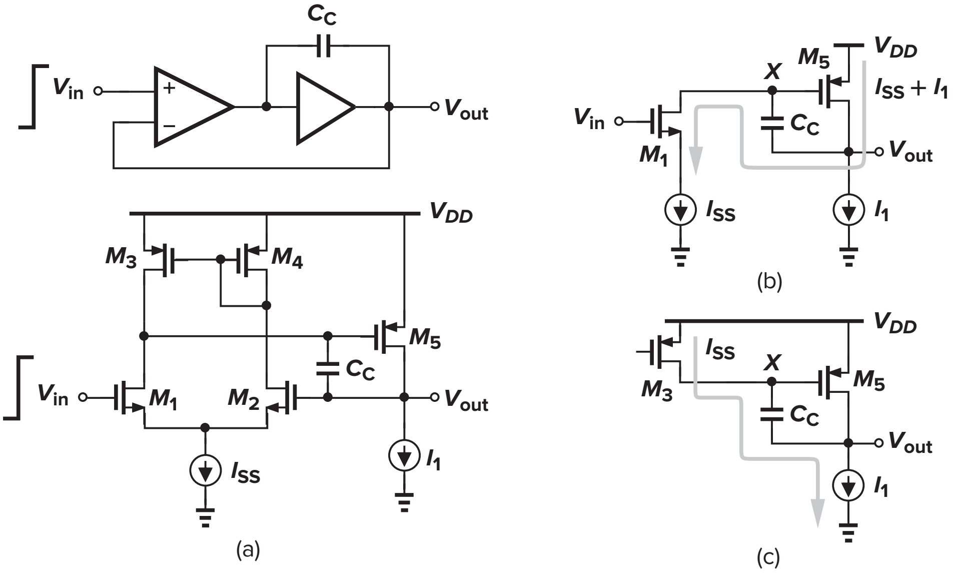


Figure 10.38 (a) Simple two-stage op amp, (b) simplified circuit during positive slewing, and (c) simplified circuit during negative slewing.

For the negative slew rate, we simplify the circuit as shown in Fig. 10.38(c). Here I_1 must support both I_{SS} and I_{D5}. For example, if $I_1 = I_{SS}$, then V_X rises so as to turn off M_5. If $I_1 < I_{SS}$, then M_3 enters the triode region and the slew rate is given by I_{D3}/C_C.

[7] The term “positive” refers to the slope of the waveform at the output of the op amp.

▶ Example 10.8

Op amps typically drive a heavy load capacitance. Repeat the slew rate analysis if the circuit of Fig. 10.37(b) sees a load capacitance of C_L. For simplicity, neglect channel-length modulation.

Solution

We consider two cases: I_{in} flows into or out of node X. With $\lambda = 0$, the steady-state gain from V_X to V_{out} is infinite, forcing X to be a virtual ground node. In the first case [Fig. 10.39(a)], $I_{in} = I_{SS}u(t)$ flows through C_F, generating a ramp voltage across it. Since V_X is constant, the voltage at the right terminal of C_F, V_{out}, must *fall* at a rate of I_{SS}/C_F. This also means that C_L is discharged at the same rate, requiring that the transistor draw three currents: I_1, I_{SS}, and $C_L dV_{out}/dt = (C_L/C_F)I_{SS}$. Thus, so long as M_{out} remains in saturation, the output slew rate is approximately equal to I_{SS}/C_F.

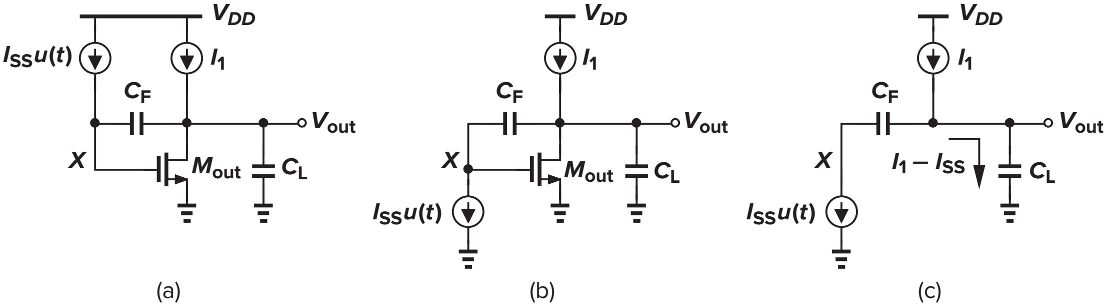


Figure 10.39

Now, let us study the second case [Fig. 10.39(b)]. If X is a virtual ground, V_{out} must rise at a rate of I_{SS}/C_F, and C_L must also receive a current of $C_L dV_{out}/dt = (C_L/C_F)I_{SS}$. We observe that, if $I_1 > I_{SS}(C_L/C_F) + I_{SS}$, then M_{out} remains on, V_X varies little, and the output slew rate is equal to I_{SS}/C_F. On the other hand, if $I_1 < (1 + C_L/C_F)I_{SS}$, M_{out} turns off, the difference between I_1 and I_{SS} charges C_L [Fig. 10.39(c)], and the slew rate is given by $(I_1 - I_{SS})/C_L$, a low value. ◀

Two-Stage Class-AB Op Amps The two-stage class-AB op amp studied in Chapter 9 can incorporate Miller compensation as well. Recall, however, that the current mirrors in the signal path contribute an additional pole, degrading the phase margin. For this reason, two-stage class-AB op amps are typically slower than their class-A counterparts.

We wish to compute the slew rate of two-stage class-AB op amps. Let us redraw the circuit of Fig. 10.39(b) for this op amp topology (Fig. 10.40). In this case, too, the slew rate is equal to $(I_1 - I_{SS})/C_L$ if M_{out} turns off, but, by virtue of class-AB operation, I_1 itself can be quite large. The current mirror action yields $I_1 = (W_{p1}/W_{p2})\alpha I_{in}$ and hence a slew rate of $[\alpha(W_{p1}/W_{p2}) - 1]I_{SS}/C_L$.

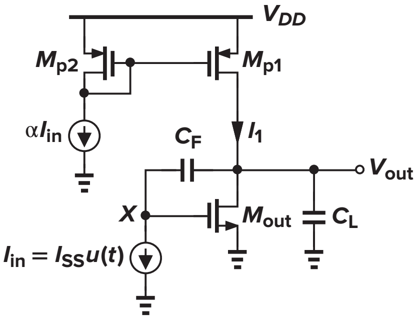
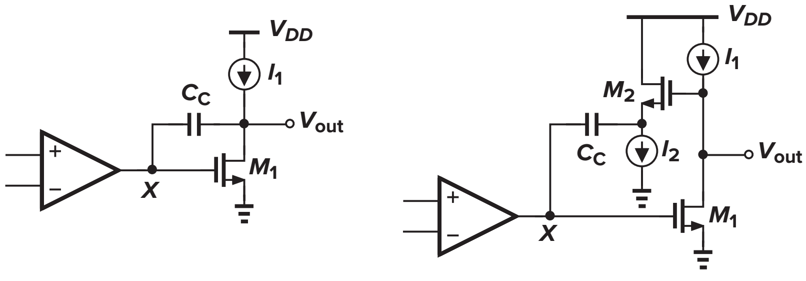


Figure 10.40 Simplified class-AB op amp.

10.7 ■ Other Compensation Techniques

The difficulty in compensating two-stage CMOS op amps arises from the feedforward path formed by the compensation capacitor [Fig. 10.41(a)]. If C_C could conduct current from the output node to node X but not vice versa, then the zero would move to a very high frequency. As shown in Fig. 10.41(b), this can be accomplished by inserting a source follower in series with the capacitor. Since the gate-source capacitance of M_2 is typically much less than C_C, we expect the right-half-plane zero to occur at high frequencies. Assuming that $\gamma = \lambda = 0$ for the source follower, neglecting some of the device capacitances, and simplifying the circuit as shown in Fig. 10.42, we can write $-g_{m1}V_1 = V_{out}(R_L^{-1} + C_L s)$, and hence

$$V_1 = \frac{-V_{out}}{g_{m1}R_L}(1 + R_L C_L s) \tag{10.39}$$



(a) (b)

Figure 10.41 (a) Two-stage op amp with right-half-plane zero due to C_C; (b) addition of a source follower to remove the zero.

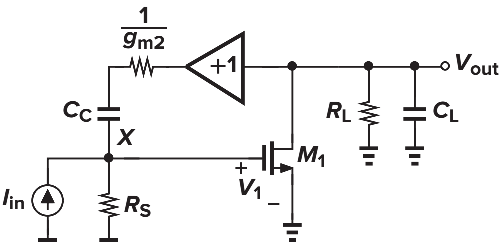


Figure 10.42 Simplified equivalent circuit of Fig. 10.41(b).

We also have

$$\frac{V_{out} - V_1}{\dfrac{1}{g_{m2}} + \dfrac{1}{C_C s}} + I_{in} = \frac{V_1}{R_S} \tag{10.40}$$

Substituting for V_1 from (10.39) yields

$$\frac{V_{out}}{I_{in}} = \frac{-g_{m1}R_L R_S(g_{m2} + C_C s)}{R_L C_L C_C(1 + g_{m2}R_S)s^2 + [(1 + g_{m1}g_{m2}R_L R_S)C_C + g_{m2}R_L C_L]s + g_{m2}} \tag{10.41}$$

Thus, the circuit contains a zero in the *left* half plane, which can be chosen to cancel one of the poles. The zero can also be derived as illustrated in Fig. 6.18.

We can also compute the magnitudes of the two poles, assuming that they are widely separated. Since typically $1 + g_{m2}R_S \gg 1$ and $(1 + g_{m1}g_{m2}R_LR_S)C_C \gg g_{m2}R_LC_L$, we have

$$\omega_{p1} \approx \frac{g_{m2}}{g_{m1}g_{m2}R_LR_SC_C} \tag{10.42}$$

$$\approx \frac{1}{g_{m1}R_LR_SC_C} \tag{10.43}$$

and

$$\omega_{p2} \approx \frac{g_{m1}g_{m2}R_LR_SC_C}{R_LC_LC_Cg_{m2}R_S} \tag{10.44}$$

$$\approx \frac{g_{m1}}{C_L} \tag{10.45}$$

Thus, the new values of ω_{p1} and ω_{p2} are similar to those obtained by simple Miller approximation. For example, the output pole has moved from $(R_LC_L)^{-1}$ to g_{m1}/C_L.

The primary issue in the circuit of Fig. 10.41(b) is that the source follower limits the lower end of the output voltage to $V_{GS2} + V_{I2}$, where V_{I2} is the voltage required across I_2. For this reason, it is desirable to utilize the compensation capacitor to isolate the dc levels in the active feedback stage from that at the output. Such a topology is depicted in Fig. 10.43, where C_C and the common-gate stage M_2 convert the output voltage swing to a current, returning the result to the gate of M_1 [4]. If V_1 changes by ΔV and V_{out} by $A_v\Delta V$, then the current through the capacitor is nearly equal to $A_v\Delta VC_Cs$ because $1/g_{m2}$ can be relatively small. Thus, a change ΔV at the gate of M_1 creates a current change of $A_v\Delta VC_Cs$, providing a capacitor multiplication factor equal to A_v.

Assuming that $\lambda = \gamma = 0$ for the common-gate stage, we redraw the circuit of Fig. 10.43 in Fig. 10.44, where we have

$$V_{out} + \frac{g_{m2}V_2}{C_Cs} = -V_2 \tag{10.46}$$

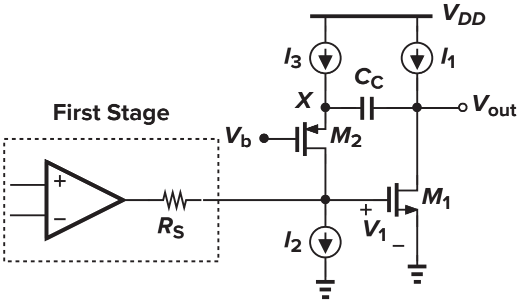


Figure 10.43 Compensation technique using a common-gate stage.

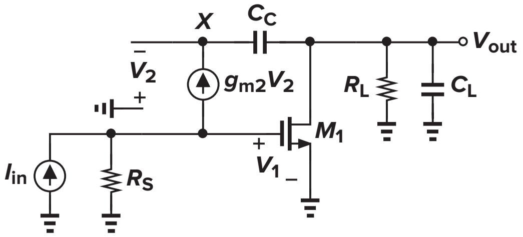


Figure 10.44 Simplified equivalent circuit of Fig. 10.43.

and hence

$$V_2 = -V_{out} \frac{C_C s}{C_C s + g_{m2}} \tag{10.47}$$

Also,

$$g_{m1} V_1 + V_{out} \left(\frac{1}{R_L} + C_L s \right) = g_{m2} V_2 \tag{10.48}$$

and $I_{in} = V_1/R_S + g_{m2} V_2$. Solving these equations, we obtain

$$\frac{V_{out}}{I_{in}} = \frac{-g_{m1} R_S R_L (g_{m2} + C_C s)}{R_L C_L C_C s^2 + [(1 + g_{m1} R_S) g_{m2} R_L C_C + C_C + g_{m2} R_L C_L] s + g_{m2}} \tag{10.49}$$

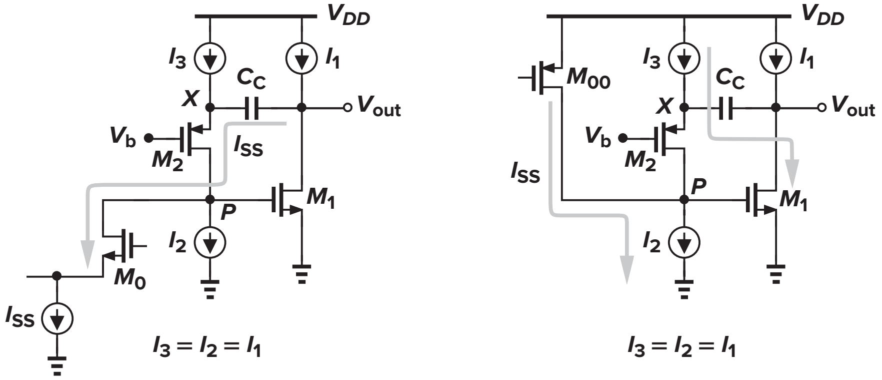
As with the circuit of Fig. 10.41(b), this topology contains a zero in the left half plane. Using similar approximations, we compute the poles as

$$\omega_{p1} \approx \frac{1}{g_{m1} R_L R_S C_C} \tag{10.50}$$

$$\omega_{p2} \approx \frac{g_{m2} R_s g_{m1}}{C_L} \tag{10.51}$$

Interestingly, the second pole has considerably risen in magnitude — by a factor of $g_{m2} R_S$ with respect to that of the circuit of Fig. 10.41. This is because at very high frequencies, the feedback loop consisting of M_2 and R_S in Fig. 10.43 lowers the output resistance by the same factor. Of course, if the capacitance at the gate of M_1 is taken into account, pole splitting is less pronounced. Nevertheless, this technique can potentially provide a high bandwidth in two-stage op amps.

The op amp of Fig. 10.43 entails important slewing issues. For positive slewing at the output, the simplified circuit of Fig. 10.45(a) suggests that M_2 and hence I_1 must support I_{SS}, requiring that $I_1 \geq I_{SS} + I_{D1}$. If I_1 is less, then V_P drops, turning M_1 off, and if $I_1 < I_{SS}$, M_0 and its tail current source must enter the triode region, yielding a slew rate equal to I_1/C_C.



(a) (b)

Figure 10.45 Circuit of Fig. 10.43 during (a) positive and (b) negative slewing.

For negative slewing, I_2 must support both I_{SS} and I_{D2} [Fig. 10.45(b)]. As I_{SS} flows into node P, V_P tends to rise, increasing I_{D1}. Thus, M_1 absorbs the current produced by I_3 through C_C, turning off M_2 and opposing the increase in V_P. We can therefore consider P a virtual ground node. This means that, for equal positive and negative slew rates, I_3 (and hence I_2) must be as large as I_{SS}, raising the power dissipation.

Op amps using a cascode topology as their first stage can incorporate a variant of the technique illustrated in Fig. 10.43. Shown in Fig. 10.46(a), this approach places the compensation capacitor between the *source* of the cascode devices and the output nodes. Using the simplified model of Fig. 10.46(b) and the method of Fig. 6.18, the reader can prove that the zero appears at $(g_{m4}R_{eq})(g_{m9}/C_C)$, a much greater magnitude than g_{m9}/C_C. If other capacitances are neglected, it can also be proved that the dominant pole is located at approximately $(R_{eq}g_{m9}R_LC_C)^{-1}$, as if C_C were connected to the gate of M_9 rather the source of M_4. The first nondominant pole is given by $g_{m4}g_{m9}R_{eq}/C_L$, an effect similar to that described by Eq. (10.51). In reality, the capacitance at X may create a significant pole because the resistance seen at this node is quite large. The analysis of the slew rate is left as an exercise for the reader. (One can also insert a resistor in series with each C_C to move the zero frequency.)

It is possible to combine two compensation techniques. As shown in Fig. 10.46(a), both C_C and C'_C provide greater flexibility in the design.

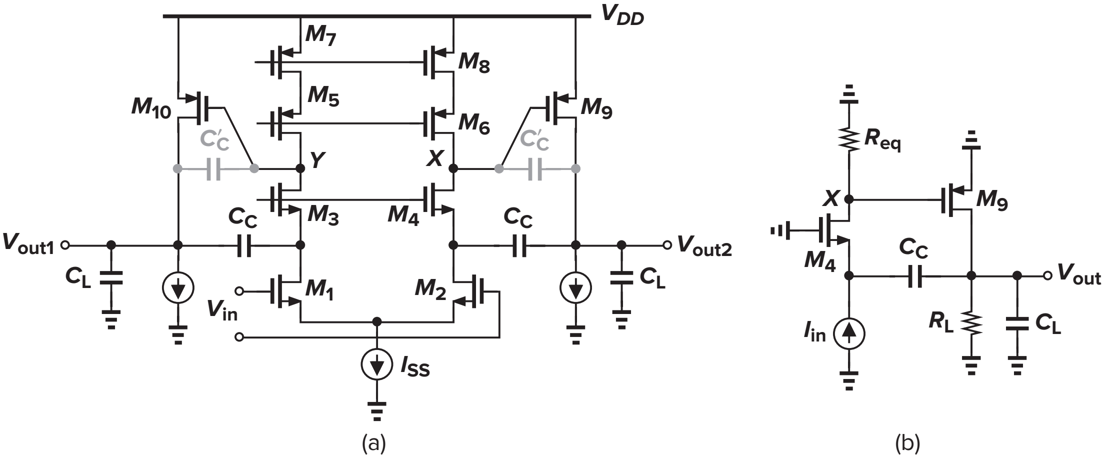


Figure 10.46 (a) Alternative method of compensating two-stage op amps; (b) simplified equivalent circuit of (a).

10.8 ■ Nyquist's Stability Criterion[8]

10.8.1 Motivation

Our analysis of stability in negative-feedback systems has drawn upon Bode's view of the loop transmission, namely, the magnitude and phase plots as a function of frequency, but only for $s = j\omega$. To understand the shortcomings of this approach, let us consider the loop transmission plots shown in Fig. 10.47, where $\beta = 1$ and $|H|$ is equal to 3 at the phase crossover frequency, ω_0. Our previous studies suggest that such a feedback system is unstable because it has a negative phase margin. However, if we write the closed-loop transfer function as $Y/X = H(s)/[1 + H(s)]$ and assume that $s = j\omega_0$, then we have

$$\frac{Y}{X}(j\omega_0) = \frac{-3}{1-3} \tag{10.52}$$

$$= \frac{3}{2} \tag{10.53}$$

[8] This section can be skipped in a first reading.

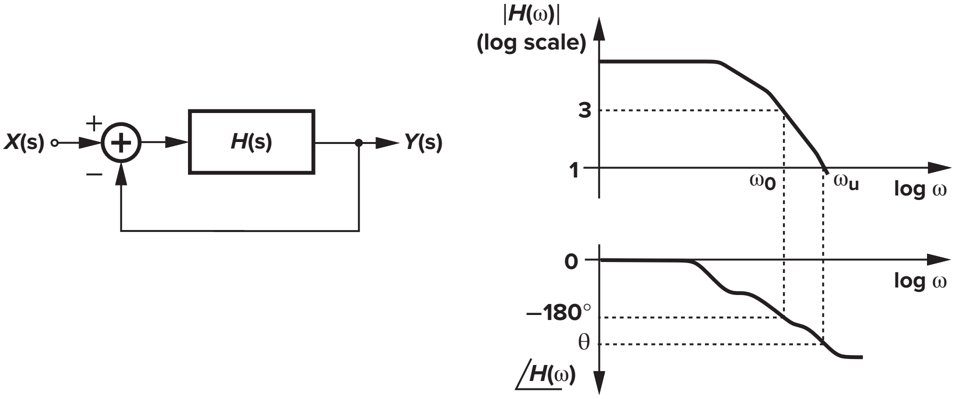


Figure 10.47 Unstable system Bode plots.

Since the closed-loop gain is less than infinity at ω_0, the circuit cannot oscillate at this frequency. In fact, for no value of $s = j\omega$ in Fig. 10.47 can we find the condition $Y/X = \infty$. For example, at ω_u, we have $Y/X = \exp(j\theta)/[1 + \exp(j\theta)] < \infty$ if $\theta \neq x \times 180°$.

Should we conclude that this system does not oscillate?! This difficulty arises because Bode plots confine s to imaginary values, i.e., they predict the behavior with only simple sinusoids. Indeed, this study shows that no simple sinusoid can circulate around the loop indefinitely. This, however, does not preclude other unstable waveforms. For example, suppose s is equal to $\sigma_1 + j\omega_1$ with $\sigma_1 > 0$, representing a *growing* sinusoid. It is possible that in the system of Fig. 10.47, $H(s = \sigma_1 + j\omega_1) = -1$; that is, Y/X goes to infinity for a growing sinusoid, allowing such a waveform to survive. Whether or not $s = \sigma_1 + j\omega_1$ exists is predicted by Nyquist's theorem but not by Bode plots.

We can exploit Nyquist's stability analysis, for it provides greater insight and, more important, tackles complex circuits more clearly. For a loop transmission $\beta H(s)$, this analysis predicts how many zeros $1 + \beta H(s)$ has in the right half plane (RHP) or on the $j\omega$ axis. If it has none, then the closed-loop system is stable.

Nyquist's method, however, is less intuitive and demands additional background in complex number theory. The reader should study this section patiently. We remind the reader that the poles and zeros of a transfer function can be shown on the complex s plane (Fig. 10.48).

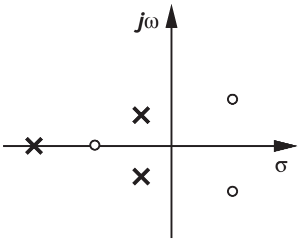


Figure 10.48 The s plane with poles and zeros.

10.8.2 Basic Concepts

Bode's approach to stability analysis plots the magnitude and phase of the loop gain versus frequency in Cartesian coordinates. We can also plot these two parameters in polar coordinates, in which every point is defined by an angle, θ, and a radius, r, rather than by x and y [Fig. 10.49(a)]. As the frequency varies, so do $\angle H$ and $|H|$, creating a "contour" in these coordinates [Fig. 10.49(b)]. The horizontal and vertical axes in Fig. 10.49(a) also carry a meaning: the projections of the vector on the two are expressed as

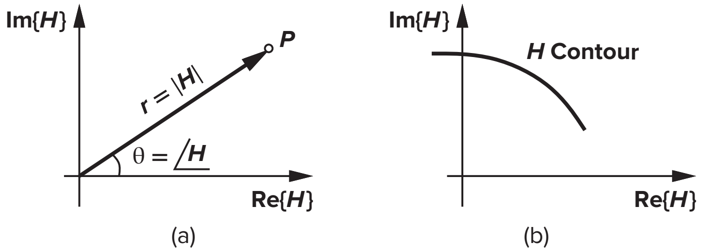


Figure 10.49 (a) One value of $H(s)$ shown in polar coordinates, and (b) contour of $H(s)$ as the frequency varies.

$|H|\cos(\angle H)$ and $|H|\sin(\angle H)$, respectively, with the former being the real part of H and the latter, the imaginary part. We thus denote the two axes by $Re\{H\}$ and $Im\{H\}$, respectively. We call the polar plot of $H(s)$ the "H contour." We initially assume that $s = j\omega$, but later allow it to become complex.

As an example, let us plot $H(s) = A_0/(1 + s/\omega_p)$ in polar coordinates if s is replaced with $j\omega$ and ω varies from 0 to $+\infty$. The phase, $-\tan^{-1}(\omega/\omega_p)$, begins at zero and approaches -90° while the magnitude, $A_0/\sqrt{1 + \omega^2/\omega_p^2}$, varies from A_0 to zero. Figure 10.50 sketches both the Bode plots and the polar plot, highlighting the corresponding points at $\omega = 0$ (M) and $\omega = \infty$ (N). The reader may wonder how we know that the polar plot is a semicircle. It is possible to prove this by calculating $Im\{H\}$ and $Re\{H\}$, but, as seen later, we do not have any interest in the actual shape. The beauty of Nyquist's method is that it primarily considers $\omega = 0$ and $\omega = \pm\infty$, avoiding the need for lengthy algebra.

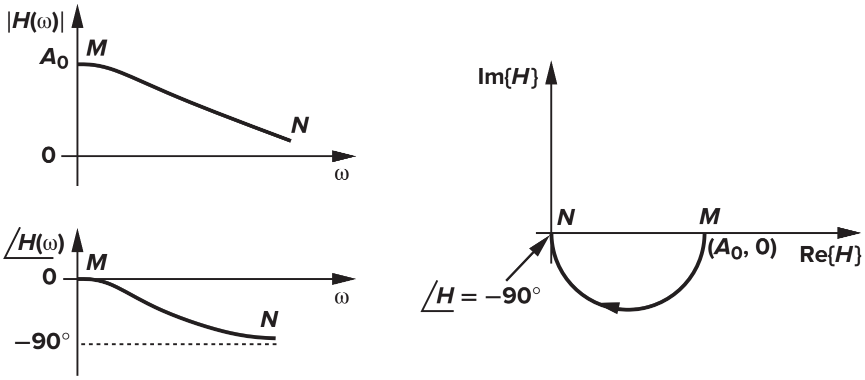


Figure 10.50 Bode and polar plots of H as $s = j\omega$ goes from zero to infinity.

This simple example readily demonstrates various decisions that we must make while plotting the H contour: (1) the contour begins at $(A_0, 0)$ for $\omega = 0$ and travels to the *left* because it must eventually reach the origin for $\omega = \infty$; (2) the contour falls *below* the horizontal axis because $\angle H$ is negative; and (3) the contour approaches the origin at an angle of -90°.

Since calculating $|H|$ and $\angle H$ is generally cumbersome, we wish to construct polar plots by considering only the poles and zeros of the transfer function. To understand the objective, consider the complex s plane for the above example, where the pole is real and equal to $-\omega_p$ (Fig. 10.51). As ω goes from 0 to $+\infty$, we begin at the origin and travel upward on the $j\omega$ axis.[9] Can we construct the polar plot of H by examining what happens in the s plane?

Let us first see whether $\angle H$ can be directly computed in the s plane. To this end, we consider one value for s and denote it by $s_1 = \sigma_1 + j\omega_1$, a complex value. Shown in Fig. 10.52 for a one-pole system,

[9] Recall that $H(\omega)$ is in fact $H(s = j\omega)$; i.e., the s values are confined to the $j\omega$ axis and the input is assumed to be a sinusoid.

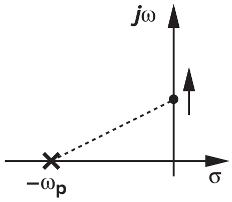


Figure 10.51 The s plane with one pole.

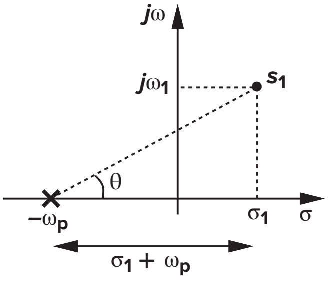


Figure 10.52 Phase shift produced by a pole at frequency s_1.

the s plane can yield a value for $\angle H(s = s_1)$. Since

$$H(s_1) = \frac{A_0}{1 + \dfrac{\sigma_1 + j\omega_1}{\omega_p}} \tag{10.54}$$

$$= \frac{A_0\omega_p}{\sigma_1 + \omega_p + j\omega_1} \tag{10.55}$$

we have

$$\angle H(s_1) = -\tan^{-1}\frac{\omega_1}{\sigma_1 + \omega_p} \tag{10.56}$$

Thus, the angle θ in Fig. 10.52 is equal to $-\angle H(s_1)$. That is, to determine $\angle H(s_1)$ in the s plane, we draw a vector from the pole to s_1, measure the angle of this vector with respect to the *positive* σ axis, and multiply the result by -1. For the phase contributed by a zero, the procedure is the same, except that the result is not multiplied by -1. If H contains multiple poles and zeros, then their phase contributions simply add algebraically.

It is possible to calculate $|H|$ from the s plane as well,[10] but, fortunately, the exact knowledge of $|H|$ is not necessary in Nyquist's approach.

10.8.3 Construction of Polar Plots

In this section, we study examples of plotting $H(s)$ in polar coordinates so as to prepare ourselves for Nyquist's stability criterion.

General First-Order System Suppose $H(s) = A_0(1 + s/\omega_z)/(1 + s/\omega_p)$ and $\omega_p > \omega_z$. We first plot H for $s = j\omega$ as ω varies from 0 to $+\infty$. At $\omega = 0$, $|H(s)| = A_0$. Also, as shown in Fig. 10.53(a), the vectors going from the pole and the zero to $s = 0$ contribute equal and opposite angles, yielding

[10] The magnitude of H can be determined as the product of the lengths of the vectors emanating from the zeros divided by the product of the lengths of the vectors emanating from the poles.

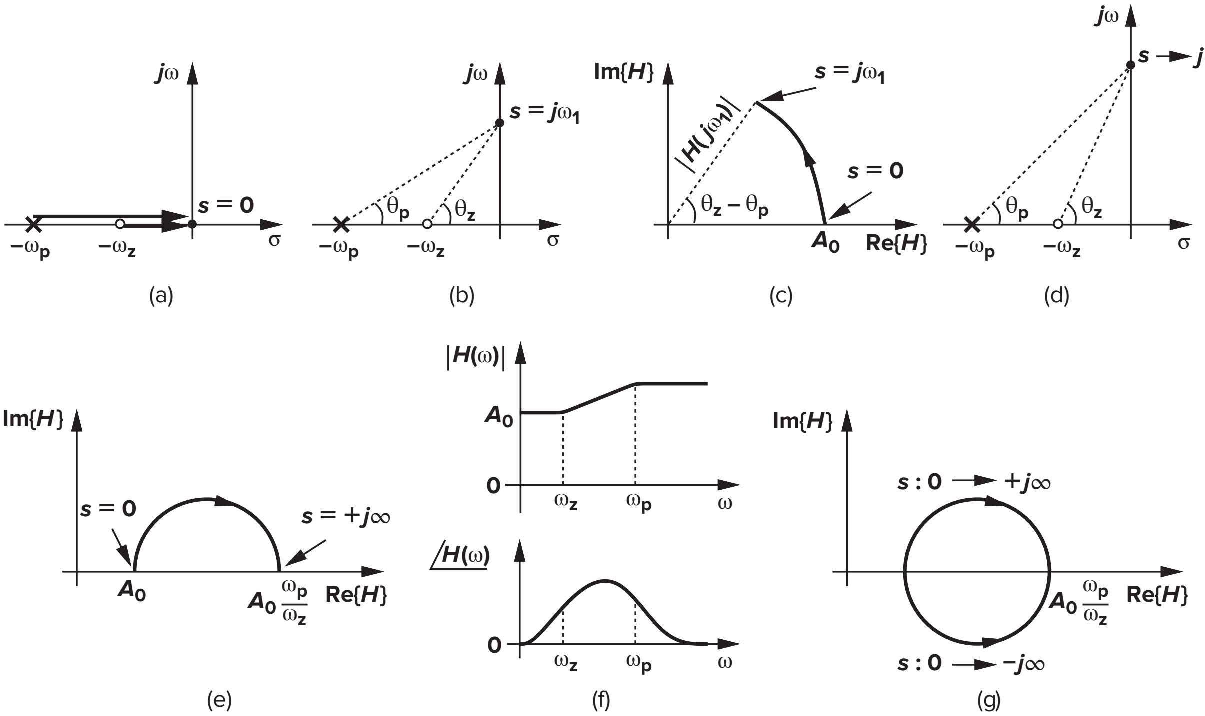


Figure 10.53 (a) Phase shifts contributed by a pole and a zero at $s = 0$, (b) phase shifts at $s = j\omega_1$, (c) contour of H as s goes from zero to $j\omega_1$, (d) phase shifts as $s \to j\infty$, (e) correct contour of H, (f) corresponding Bode plots, and (g) complete H contour.

$\angle H(0) = 0$. Now, if s rises to $j\omega_1$ [Fig. 10.53(b)], the angle contributed by the zero, θ_z, is *greater* than that contributed by the pole, θ_p. That is, $\angle H = \theta_z - \theta_p$ remains positive. We have thus far constructed the H contour shown in Fig. 10.53(c). The reader may wonder whether $|H(j\omega_1)|$ is greater or less than A_0, but we do not concern ourselves at this point.

What happens as s goes toward $+j\infty$? As depicted in Fig. 10.53(d), the angles arising from the zero and the pole approach 90°, producing a net value of 0 for $\angle H$. The magnitude of H, on the other hand, approaches $A_0\omega_p/\omega_z > A_0$. This means that the H contour returns to the σ axis, but at a more *positive* real value. Our guess in Fig. 10.53(c) is therefore not quite correct and must be revised to that in Fig. 10.53(e). For completeness, we also show the Bode plots in Fig. 10.53(f).

It is necessary to repeat this procedure as $s = j\omega$ goes from 0 to $-j\infty$. Since the s plane contents are always symmetric with respect to the σ axis for a physical system (due to conjugate symmetry of the poles and zeros), the polar plot is also symmetric, emerging as shown in Fig. 10.53(g).

What if $\omega_p < \omega_z$? As shown in Fig. 10.54(a), the net angle is now negative as s travels upward on the $j\omega$ axis. Moreover, since $|H(j\omega = 0)| = A_0$ and $|H(j\omega = +j\infty)| = A_0\omega_p/\omega_z < A_0$, the H contour begins from A_0 on the real axis, rotates downward, and shrinks in magnitude [Fig. 10.54(b)]. For $s = 0$ to $-j\infty$, this plot is reflected around the real axis in a manner similar to that in Fig. 10.53(g). The Bode plots are also constructed in Fig. 10.54(c) to highlight the correspondences.

We have mostly confined the s values in $H(s)$ to the $j\omega$ axis. In general, however, s can travel on an arbitrary path (contour) in the s plane, assuming complex, real, or imaginary values. It is therefore beneficial to consider the behavior of the foregoing first-order system in such a case. For example, suppose s travels clockwise on a closed contour in the right half plane [Fig. 10.55(a)]. How does $H(s) = A_0(1 + s/\omega_z)/(1 + s/\omega_p)$ behave if $\omega_p > \omega_z$? At point M, s is real and equal to σ_M, yielding

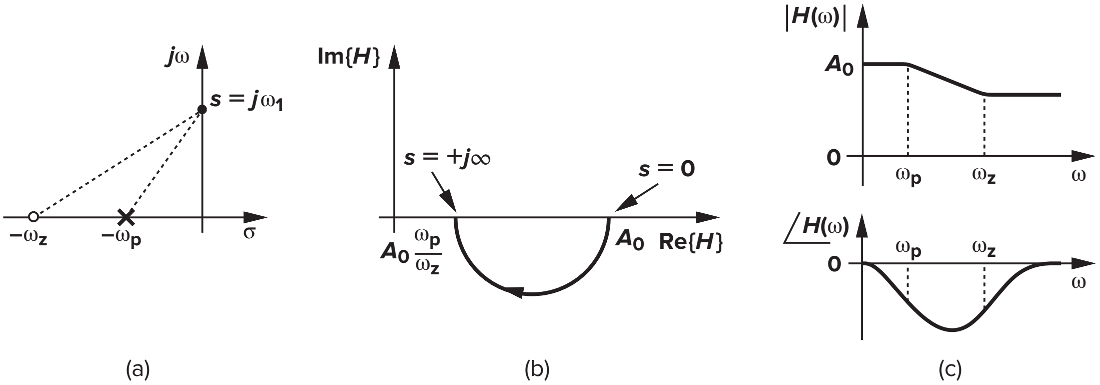


Figure 10.54 System with $\omega_p < \omega_z$, (b) contour of H, and (c) corresponding Bode plots.

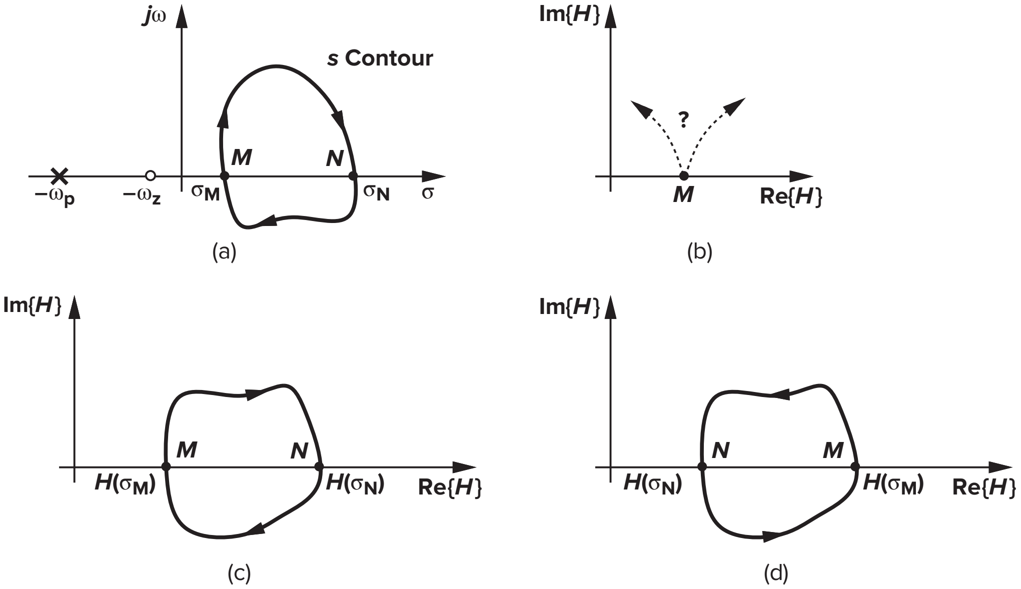


Figure 10.55 (a) s contour excluding pole and zero, (b) possible trajectories for H, (c) actual H contour, and (d) H contour if $-\omega_z < -\omega_p$.

$H(s) = A_0(1+\sigma_M/\omega_z)/(1+\sigma_M/\omega_p)$, a real point in the polar plot of H [Fig. 10.55(b)]. As s departs from M, the net angle becomes more positive because the zero contributes more phase than the pole does, but we do not know whether the H contour rises to the left or to the right. We therefore continue the s contour to point N, noting that the angle returns to zero and $H(s) = A_0(1+\sigma_N/\omega_z)/(1+\sigma_N/\omega_p)$, which is *greater* than $H(s = \sigma_M)$ if $\omega_p > \omega_z$. Thus, the H contour must rise to the right, i.e., rotate clockwise. Figure 10.55(c) depicts the complete plot as s traverses clockwise the contour in the s plane from M to N and from N back to M. For the case of $\omega_p < \omega_z$, H rotates counterclockwise [Fig. 10.55(d)] because $H(\sigma_N) < H(\sigma_M)$. The reader is encouraged to repeat this analysis for $H(s) = A_0/(1+s/\omega_p)$.

Let us consider another s contour that *encloses* the pole and the zero of the transfer function. As illustrated in Fig. 10.56(a), we begin at point M and observe a net angle of zero and $H(\sigma_M) = A_0(1+\sigma_M/\omega_z)/(1+\sigma_M/\omega_p)$. Since σ_M is more negative than $-\omega_z$ and $-\omega_p$, $H(\sigma_M) > 0$, yielding a point on the real axis for the polar plot of H [Fig. 10.56(b)]. As we travel clockwise on the s contour, say to point s_1, the net angle becomes positive (why?), eventually returning to zero as we reach point N. Since σ_N is less negative than $-\omega_z$ and $-\omega_p$, $H(\sigma_N) > 0$. The reader can prove that $H(\sigma_N) > H(\sigma_M)$. If we now continue on the s contour from N toward M, the polar plot of H becomes negative and returns to zero. It is important to note that the H contour does not enclose the origin. We say that H does not "encircle" the origin. The s contours in both Figs. 10.55(a) and 10.56(a) lead to H contours that do not encircle the origin. The significance of this point becomes clear later.

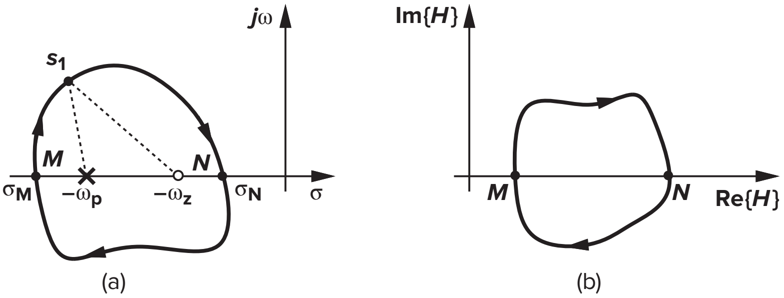


Figure 10.56 (a) s contour enclosing pole and zero, and (b) H contour.

What happens if the first-order system has no zero? As shown in Fig. 10.57(a), $\angle H$ is equal to -180° at point M, reaching -90° at s_1 and zero at N. Also, $H(\sigma_M) = A_0/(1 + \sigma_M/\omega_p) < 0$ and $H(\sigma_N) > 0$ [Fig. 10.57(b)]. We thus observe that $H(s)$ encircles the origin in this case, and that the encirclement is in the counterclockwise direction. Similarly, if the system has only one zero and no pole, a clockwise s contour containing the zero maps to an H contour that encircles the origin in the clockwise direction [Fig. 10.57(c)]. We hereafter assume that the s contours are symmetric around the σ axis, obtaining polar plots that are symmetric around the real axis.

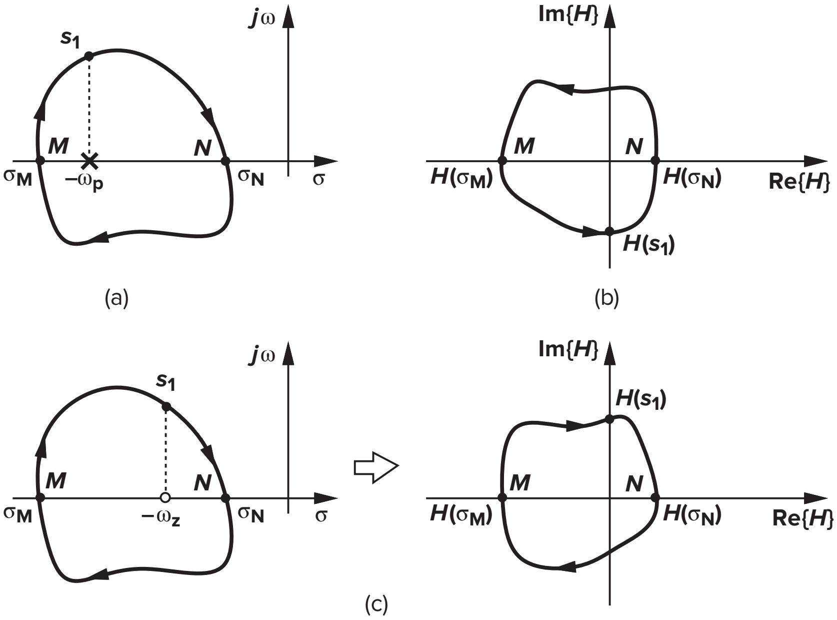


Figure 10.57 (a) System with one pole, (b) H contour, and (c) s and H contours if system has only one zero.

We now study a first-order system that has both a zero and a pole while the s contour encircles only the pole. We note from Fig. 10.58(a) that the polar plot of $H(s)$ assumes a positive value at M, as it did in Fig. 10.56(b). As s begins from point M and traverses the contour clockwise, $\angle H$ becomes more positive, reaching 180° at point N [Fig. 10.58(b)]. (It is helpful as a crosscheck to show that $H(\sigma_N) < 0$.) Thus, the H contour encircles the origin counterclockwise in a manner similar to that in Fig. 10.57(b). The reader can repeat this exercise with an s contour enclosing only the zero and prove that the plot of $H(s)$ encircles the origin clockwise.

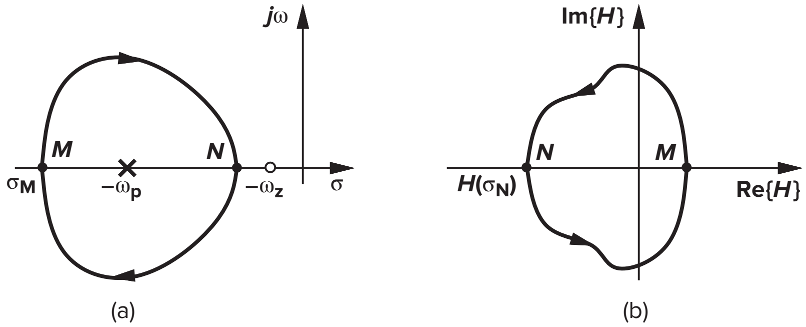


Figure 10.58 (a) s contour enclosing only one pole, and (b) H contour.

Second-Order System Consider $H(s) = A_0[(1 + s/\omega_{p1})(1 + s/\omega_{p2})]^{-1}$ and assume that s travels upward on the $j\omega$ axis [Fig. 10.59(a)]. We recognize that $|H(s)|$ begins at A_0 and falls as $s \rightarrow +j\infty$. Also, $\angle H(s)$ begins at 0 and becomes more negative, reaching −90° and, asymptotically, −180°. As depicted in Fig. 10.59(b), the H contour begins at A_0, rotates clockwise, crosses the $j\omega$ axis when

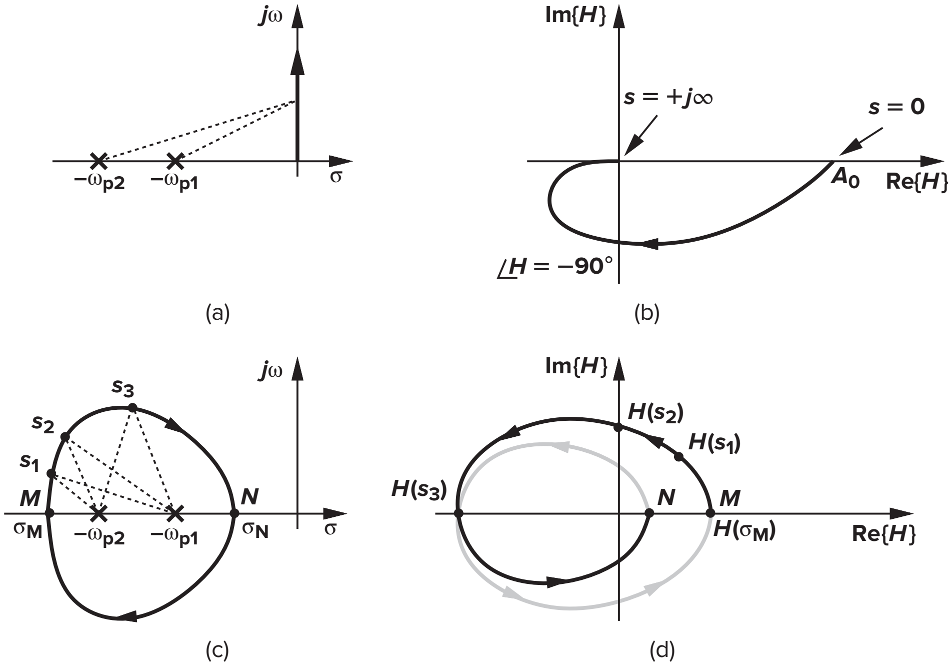


Figure 10.59 (a) Two-pole system, (b) H contour as s travels up on the $j\omega$ axis, (c) s contour chosen to enclose both poles, and (d) corresponding H contour.

$\angle H = -90^\circ$, enters the third quadrant, and eventually returns to the origin at a 180° angle. For $s = 0$ to $-j\infty$, this plot is reflected around the real axis.

What if the contour of s encloses both poles? From Fig. 10.59(c), we note that $\angle H = -360^\circ$ at M and $H(\sigma_M) > 0$. As we travel clockwise on the s contour to some point s_1, the angle becomes less negative, e.g., equal to $-320^\circ = +40^\circ$. Thus, the H contour rotates counterclockwise [Fig. 10.59(d)]. At some point, s_2, the net angle is around $-270^\circ = +90^\circ$, and at some other point, s_3, we have $\angle H(s_3) = -180^\circ$. As s approaches point N, $H(s)$ reaches a real, positive value. The other symmetric half is shown in gray for clarity. We observe that the polar plot encircles the origin *twice* in the counterclockwise direction if the s contour encircles *two* poles in the clockwise direction.

10.8.4 Cauchy's Principle

From the foregoing studies, we postulate that, if the s contour encircles P poles and Z zeros of $H(s)$ in the clockwise direction, then the polar plot of $H(s)$ encircles the origin $Z - P$ times in the same direction. This is known as "Cauchy's Principle of Argument." For example, if the s contour encircles clockwise three zeros and no poles, then the H contour encircles the origin $3 - 0 = 3$ times clockwise. As seen earlier, the H contour is constructed primarily from the angles contributed by the poles and zeros, with little need for the exact knowledge of $|H|$.

We have thus far assumed that we know the locations of the poles and zeros of a transfer function and we construct the polar plot to see how many times it encircles the origin. One can embark on a different task: suppose we know that an s contour contains P poles but do *not* know the number of zeros within the contour. If we still manage to draw the polar plot of the transfer function and find that it encircles the origin clockwise N times, we can conclude that the number of zeros within the s contour is equal to $Z = N + P$. This is the key to Nyquist's stability theorem.

10.8.5 Nyquist's Method

Having studied the foregoing concepts patiently, the reader is now ready to learn Nyquist's stability analysis. A negative-feedback system whose closed-loop transfer function is given by

$$\frac{Y}{X}(s) = \frac{H(s)}{1 + \beta H(s)} \tag{10.57}$$

becomes unstable if it has any poles on the $j\omega$ axis or in the right half plane, both of which we call herein the "critical region." In other words, if $1 + \beta H(s)$ has any *zeros* in this critical region, then the system is unstable.

How do we determine whether $1 + \beta H(s)$ has any zeros in the critical region? Let us construct an s contour containing this region (Fig. 10.60). From Cauchy's principle, we know that the polar plot of

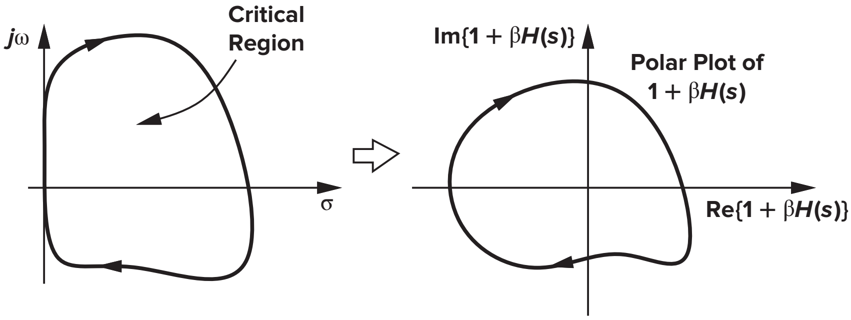


Figure 10.60 Critical region in the s plane and the corresponding H contour.

$1 + \beta H(s)$ encircles the origin $Z - P$ times, where Z and P respectively denote the number of zeros and poles that $1 + \beta H(s)$ contains within the s contour. We thus proceed as follows: (1) independently determine P, (2) draw the polar plot of $1 + \beta H(s)$ as s traverses the contour shown in Fig. 10.60, (3) determine the number of times, N, that $1 + \beta H(s)$ encircles the origin clockwise, and (4) find $P + N$ as the number of zeros that $1 + \beta H(s)$ has in the critical region.

We must recognize a point that simplifies our task. The poles of $1 + \beta H(s)$ are in fact the same as the poles of $H(s)$. If the open-loop system is stable (as is the case in most of our circuits), then $H(s)$ has no poles in the critical region and $N = Z$. Unless otherwise stated, we assume this to be true.

Before studying examples of the above procedure, we make one change that leads us to Nyquist's theorem: if the polar plot of $1 + \beta H(s)$ encircles the origin, then the polar plot of $\beta H(s)$ encircles the point (−1, 0) (Fig. 10.61) because the latter is obtained by shifting the former to the left by one unit. Nyquist's theorem articulates this result as for a closed-loop system, $H(s)/[1 + \beta H(s)]$, to be stable, the polar plot of $\beta H(s)$ must *not* encircle the point (−1, 0) clockwise as s traverses a contour around the critical region clockwise.

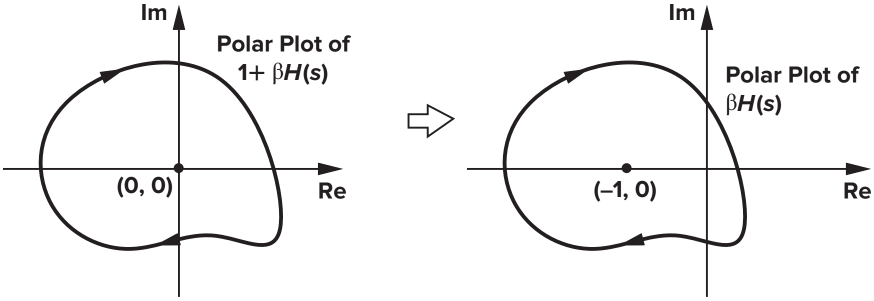


Figure 10.61 Polar plots of $1 + \beta H(s)$ and $\beta H(s)$.

In applying Nyquist's theorem, we must choose the s contour so as to minimize the mathematical labor. One possibility is depicted in Fig. 10.62: we begin at the origin, travel *on* the $j\omega$ axis to $+j\infty$, go around the RHP on a very large radius, continue to $j\omega = -j\infty$, and return to the origin *on* the $j\omega$ axis. The reader may wonder what exactly happens now that the contour does not *enclose* the $j\omega$ axis. If $1 + \beta H(s)$ has any zeros on this axis, then the polar plot of $\beta H(s)$ goes *through* the point (−1, 0) rather than encircle it. [Recall from Bode plots that $1 + \beta H(j\omega_1) = 0$ translates to $|\beta H(j\omega_1)| = 1$ and $\angle H(j\omega_1) = 180°$.] Since the s contour is symmetric around the σ axis, we construct the βH contour only as s goes from the origin to M and N, and simply reflect the result around the real axis to complete the task.

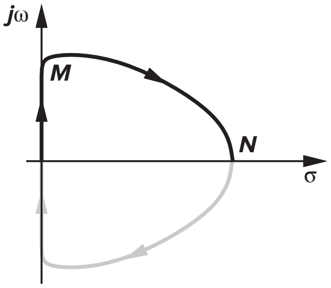


Figure 10.62 Simple contour enclosing the $j\omega$ axis and RHP.

▶ **Example 10.9**

Study the closed-loop stability if $H(s) = A_0/(1 + s/\omega_{p1})$.

Solution

For the s contour shown in Fig. 10.63(a), $\beta H(s)$ begins at βA_0 for $s = 0$. As $s = j\omega$ moves upward, the phase becomes more negative. At $+j\infty$, the phase goes to $-90°$ and the magnitude drops to zero, i.e., the polar plot reaches the origin at an angle of $-90°$ [Fig. 10.63(b)]. What happens as s enters the right half plane? Traveling in the RHP at a very long radius, s keeps $\beta H(s)$ at zero. That is, the entire RHP contour from M to N maps to the origin. We reflect this polar plot around the real axis to obtain the complete βH contour. Since the contour does not encircle $(-1, 0)$, the closed-loop system is always stable.

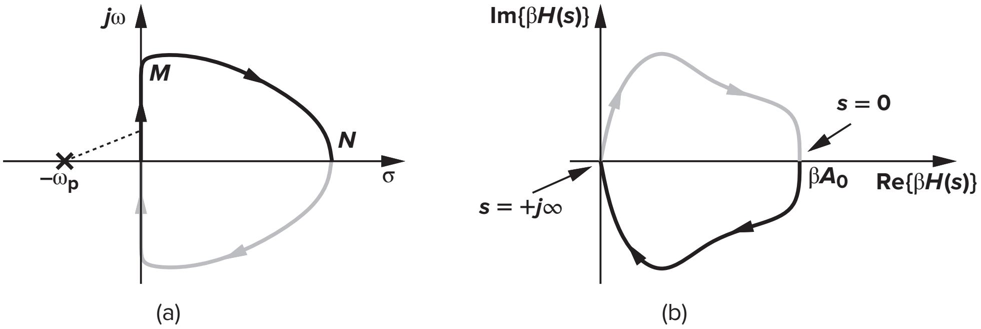


Figure 10.63 (a) s contour for a one-pole system, and (b) βH contour.

Example 10.10

Study the closed-loop stability if $H(s) = A_0/[(1 + s/\omega_{p1})(1 + s/\omega_{p2})]$.

Solution

At $s = 0$, $\beta H(s) = \beta A_0$. As $s = j\omega$ moves upward, the two poles contribute negative phase (Fig. 10.64). At $s = +j\infty$, the phase goes to $-180°$ and the magnitude falls to zero, i.e., the polar plot reaches the origin at an angle of $180°$. The βH contour remains at the origin as s traverses the RHP at a very long radius from M to N. Since the contour does not encircle $(-1, 0)$, the closed-loop system is stable for any value of the feedback factor, β. The reader is encouraged to repeat this exercise for increasingly larger values of ω_{p2}.

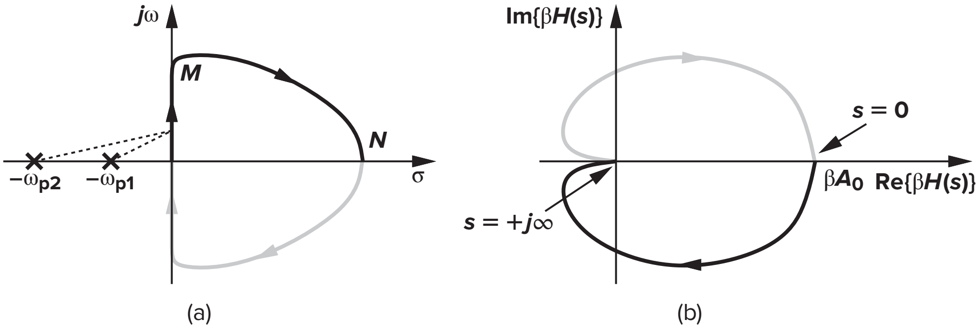


Figure 10.64 s plane and βH contours for a two-pole system.

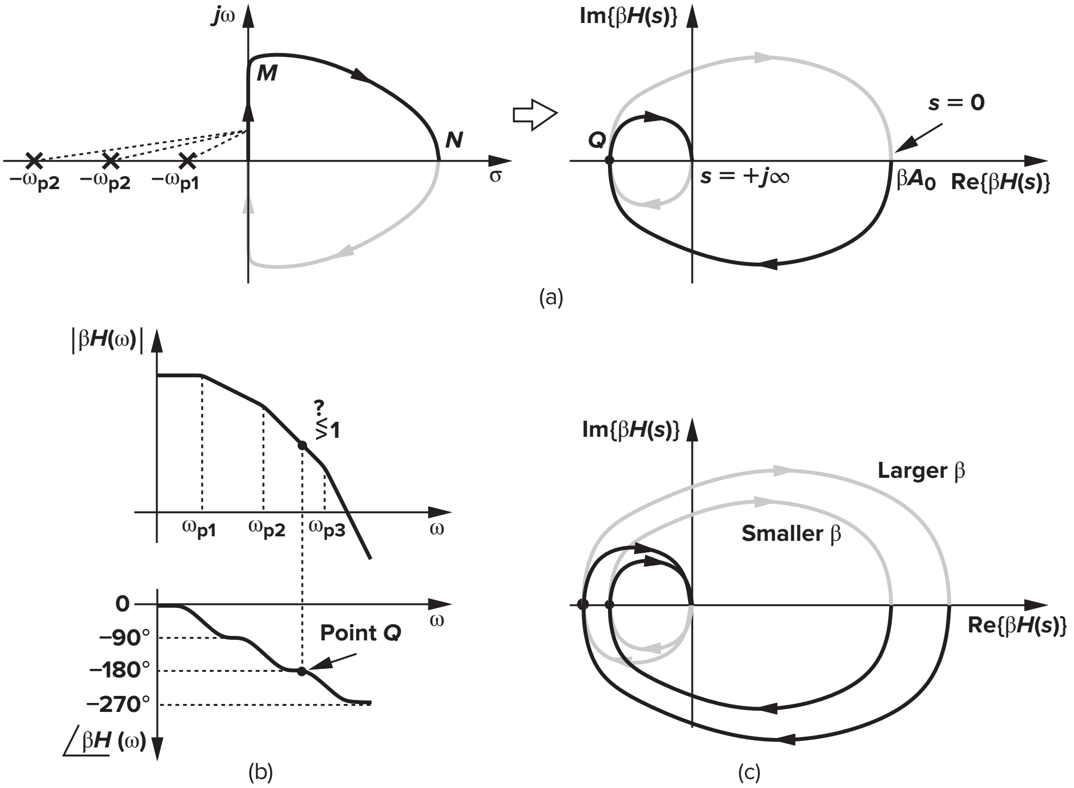
Example 10.11

Study the closed-loop stability if $H(s) = A_0/[(1 + s/\omega_{p1})(1 + s/\omega_{p2})(1 + s/\omega_{p3})]$.

Solution

The polar plot of $\beta H(s)$ begins at βA_0 and rotates clockwise, reaching an angle of $-270^\circ = +90^\circ$ and a magnitude of zero at $s = +j\infty$ [Fig. 10.65(a)]. The reflection of this half around the real axis completes the plot, revealing that the βH contour *can* encircle $(-1, 0)$ depending on the location of the intersection point, Q. At this point, $\angle\beta H = -180^\circ$, i.e., $\tan^{-1}(\omega_Q/\omega_{p1}) + \tan^{-1}(\omega_Q/\omega_{p2}) + \tan^{-1}(\omega_Q/\omega_{p3}) = 180^\circ$. With the pole values known, one can compute ω_Q and hence $|\beta H(s = j\omega_Q)|$ so as to determine whether point Q is to the right or to the left of $(-1, 0)$. The corresponding calculation on Bode plots is illustrated in Fig. 10.65(b).

What happens to the βH contour in Fig. 10.65(a) if different values of β are chosen? Since the radius at every point on the plot is proportional to β, the contour contracts as β decreases and expands as β increases. Illustrated in Fig. 10.65(c), this trend confirms that a higher feedback factor can make a three-pole system unstable.



(a)

(b)

(c)

Figure 10.65 (a) s and H contours for three-pole system, (b) corresponding Bode plots, and (c) H contour for different values of β.

In some cases, it may not be straightforward to determine how many times the βH contour encircles $(-1, 0)$ clockwise. The general procedure for counting the number of encirclements is as follows: (1) draw a straight line from $(-1, 0)$ to infinity in any direction, (2) count the number of times the contour crosses this line in clockwise and counterclockwise directions, and (3) subtract the latter from the former.

10.8.6 Systems with Poles at Origin

Some open-loop systems contain one or more poles at the origin. For example, the integrator shown in Fig. 10.66 has the following transfer function:

$$H(s) = \frac{-1}{R_1C_1s} \tag{10.58}$$

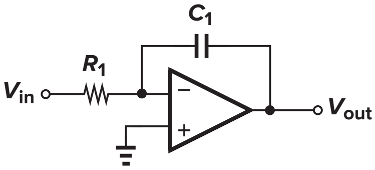


Figure 10.66 Integrator.

if the op amp is ideal. When such systems are placed in a negative-feedback loop, their Nyquist stability analysis must choose a slightly different s contour. We begin with a one-pole system as depicted in Fig. 10.67(a) and seek a contour that does not go through the origin so as to avoid an infinite value for $\beta H(s)$. Rather than begin at (0, 0), we travel on an infinitesimally small circle around it given by $\epsilon\exp(j\phi)$ until we reach the $j\omega$ axis and then move upward. The key point here is that ϵ is very small, simplifying the calculations.

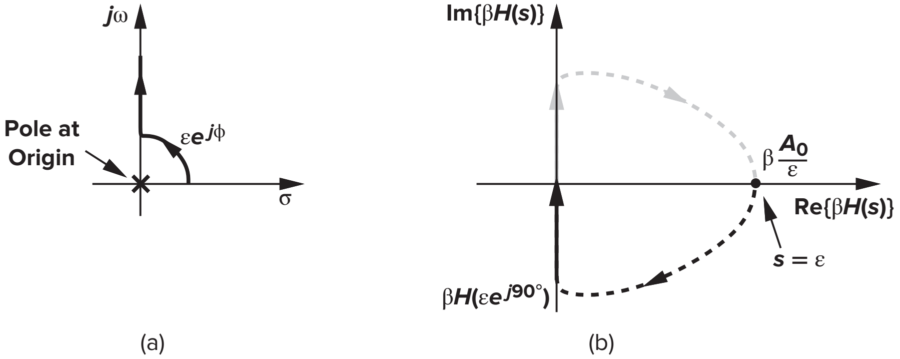


Figure 10.67 (a) s plane contour bypassing pole at origin, and (b) corresponding βH contour.

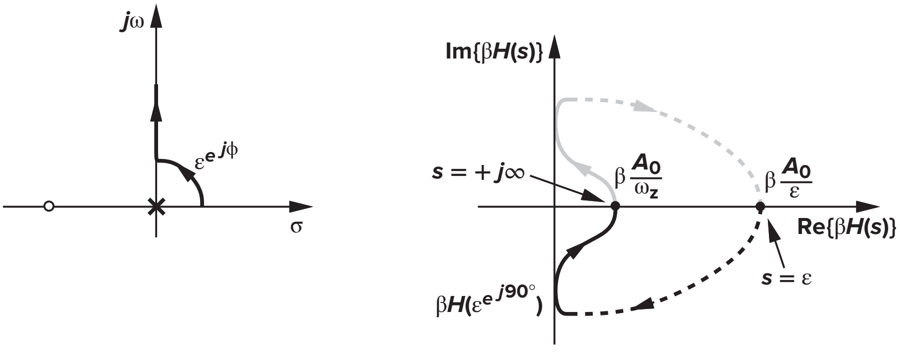
If $H(s) = A_0/s$ and we choose $s = \epsilon\exp(j\phi)$, then $\beta H(s) = \beta(A_0/\epsilon)\exp(-j\phi)$. At $\phi = 0$, $s = \epsilon$, and βH is real and very large [Fig. 10.67(b)]. As s traverses the circle, ϕ rotates toward +90° and $\beta(A_0/\epsilon)\exp(-j\phi)$ remains at a very large radius, approaching −90°. This behavior is indicated by a dashed curve in Fig. 10.67(b) to emphasize the large radius. Now, s travels upward on the $j\omega$ axis, still retaining a phase of −90° (due to the pole at the origin), while $|\beta H|$ falls, i.e., βH goes toward the origin at an angle of −90° and remains at (0, 0) as s enters the RHP (not shown in the s plane). The other half of the contour is obtained if s begins from the RHP (not shown), arrives at $-j\infty$, travels toward the origin on the $j\omega$ axis, and traverses the circle $\epsilon\exp(j\phi)$ from $\phi = -90°$ to $\phi = 0$. Note that the polar plot of βH does not encircle (−1, 0).

▶ Example 10.12

Analyze the closed-loop stability of $H(s) = A_0(1 + s/\omega_z)/s$. The zero can be created, for example, by inserting a resistor in series with C_1 in Fig. 10.66.

Solution

With $s = \epsilon\exp(j\phi)$ in Fig. 10.68(a), $\beta H(s) \approx \beta(A_0/\epsilon)\exp(-j\phi)$ because ϵ is small. Shown in Fig. 10.68(b) is the βH contour. Even at $\phi = 90°$, the zero contributes negligible phase because the circle's radius is very small. As we travel upward on the $j\omega$ axis, the zero begins to add positive phase, and $\beta H(s) = \beta A_0(1 + s/\omega_z)/s$ approaches a real value equal to $\beta A_0/\omega_z$ at $s = +j\infty$. In contrast to the case illustrated in Fig. 10.67, this contour is deflected away from the origin by the zero.


jω
$\varepsilon e^{j\phi}$
σ
Im{βH(s)}
s = +j∞
$\beta\frac{A_0}{\omega_z}$
$\beta\frac{A_0}{\varepsilon}$
Re{βH(s)}
s = ε
$\beta H(\varepsilon e^{j90°})$

(a) (b)

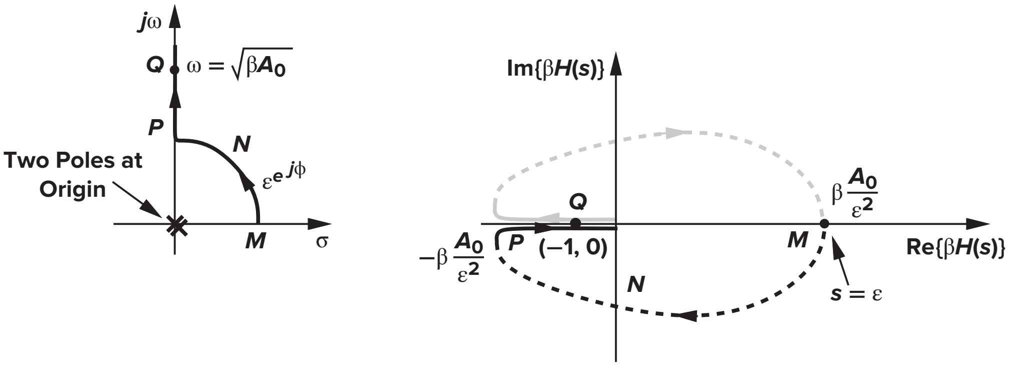
Figure 10.68

Example 10.13

A negative-feedback loop employs two ideal integrators, i.e., $H(s) = A_0/s^2$. Study the closed-loop stability of the system.

Solution

We begin with $s = \epsilon\exp(j\phi)$, $\phi = 0$, and hence $\beta H = \beta A_0/\epsilon^2$ (Fig. 10.69). As ϕ goes to $+45°$ (point N), $\beta H(s) = \beta(A_0/\epsilon^2)\exp(-2j\phi)$ rotates by $-90°$, still at a very large radius. For $\phi = +90°$ (point P), $\beta H(s)$ returns to the real axis. Now, as $s = +j\omega$ travels upward, the angle remains unchanged, but the magnitude, $\beta|H(j\omega)| = \beta A_0/\omega^2$, falls. That is, βH continues on the real axis toward the origin, *passing* through $(-1, 0)$ at $\omega = \sqrt{\beta A_0}$. The closed-loop system therefore contains two poles *on* the $j\omega$ axis because it crosses $(-1, 0)$ twice. After all, we can write $H(s)/[1 + \beta H(s)] = A_0/(s^2 + \beta A_0)$, observing two imaginary poles at $\pm j\sqrt{\beta A_0}$. Thus, a two-pole system *can* oscillate if it has two poles at the origin.


jω
Q
$\omega = \sqrt{\beta A_0}$
P
N
Two Poles at Origin
$\varepsilon e^{j\phi}$
M
σ
Im{βH(s)}
Q
$\beta\frac{A_0}{\varepsilon^2}$
$-\beta\frac{A_0}{\varepsilon^2}$
P
(−1, 0)
M
Re{βH(s)}
N
s = ε

(a) (b)

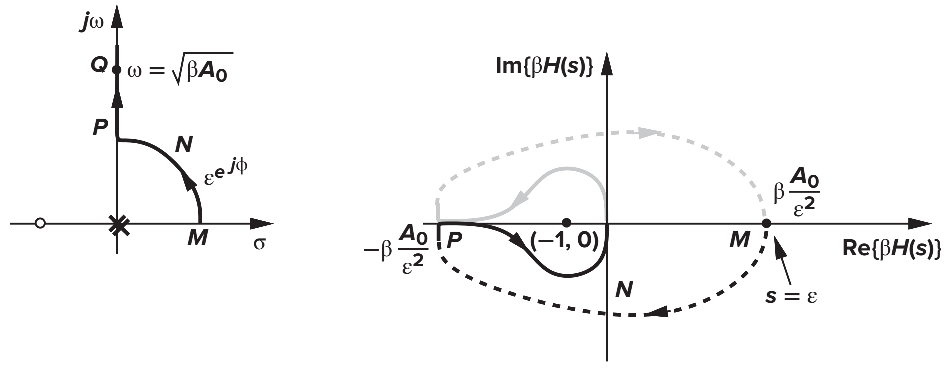
Figure 10.69

Example 10.14

Repeat the previous example if a zero is added to one of the integrators, i.e., $H(s) = A_0(1 + s/\omega_z)/s^2$.

Solution

With $s = \epsilon \exp(j\phi)$ and $\phi = 0$, we have $\beta H \approx \beta A_0/\epsilon^2$. The behavior of βH is similar to that in the previous example up to point P (Fig. 10.70). As $s = +j\omega$ travels upward, the zero begins to contribute appreciable phase and $|\beta H(j\omega)| = \beta A_0\sqrt{1+\omega^2/\omega_z^2}/\omega^2$ continues to fall. As $s \to +j\infty$, $\angle\beta H$ approaches $-90°$, suggesting that the βH contour must reach the origin at an angle of $-90°$. As shown in Fig. 10.70, the zero ensures that βH does not cross or encircle $(-1, 0)$, stabilizing the closed-loop system.



(a) (b)

Figure 10.70

The foregoing example sheds light on a common paradox that the Bode plots of the two-integrator system conjure up, especially in the context of phase-locked loops (Chapter 16). As shown in Fig. 10.71, it appears that the above closed-loop system is capable of oscillation at a frequency ω_1, at which $|\beta H|$ is greater than unity and $\angle\beta H = -180°$. But we observe from the Nyquist plot in Fig. 10.70 or from $\beta H(j\omega_1) = -\beta A_0(1 + j\omega_1/\omega_z)/\omega_1^2$ that, owing to the zero, the phase of βH *never* reaches exactly 180°. That is, at point P in Fig. 10.70, the infinitesimal phase contributed by the zero causes $|\angle\beta H|$ to be less than 180°. Similarly, even though the approximate Bode plots of Fig. 10.71 suggest a phase of 180° for $\omega_1 \ll \omega_z$, in reality this amount of phase occurs only at $\omega = 0$. The story, however, does not end here. The next section provides a more fundamental understanding.

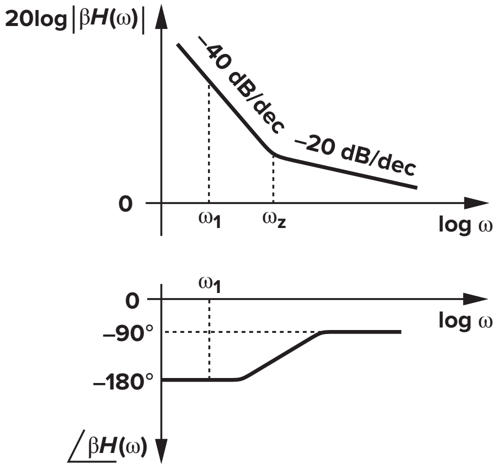


Figure 10.71 Bode plots of a system with two poles at origin and one zero.

10.8.7 Systems with Multiple 180° Crossings

Consider a system whose loop transmission has three poles and two zeros as shown in Fig. 10.72(a). Illustrated in Fig. 10.72(b), the Bode plots reveal that the phase crosses $-180°$ *twice* while the gain remains higher than unity. Is this system stable when placed in a negative-feedback loop?

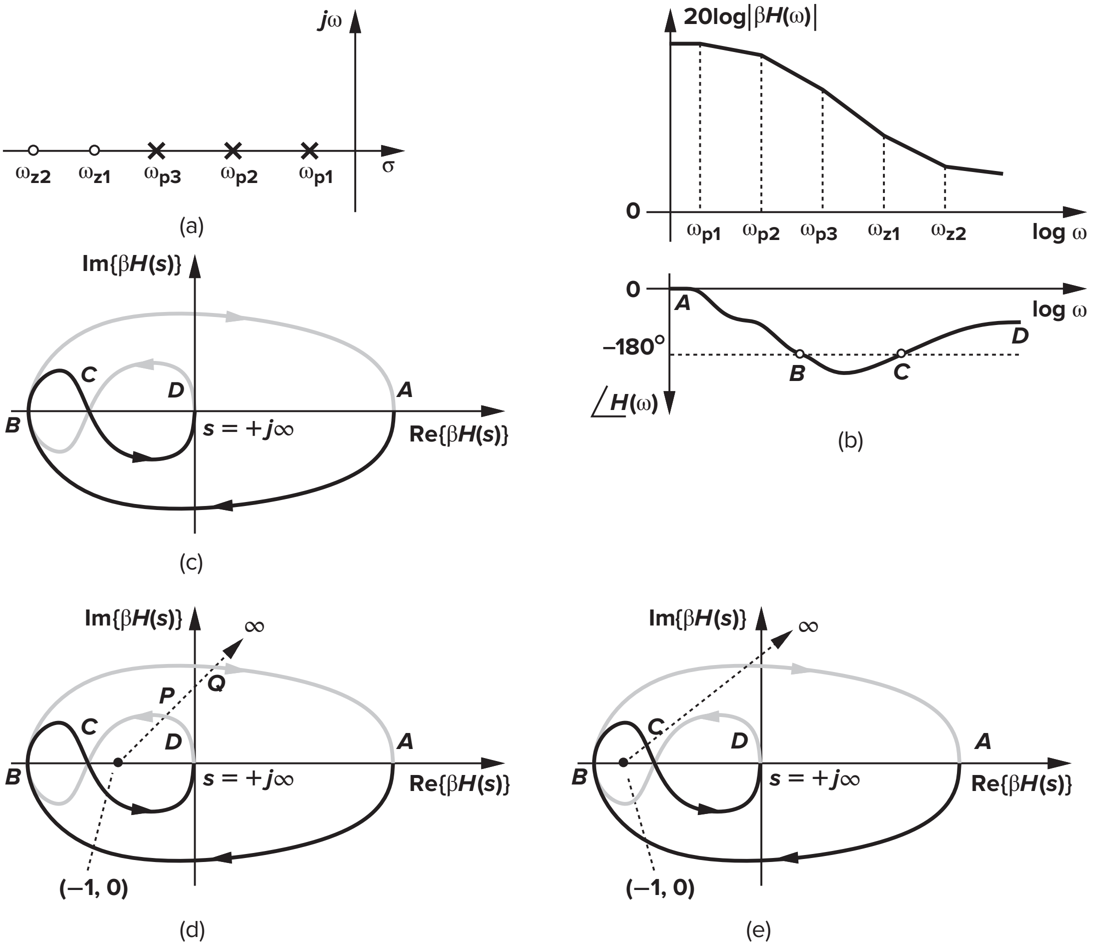


Figure 10.72 (a) System with three poles and two zeros, (b) Bode plots, (c) βH contour, (d) case where C is to the left of $(-1, 0)$, and (e) case where C is to the right of $(-1, 0)$.

In the absence of the zeros, the βH contour crosses $-180°$ and approaches the origin at an angle of $-270°$ [Fig. 10.65(a)]. We now construct the Nyquist plot as follows. As we begin from the origin in Fig. 10.72(a) and travel up on the $j\omega$ axis, the phase starts at zero and the βH contour at point A in Fig. 10.72(c). Due to the higher number of poles, $\angle\beta H$ becomes negative, reaching $-180°$ (point B) for some value of $j\omega$. As ω increases further, $\angle\beta H$ becomes more negative, but, due to the contribution of the zeros, it deflects, forcing the βH contour to return to the real axis (point C). The phase then becomes more positive and, for $\omega \to \infty$, approaches $-90°$ (point D). Drawing the other symmetric half, we distinguish between two cases.

1. Point C is to the left of $(-1, 0)$ [Fig. 10.72(d)]; if we draw a line from $(-1, 0)$ to infinity, it crosses the contour twice (at P and Q), but the contour has opposite directions at these two points. The

closed-loop system therefore has no poles in the RHP. Since both B and C are to the right of $(-1, 0)$, this case corresponds to the Bode plots of Fig. 10.72(b).

2. Point $(-1, 0)$ lies between C and B [Fig. 10.72(e)]. In this case, the system is unstable.

We summarize the above results as follows. If $\angle\beta H$ crosses 180° an even (odd) number of times while $|\beta H| > 1$, then the system is stable (unstable).

References

[1] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed. New York: Wiley, 1993.

[2] W. C. Black, D. J. Allstot, and R. A. Reed, "A High Performance Low Power CMOS Channel Filter," *IEEE J. of Solid-State Circuits*, vol. 15, pp. 929–938, December 1983.

[3] R. M. Ziazadeh, H.-T. Ng, and D. J. Allstot, "A Multistage Amplifier Topology with Embedded Tracking Compensation," *CICC Proc.*, pp. 361–364, May 1998.

[4] B. K. Ahuja, "An Improved Frequency Compensation Technique for CMOS Operational Amplifiers," *IEEE J. of Solid-State Circuits*, vol. 18, pp. 629–633, December 1983.

[5] P. R. Gray and R. G. Meyer, "MOS Operational Amplifier Design—A Tutorial Overview," *IEEE J. of Solid-State Circuits*, vol. 17, pp. 969–982, December 1982.

[6] B. Y. Kamath, R. G. Meyer, and P. R. Gray, "Relationship Between Frequency Response and Settling Time of Operational Amplifiers," *IEEE J. of Solid-State Circuits*, vol. 9, pp. 347–352, December 1974.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

10.1. An amplifier with a forward gain of A_0 and two poles at 10 MHz and 500 MHz is placed in a unity-gain feedback loop. Calculate A_0 for a phase margin of 60°.

10.2. An amplifier with a forward gain of A_0 has two coincident poles at ω_p. Calculate the maximum value of A_0 for a 60° phase margin with a closed-loop gain of **(a)** unity and **(b)** 4.

10.3. An amplifier has a forward gain of $A_0 = 1000$ and two poles at ω_{p1} and ω_{p2}. For $\omega_{p1} = 1$ MHz, calculate the phase margin of a unity-gain feedback loop if **(a)** $\omega_{p2} = 2\omega_{p1}$ and **(b)** $\omega_{p2} = 4\omega_{p1}$.

10.4. A unity-gain closed-loop amplifier exhibits a frequency peaking of 50% in the vicinity of the gain crossover. What is the phase margin?

10.5. Consider the transimpedance amplifier shown in Fig. 10.73, where $R_D = 1\ \text{k}\Omega$, $R_F = 10\ \text{k}\Omega$, $g_{m1} = g_{m2} = 1/(100\ \Omega)$, and $C_A = C_X = C_Y = 100$ fF. Neglecting all other capacitances and assuming that $\lambda = \gamma = 0$, compute the phase margin of the circuit. (Hint: break the loop at node X.)

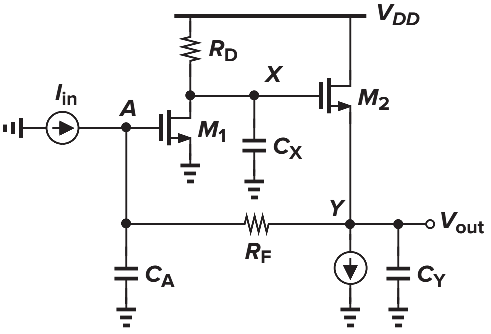


Figure 10.73

10.6. In Problem 10.5, what is the phase margin if R_D is increased to 2 kΩ?

10.7. If the phase margin required of the amplifier of Problem 10.5 is 45°, what is the maximum value of (a) C_Y, (b) C_A, and (c) C_X while the other two capacitances remain constant?

10.8. Prove that the zero of the circuit shown in Fig. 10.32 is given by Eq. (10.30). Apply the technique illustrated in Fig. 6.18.

10.9. Consider the amplifier of Fig. 10.74, where $(W/L)_{1-4} = 50/0.5$ and $I_{SS} = I_1 = 0.5$ mA.

(a) Estimate the poles at nodes X and Y by multiplying the small-signal resistance and capacitance to ground. Assume that $C_X = C_Y = 0.5$ pF. What is the phase margin for unity-gain feedback?

(b) If $C_X = 0.5$ pF, what is the maximum tolerable value of C_Y that yields a phase margin of 60° for unity-gain feedback?

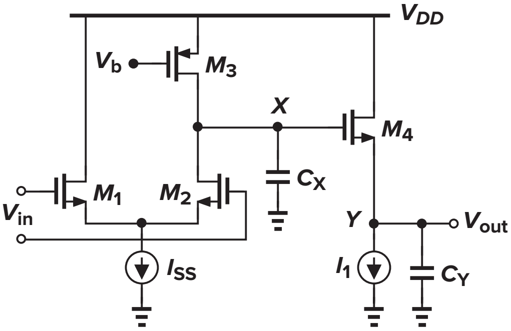


Figure 10.74

10.10. Estimate the slew rate of the op amp of Problem 10.9 for both parts **(a)** and **(b)**.

10.11. In the two-stage op amp of Fig. 10.75, $W/L = 50/0.5$ for all transistors except for $M_{5,6}$, for which $W/L = 60/0.5$. Also, $I_{SS} = 0.25$ mA and each output branch is biased at 1 mA.

(a) Determine the CM level at nodes X and Y.

(b) Calculate the maximum output voltage swing.

(c) If each output is loaded by a 1-pF capacitor, compensate the op amp by Miller multiplication for a phase margin of 60° in unity-gain feedback. Calculate the pole and zero positions after compensation.

(d) Calculate the resistance that must be placed in series with the compensation capacitors to position the zero atop the nondominant pole.

(e) Determine the slew rate.

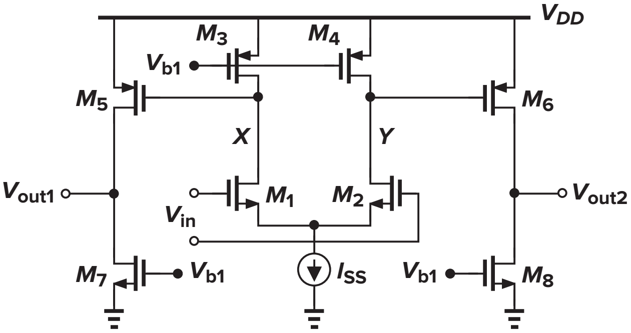


Figure 10.75

10.12. In Problem 10.11, the pole-zero cancellation resistor is implemented with a PMOS device as in Fig. 10.34. Calculate the dimensions of M_{13}–M_{15} if $I_1 = 100\ \mu$A.

10.13. Calculate the input-referred thermal noise voltage of the op amp shown in Fig. 10.75.

10.14. Figure 10.76 depicts a transimpedance amplifier employing voltage-current feedback. Note that the feedback factor may exceed unity because of M_3. Assume that I_1–I_3 are ideal, $I_1 = I_2 = 1$ mA, $I_3 = 10\ \mu$A, $(W/L)_{1,2} = 50/0.5$, and $(W/L)_3 = 5/0.5$.

(a) Breaking the loop at the gate of M_3, estimate the poles of the open-loop transfer function.

(b) If the circuit is compensated by adding a capacitor C_C between the gate and the drain of M_1, what value of C_C achieves a phase margin of 60°? Determine the poles after compensation.

(c) What resistance must be placed in series with C_C to position the zero of the output stage atop the first nondominant pole?

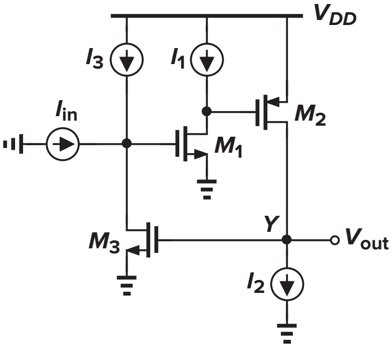


Figure 10.76

10.15. Repeat Problem 10.14 if the output node is loaded by a 0.5-pF capacitor.

10.16. Suppose that in the circuit of Fig. 10.76, a large negative input current is applied such that M_1 turns off momentarily. What is the slew rate at the output?

10.17. Explain why, in the circuit of Fig. 10.76, the compensation capacitor should not be placed between the gate and the drain of M_2 or M_3.

10.18. Determine the input-referred noise current of the circuit shown in Fig. 10.76 and described in Problem 10.14.

10.19. The cancellation of a pole by a zero, e.g., in a two-stage op amp, entails an issue called the "doublet problem" [5, 6]. If the pole and the zero do not exactly coincide, we say that they constitute a doublet. The step response of feedback circuits in the presence of doublets is of great interest. Suppose the open-loop transfer function of a two-stage op amp is expressed as

$$H_{open}(s) = \frac{A_0\left(1+\dfrac{s}{\omega_z}\right)}{\left(1+\dfrac{s}{\omega_{p1}}\right)\left(1+\dfrac{s}{\omega_{p2}}\right)} \tag{10.59}$$

Ideally, $\omega_z = \omega_{p2}$ and the feedback circuit exhibits a first-order behavior, i.e., its step response contains a single time constant and no overshoot.

(a) Prove that the transfer function of the amplifier in a unity-gain feedback loop is given by

$$H_{closed}(s) = \frac{A_0\left(1+\dfrac{s}{\omega_z}\right)}{\dfrac{s^2}{\omega_{p1}\omega_{p2}}+\left(\dfrac{1}{\omega_{p1}}+\dfrac{1}{\omega_{p2}}+\dfrac{A_0}{\omega_z}\right)s+A_0+1} \tag{10.60}$$

(b) Determine the two poles of $H_{closed}(s)$, assuming they are widely spaced.

(c) Assuming $\omega_z \approx \omega_{p2}$ and $\omega_{p2} \ll (1 + A_0)\omega_{p1}$, write $H_{closed}(s)$ in the form

$$H_{closed}(s) = \frac{A\left(1 + \dfrac{s}{\omega_z}\right)}{\left(1 + \dfrac{s}{\omega_{pA}}\right)\left(1 + \dfrac{s}{\omega_{pB}}\right)} \tag{10.61}$$

and determine the small-signal step response of the closed-loop amplifier.

(d) Prove that the step response contains an exponential term of the form $(1 - \omega_z/\omega_{p2})\exp(-\omega_{p2}t)$. This is an important result, indicating that if the zero does not exactly cancel the pole, the step response exhibits an exponential with an amplitude proportional to $1 - \omega_z/\omega_{p2}$ (which depends on the mismatch between ω_z and ω_{p2}) and a time constant of $1/\omega_z$.

10.20. Using the results of the previous problem, determine the step response of the amplifier described in Problem 10.11 with **(a)** perfect pole-zero cancellation and **(b)** 10% mismatch between the pole and the zero magnitudes.

10.21. It is possible to raise the voltage gain of a folded-cascode op amp by adding a secondary path. As shown in Fig. 10.77 by the gray section, the input signal can also travel through a differential pair with current-source loads, I_1 and I_2, and drive the current sources in the original op amp. Of course, nodes X and Y exhibit a relatively high impedance, thus contributing a pole that significantly degrades the phase margin.

(a) Neglecting channel-length modulation in I_1 and I_2, determine the low-frequency gain of the op amp.

(b) Considering only the capacitances at X, Y, P, Q, and the output nodes, compute the overall transfer function. Is it possible for the zero to cancel one of the poles?

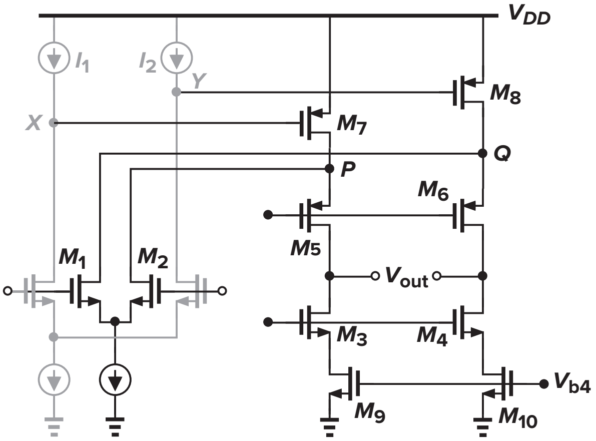


Figure 10.77

10.22. Consider the circuit of Fig. 10.37(b) and assume that $I_{in} = I_{SS}u(t)$. Also, assume that a load capacitance of C_L is tied from the drain to ground. Write a KCL at the output node and derive a differential equation in terms of V_{out}. Taking the Laplace transform and using partial fractions, prove that

$$V_{out}(t) = \frac{I_{SS}}{C_F}tu(t) - \frac{I_{SS}}{g_m}\left(1 + \frac{C_L}{C_F}\right)u(t) - \frac{I_{SS}}{g_m}\left(1 + \frac{C_L}{C_F}\right)\exp\frac{-t}{\tau}u(t) \tag{10.62}$$

where $\tau = C_L/g_m$. Plot these three terms as a function of time and determine the time at which $V_{out}(t)$ reaches a minimum. This result indicates that the output initially falls and then assumes a ramp behavior.

10.23. A two-stage op amp is compensated for a phase margin of 60° with $\beta = 1$. If β is reduced to $\beta_1 < 1$, determine the new phase margin.

CHAPTER

11

Nanometer Design Studies

The previous chapters of this book have taken us on a "scenic" route through the world of analog circuits, presenting important concepts and useful topologies. We have occasionally made *design* efforts, but only on a small scale. In this chapter, we embark upon two comprehensive designs so as to appreciate the mindset that an analog designer must uphold and the multitude of tasks that he or she must complete for a given circuit. The designs are carried out in 40-nm CMOS technology with a 1-V supply. The reader is encouraged to review the op amp design examples in Chapter 9 before starting this chapter.

We begin with a brief look at the imperfections of nanometer devices and the design procedures to achieve certain transistor parameters. We then delve into the design of an op amp and, through simulations, optimize its performance. Finally, we deal with the design of a high-speed, high-precision amplifier and pursue various techniques to achieve a low power dissipation.

11.1 ■ Transistor Design Considerations

In Chapter 2, we studied the basic operation of MOSFETs and included a few second-order effects. Our investigation has produced a large-signal model (consisting of the triode-region quadratic equation and the saturation-region square-law relation), which becomes necessary in two cases: (1) when the transistor experiences large voltage (or current) changes due to the input or output signals, disobeying the small-signal model, or (2) when the transistor must be biased, requiring certain terminal voltages so as to carry a specified current. In analog design, the former case occurs occasionally, while the latter almost always.

The large-signal behavior of nanometer MOSFETs significantly departs from the "long-channel" model that we have developed. As a result of technology scaling, i.e., the shrinkage of MOS dimensions, several effects besides those studied in Chapter 2 manifest themselves, thereby altering the I/V characteristics. As an example, Fig. 11.1 plots the actual I_D-V_{DS} characteristics of an NFET with $W/L = 5\ \mu$m/40 nm and $V_{TH} \approx 300$ mV (using a BSIM4 model) against a "best-fit" long-channel square-law approximation. We observe that the two diverge considerably. Thus, even if we are not interested in the large-signal analysis of a circuit, we still face the problem of bias calculations using the square-law model.

In this section, we briefly consider a few "short-channel" effects that make the long-channel model inaccurate. A detailed treatment of short-channel effects is deferred to Chapter 17. It is important to note that the MOS small-signal model developed in Chapter 2 still holds for short-channel devices and, as seen throughout this book, suffices for the initial analysis of many analog circuit blocks. However, the expressions relating g_m and r_O to the bias conditions must be revised.

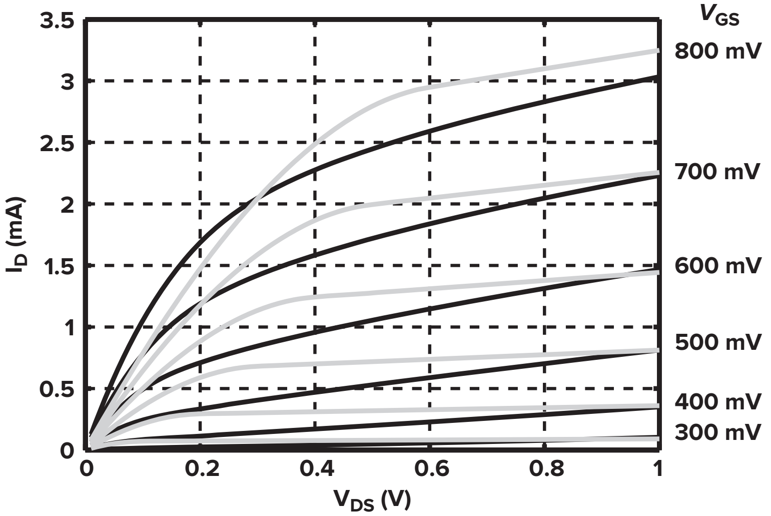


Figure 11.1 I-V characteristics of an actual 5-μm/40-nm device (black curves) and a best-fit square-law device (gray curves). (V_{GS} is incremented from 300 mV to 800 mV in 100-mV steps.)

The characteristics shown in Fig. 11.1 exhibit severe channel-length modulation for the actual 40-nm devices, making it difficult to distinguish between the triode and saturation regions. But we can associate a "knee" point with each curve as a rough boundary. Figure 11.2 plots the actual 40-nm device characteristics for a narrower V_{GS} range, namely, $V_{GS} - V_{TH} = 50$ mV, 100 mV, $\cdots$, 350 mV. We observe knee points below $V_{DS} = 0.2$ V. (Here, $W = 5$ μm and $V_{TH} \approx 200$ mV.)

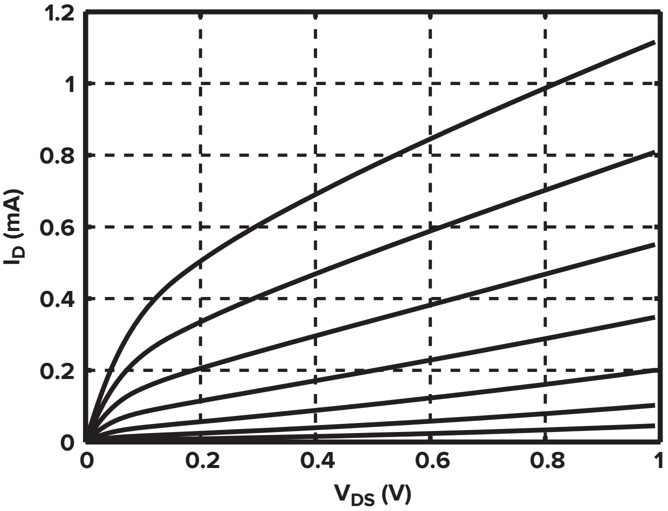


Figure 11.2 I-V characteristics of a 5-μm/40-nm device for $V_{GS} - V_{TH} = 50, \cdots, 350$ mV.

11.2 ■ Deep-Submicron Effects

Among various short-channel effects, two are particularly important at this stage of our studies; both relate to the mobility of the carriers in the channel. Recall that we have assumed that the carrier velocity is given by $v = \mu E$, where E denotes the electric field. We revisit this assumption here.

Velocity Saturation In a MOSFET, as V_{DS} and hence the electric field along the source-drain path increase, v does not rise proportionally (Fig. 11.3).

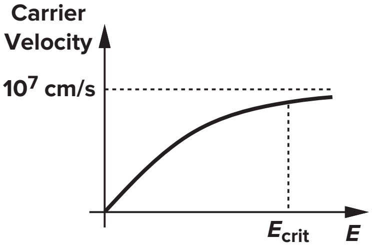


Figure 11.3 Velocity saturation at high electric fields.

We say that the carriers experience "velocity saturation" or, equivalently, that the mobility (the slope of v versus E) *falls*. This effect has arisen because the length of MOSFETs has shrunk from, say, 1 μm to 40 nm (a factor of 25) while the allowable drain-source voltage has decreased from 5 V to about 1 V. The lateral electric field has thus exceeded E_{crit} ($\approx$ 1 V/μm) in Fig. 11.3.

We deal with the modeling of velocity saturation in Chapter 17, but let us consider an extreme case here: suppose the charge carriers reach the saturated velocity, v_{sat}, as soon as they depart from the source. Since $I = Q_d \cdot v$, where Q_d is the charge density (per unit length) and given by $WC_{ox}(V_{GS} - V_{TH})$, we have

$$I_D = WC_{ox}(V_{GS} - V_{TH})v_{sat} \tag{11.1}$$

Extreme velocity saturation therefore creates three departures from the square-law behavior. First, the device carries a current that is *linearly* proportional to the overdrive and independent of the channel length.[1] Second, I_D reaches saturation even for $V_{DS} < V_{GS} - V_{TH}$ (Fig. 11.4). As evident in Fig. 11.2, the knee points occur at relatively small V_{DS}'s even as the overdrive reaches 350 mV. Third, the transconductance of a fully velocity-saturated MOSFET emerges as

$$g_m = \frac{\partial I_D}{\partial V_{GS}}|_{V_{DSconst}} \tag{11.2}$$

$$= WC_{ox}v_{sat} \tag{11.3}$$

a relatively *constant* value versus I_D or V_{GS}. For example, in the plots of Fig. 11.2, the change in I_D is fairly constant as the overdrive increments from 250 mV to 300 mV and from 300 mV to 350 mV.

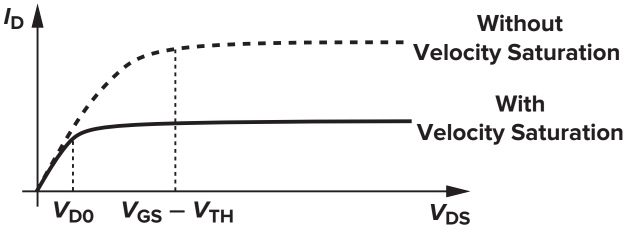


Figure 11.4 Premature saturation of drain current due to velocity saturation.

Mobility Degradation with Vertical Field The mobility of the charge carriers in the channel also declines as the gate-source voltage and the *vertical* field increase (Fig. 11.5).

[1] So long as L is small enough and V_{DS} large enough to cause velocity saturation.

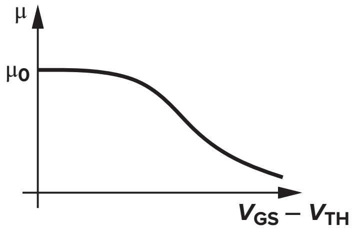


Figure 11.5 Reduction of mobility due to vertical electric field.

What is the impact of this mobility degradation on the device transconductance? We intuitively expect that g_m no longer follows the linear relationship, $g_m = \mu C_{ox}(W/L)(V_{GS} - V_{TH})$, with the overdrive voltage. Figure 11.6 displays this behavior for the 5-μm/40-nm NFET mentioned above.

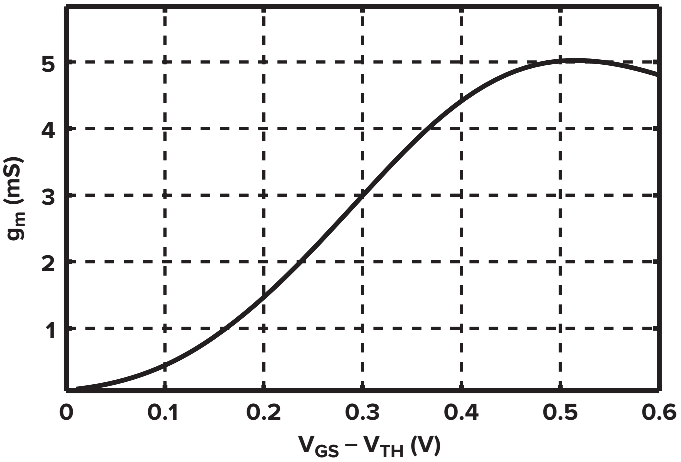


Figure 11.6 Transconductance as a function of overdrive voltage.

▶ **Example 11.1**

We approximate the mobility plot of Fig. 11.5 by

$$\mu = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})} \tag{11.4}$$

where θ is a proportionality factor with a dimension of (voltage)$^{-1}$. Determine the transconductance of a MOSFET that suffers from this type of mobility degradation.

Solution

We write

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1 + \theta(V_{GS} - V_{TH})}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{11.5}$$

and hence

$$g_m = \mu_0 C_{ox}\frac{W}{L}\frac{(\theta/2)(V_{GS} - V_{TH})^2 + V_{GS} - V_{TH}}{[1 + \theta(V_{GS} - V_{TH})]^2} \tag{11.6}$$

As expected, for $\theta(V_{GS} - V_{TH}) \ll 1$, we have $g_m \approx \mu_0 C_{ox}(W/L)(V_{GS} - V_{TH})$. At the other extreme, if $(V_{GS} - V_{TH}) \gg 2/\theta$, then g_m approaches a constant value: $g_m \approx (1/2)\mu_0 C_{ox}(W/L)/\theta$. ◀

In the general case, the degradation of the mobility due to both lateral and vertical fields (V_{DS} and V_{GS}, respectively) must be considered. Nonetheless, the simple results derived above suffice for most of our studies in analog design.

11.3 ■ Transconductance Scaling

Device transconductances manifest themselves in almost every analog circuit. Suppose a transistor operates in the saturation region but does not provide the required transconductance. The g_m equations in Chapter 2,

$$g_m = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH}) \tag{11.7}$$

$$= \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} \tag{11.8}$$

$$= \frac{2I_D}{V_{GS} - V_{TH}} \tag{11.9}$$

suggest that adjustments in three parameters, namely, W/L, $V_{GS} - V_{TH}$, or I_D, can scale g_m. We study these scenarios, assuming for now a long-channel device and hence $I_D \approx (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$. That is, $V_{GS} - V_{TH} \approx \sqrt{2I_D/(\mu_n C_{ox} W/L)}$. In each case, we keep one of the parameters constant and vary the other two.

From (11.7), we can increase W/L while keeping $V_{GS} - V_{TH}$ constant. In this case, both g_m and I_D linearly scale with W/L (why?) [Fig. 11.7(a)], and so does the power consumption. Alternatively, we

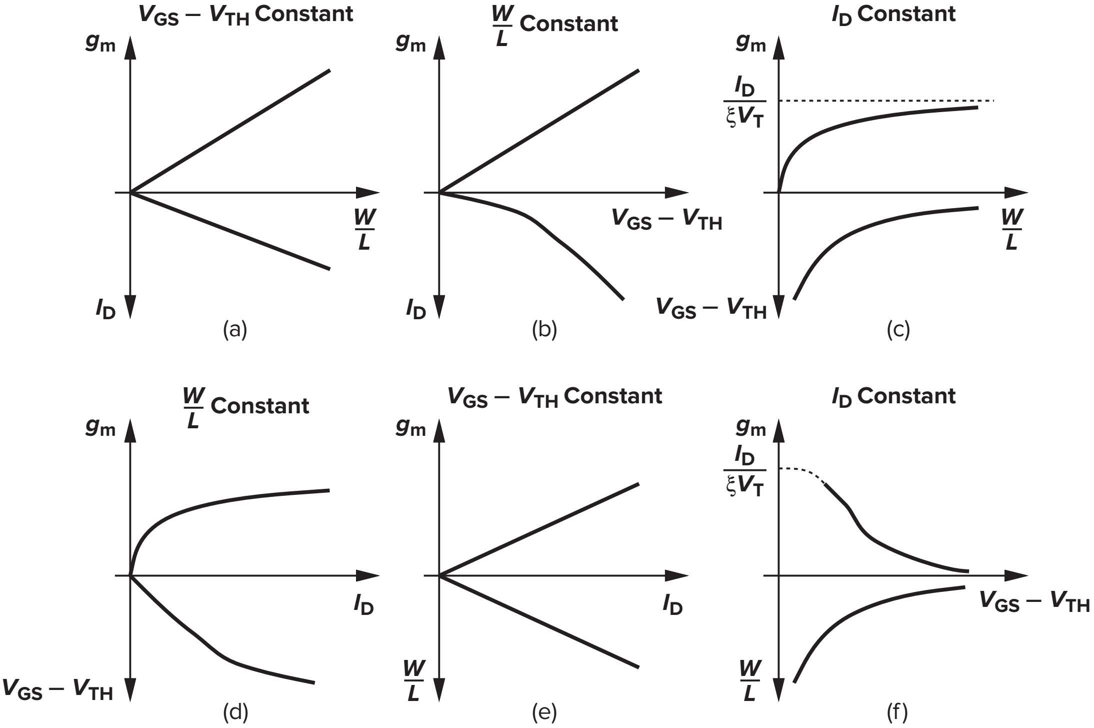


Figure 11.7 Dependence of (a) g_m and I_D upon W/L, (b) g_m and I_D upon $V_{GS} - V_{TH}$, (c) g_m and $V_{GS} - V_{TH}$ upon W/L, (d) g_m and $V_{GS} - V_{TH}$ upon I_D, (e) g_m and W/L upon I_D, and (f) g_m and W/L upon $V_{GS} - V_{TH}$.

can increase $V_{GS} - V_{TH}$ but keep W/L constant [Fig. 11.7(b)], thus requiring a higher drain current. In the former case, the device capacitances rise whereas in the latter, $V_{DS,min}$ increases. In this chapter, we use the notations $V_{DS,min}$, $V_{D,sat}$, and $V_{GS} - V_{TH}$ interchangeably.

From (11.8), we can raise W/L while I_D is constant [Fig. 11.7(c)], as a result of which $V_{GS} - V_{TH}$ is lowered (why?). Due to subthreshold conduction, however, g_m does not climb indefinitely in this case. If we keep W/L constant and increase I_D [Fig. 11.7(d)], then $V_{GS} - V_{TH}$ and hence $V_{DS,min}$ must rise.

From (11.9), we can increase I_D while $V_{GS} - V_{TH}$ is constant [Fig. 11.7(e)]. This requires that W/L increase. Alternatively, we can lower $V_{GS} - V_{TH}$ and keep I_D constant [Fig. 11.7(f)], which means that W/L must increase. For $V_{GS} - V_{TH} \approx 0$, the device enters the subthreshold region and $g_m \approx I_D/(\xi V_T)$. In both cases, the device capacitances climb.

Let us now reconsider the foregoing six scenarios for nanometer devices. We note that the plots in Fig. 11.7 still hold qualitatively, but the g_m and overdrive equations are more complex. The case of Fig. 11.7(a) is particularly interesting and useful and is studied further in the following example.

▶ Example 11.2

The linear scaling of g_m and I_D with W/L, shown in Fig. 11.7(a), holds *regardless* of the transistor characteristics. Explain why.

Solution

Consider, as an example, two identical transistors connected in parallel (Fig. 11.8), each having a transconductance of g_m. If V_{GS} changes by ΔV, then the drain current of each device changes by $g_m \Delta V$, and hence the current of the composite device changes by $2g_m \Delta V$. That is, the parallel combination exhibits a transconductance of $2g_m$. We conclude that increasing *both* the width and the drain current of the transistor by a factor of K (> 1) is equivalent to placing K transistors in parallel and raising the g_m by a factor of K. We say that the scaling preserves the device "current density" (I_D/W) in this case. Note that the bias overdrive voltage remains constant in this scenario, and so does the g_m/I_D ratio. The latter property proves useful in our studies.

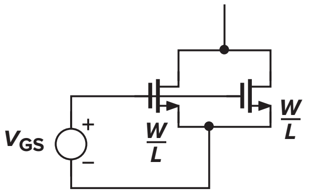


Figure 11.8

◀

Of the six scenarios depicted in Fig. 11.7, which ones are more common in practice? Since modern analog circuits must operate with low supplies (around 1 V), we often limit $V_{GS} - V_{TH}$ to a few hundred millivolts. Thus, to obtain a certain transconductance, we first keep increasing the width [Fig. 11.7(c)] to the extent that it raises the g_m significantly. As g_m approaches a constant value (in the subthreshold region), the width is no longer a determining factor, leaving the drain current as the only parameter that can increase the g_m [Fig. 11.7(d)]. However, as we increase I_D in this case, $V_{GS} - V_{TH}$ may exceed the allotted value, forcing us to resort to the scenario in Fig. 11.7(e) [which is equivalent to that in Fig. 11.7(a)]. These trials and errors seem rather haphazard, but do not despair! The remainder of this section is dedicated to developing a methodical approach to transistor design. We begin with an important example.

▶ Example 11.3

A transistor having an aspect ratio of $(W/L)_{REF}$ exhibits the g_m-I_D characteristic shown in Fig. 11.9(a).

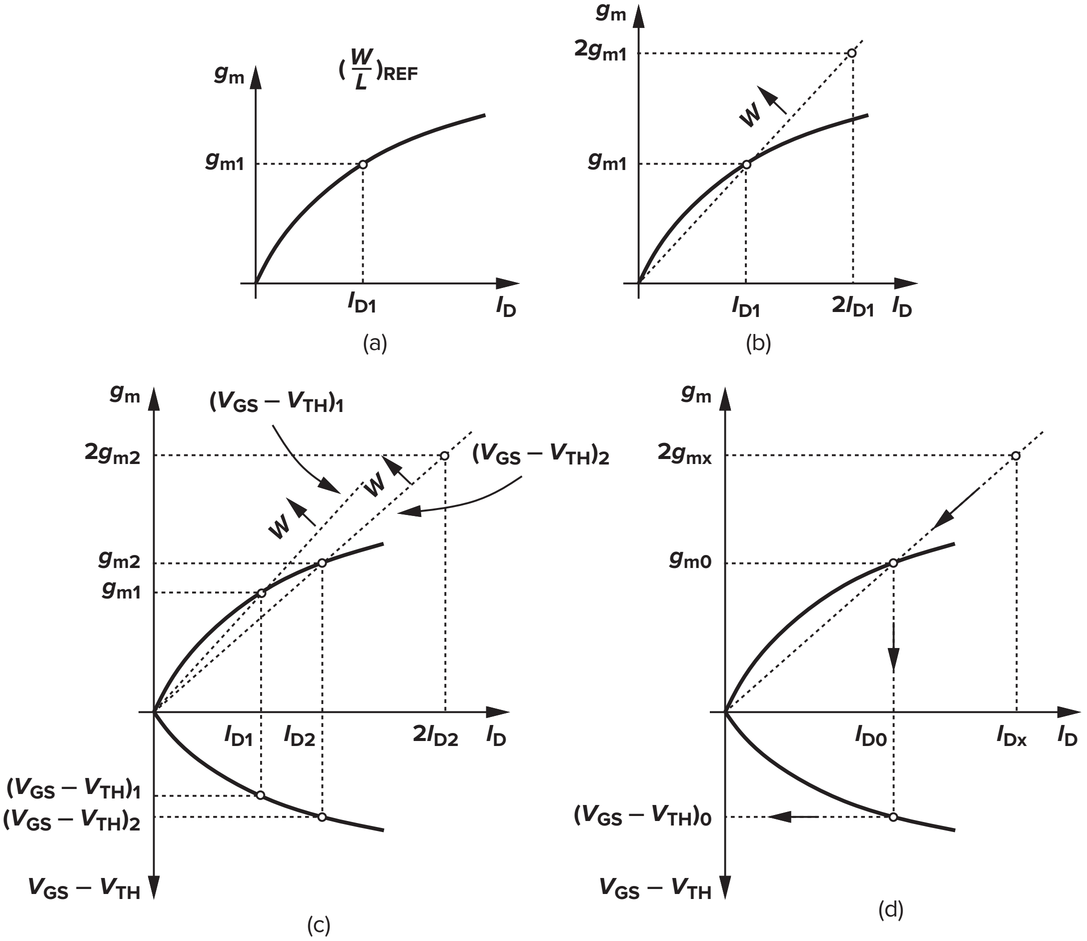


Figure 11.9

(I) Suppose the device is first biased at $I_D = I_{D1}$. What happens to the transconductance and the drain current if the width is doubled while $V_{GS} - V_{TH}$ remains constant? (II) Repeat (I) if we begin with a greater overdrive. (III) We wish to obtain a transconductance of g_{mx} at a drain current of I_{Dx}. How should the transistor be scaled?

Solution

(I) With a constant $V_{GS} - V_{TH}$, doubling the width also doubles the transconductance and the drain current (Example 11.2). Since g_m/I_D is constant, to obtain this point on the g_m-I_D plane, we pass a straight line through the origin and (I_{D1}, g_{m1}), continuing to reach $(2I_{D1}, 2g_{m1})$ [Fig. 11.9(b)]. Thus, all (I_D, g_m) combinations resulting from the scaling of W fall on this line if the overdrive is fixed.

(II) If we begin with a greater overdrive, $(V_{GS} - V_{TH})_2$, the (I_D, g_m) point is located elsewhere, at (I_{D2}, g_{m2}), on the characteristic [Fig. 11.9(c)]. We again draw a straight line through the origin and (I_{D2}, g_{m2}) and continue to $(2I_{D2}, 2g_{m2})$. Thus, each such line in the g_m-I_D plane represents the possible (I_D, g_m) combinations that can be obtained by scaling W for a given overdrive.

(III) We draw a line through the origin and the point (I_{Dx}, g_{mx}) [Fig. 11.9(d)]. The intersection of the line and the g_m plot yields a "reference" point specifying the proper overdrive voltage, $(V_{GS} - V_{TH})_0$, and an acceptable (I_D, g_m) combination, (I_{D0}, g_{m0}). If the width is scaled up by a factor of g_{mx}/g_{m0} $(= I_{Dx}/I_{D0})$, and the overdrive remains equal to $(V_{GS} - V_{TH})_0$, then the desired transconductance and current are obtained.

◀

11.4 ■ Transistor Design

The reader may have noticed by now that a given transistor in a circuit is characterized by a multitude of parameters. In this section, we assume that transistors operate in saturation and focus on two bias quantities, I_D and $V_{GS} - V_{TH}$ ($= V_{DS,min}$), one small-signal parameter, g_m, and one physical parameter, W/L. A typical transistor design problem specifies two of the first three and seeks the other two (Table 11.1). We wish to develop methodical approaches to computing these two parameters for nanometer devices. While not listed here explicitly, the output resistance, r_O, also proves important in many circuits and is eventually included in our studies in Sec. 11.4.5.

Table 11.1 Three scenarios encountered in transistor design.

	Case I	Case II	Case III
Given	I_D, $V_{DS,min}$	g_m, I_D	g_m, $V_{DS,min}$
To Be Determined	$\frac{W}{L}$, g_m	$\frac{W}{L}$, $V_{DS,min}$	$\frac{W}{L}$, I_D
Design Revision	g_m insufficient;	$V_{DS,min}$ too large;	I_D too large;
	Raise I_D and $\frac{W}{L}$	Raise $\frac{W}{L}$	Raise $\frac{W}{L}$; Lower $V_{GS} - V_{TH}$

The reader may recognize that the design problems shown in Table 11.1 are "overconstrained," i.e., the two given parameters inevitably lead to certain values for the other two—even though the results may not always be *desirable*. For example, a known I_D and $V_{DS,min}$ directly give a value for g_m that may not be sufficient for a particular circuit. In such a case, we must modify the design as prescribed in the last row of the table. We will elaborate on this row in the design studies to be followed, but let us make some preliminary remarks here. In Case I, an insufficient g_m would require a higher I_D (possibly exceeding a power budget) and a greater W/L (to satisfy the specified $V_{DS,min}$). In Case II, the given I_D and g_m may yield an unacceptably large $V_{DS,min}$, thereby dictating a greater W/L. In Case III, the necessary I_D may be excessive, demanding a greater W/L and a smaller overdrive.[2]

11.4.1 Design for Given I_D and $V_{DS,min}$

A common situation that arises in analog design is as follows. For a particular transistor in the circuit, we have chosen a bias current (perhaps according to a power budget) and a minimum V_{DS} (perhaps according to the voltage headroom, i.e., the restrictions imposed by the supply voltage and the required swings).[3] We now wish to determine the dimensions and the transconductance of the device, recognizing that the square-law equations are inaccurate. Of course, with the transistor models available, we can simulate the device and obtain these values, but we seek a more methodical and less laborious procedure. Our approach proceeds in three steps. We consider $I_D = 0.5$ mA and $V_{DS,min} = 200$ mV as an example.

Step 1 Select a "reference" transistor, with a width W_{REF} and a length equal to the minimum allowable value, L_{min} (e.g., $L_{min} = 40$ nm). Let us choose $W_{REF} = 5\ \mu$m as an example.

Step 2 Using the actual device models and a circuit simulator, plot the I_D-V_{DS} characteristics of the reference transistor for different values of $V_{GS} - V_{TH}$. In typical analog circuits, $V_{GS} - V_{TH}$ ranges from about 50 mV to about 600 mV. We can therefore construct the characteristics with the overdrive

[2] In cases I and III, raising W/L can relax the g_m-I_D trade-off only if the device does not enter the subthreshold region.

[3] We assume that the supply voltage is given.

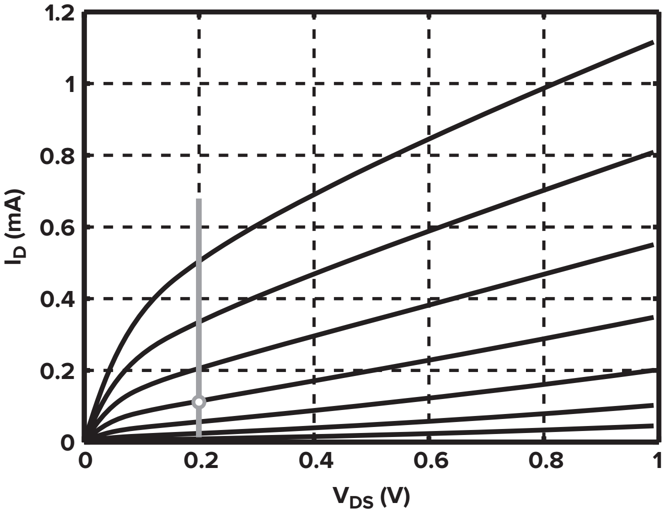


Figure 11.10 Drain current for $V_{GS} - V_{TH} = 50$ mV $\cdots$ 350 mV in steps of 50 mV for a reference device.

incrementing in steps of 50 mV.[4] Figure 11.10 shows the results for $W_{REF}/L_{min} = 5$ μm/40 nm. (Here, $V_{GS} - V_{TH}$ increments from 50 mV to 350 mV for clarity.)

Step 3 Bearing in mind that our example specifies $I_D = 0.5$ mA and $V_{DS,min} = 200$ mV, we draw a vertical line at $V_{DS} = 200$ mV (Fig. 11.10) and find its intersection with the plots. Which plot should we select? If the device obeyed the square law, we would choose the plot for $V_{GS}-V_{TH} = V_{DS,min} = 200$ mV. However, the short-channel device remains in saturation even for $V_{GS}-V_{TH} = 350$ mV at $V_{DS} = 200$ mV. The situation is therefore more complex, but let us proceed with $V_{GS} - V_{TH} = 200$ mV for now.

Step 4 The foregoing procedure has yielded, for the reference transistor, one operating point that satisfies the V_{DS} requirement. The drain current, $I_{D,REF}$, however, may not be close to the necessary value, 0.5 mA in our example. What shall we do here? We must now *scale* the width of the transistor and hence its drain current. Since in Fig. 11.10, $I_{D,REF} \approx 100$ μA, we choose a transistor width of (500 μA/100 μA)$\times W_{REF} = 5W_{REF} = 25$ μm.

How much is the transconductance of the earlier reference transistor? We recognize from the plots of Fig. 11.10 that, as $V_{GS}-V_{TH}$ is incremented from 200 mV to 250 mV, I_D changes by about 100 μA. Thus, $g_m \approx 100$ μA/50 mV= 2 mS. Since the change in the overdrive is not much less than the initial value of 200 mV, we may seek a more accurate value for g_m. To this end, let us return to the reference transistor and, using simulations, plot its transconductance as a function of $V_{GS} - V_{TH}$ with $V_{DS} = 200$ mV. For a square-law device, this plot would be a straight line, $g_m = \mu_n C_{ox}(W_{REF}/L_{min})(V_{GS} - V_{TH})$, but with short-channel effects, g_m eventually saturates. Shown in Fig. 11.11, the result predicts $g_m = 1.5$ mS for $V_{GS} - V_{TH} = 200$ mV. Now, if both the width and the drain current are scaled up by a factor of 5, then g_m also rises by the same factor (Example 11.2), reaching a value of 7.5 mS. As indicated in Table 11.1, if this transconductance is insufficient, W/L must be increased further.

With the I_D and g_m plots obtained for the reference device, we can readily perform scaling to determine the width and transconductance of other transistors in a circuit. The key point here is that the I_D and g_m simulations are performed only once (for a given channel length) but serve most of our design work.

[4] Our approach deals with only moderate and strong inversion, as is the case in most analog circuits.

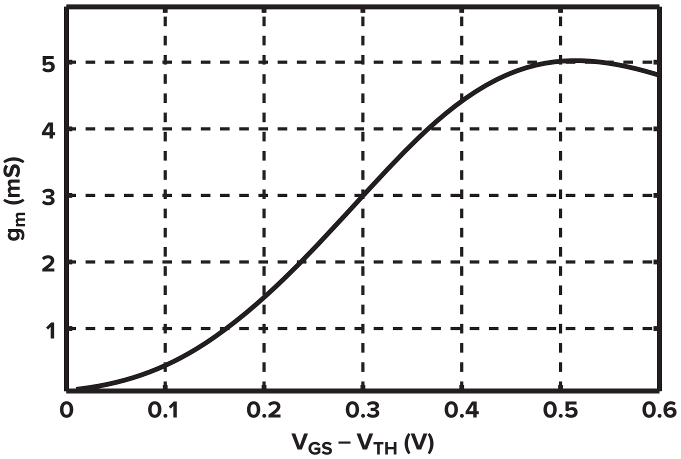


Figure 11.11 Dependence of g_m on overdrive for $W/L = 5\ \mu$m/40 nm and $V_{DS} = 200$ mV.

Can we choose a higher overdrive voltage in Fig. 11.10? Suppose we select $V_{GS} - V_{TH} = 250$ mV, obtaining $I_D = 200\ \mu$A for the reference transistor and a transconductance of about 2.3 mS from Fig. 11.11. If scaled up to 12.5 μm so as to carry 500 μA, the transistor exhibits a transconductance of 2.5×2.3 mS $= 5.75$ mS, a value *less* than that observed in the previous case (7.5 mS). This occurs because $g_m = 2I_D/(V_{GS} - V_{TH})$ in saturation. To obtain a high transconductance, therefore, we typically choose $V_{GS} - V_{TH} \approx V_{DS,min}$ even though it translates to a wider transistor.

▶ **Example 11.4**

The circuit shown in Fig. 11.12 must be designed for a power budget of 1 mW and a peak-to-peak output voltage swing of 0.8 V. Assuming $L = 40$ nm for M_1, compute its required width. Can the transistor provide a transconductance of $1/(50\ \Omega)$?

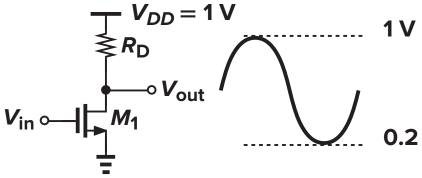


Figure 11.12

Solution

The power budget along with $V_{DD} = 1$ V translates to a drain bias current of 1 mA. For the circuit to accommodate an output swing of 0.8 V, M_1 must remain in saturation as V_{DS} falls to 0.2 V. We return to the I_D-V_{DS} characteristics of Fig. 11.10 and recall that $I_{D,REF} \approx 100\ \mu$A for $V_{DS} = V_{GS} - V_{TH} = 200$ mV. We must therefore scale W_{REF} up by a factor of 1 mA/0.1 mA, obtaining $W/L = 50\ \mu$m/40 nm. The transconductance is also multiplied by this factor, reaching 15 mS $= 1/(67\ \Omega)$. Note that these results are independent of the value of R_D.

We conclude that, if the transistor is designed simply to satisfy this example's I_D and V_{DS} specifications, then it does not necessarily achieve a transconductance of $1/(50\ \Omega)$.

◀

In addition to I_D, $V_{GS} - V_{TH}$, and g_m, the output impedance of the transistors also becomes important in many analog circuits. As explained in Chapter 17, r_O cannot be expressed as $1/(\lambda I_D)$ for short-channel devices. The value of r_O can be estimated from the slope of the I_D characteristics in Fig. 11.2, but for convenience and accuracy, we use simulations to plot r_O for the reference transistor as a function of I_D (Fig. 11.13).

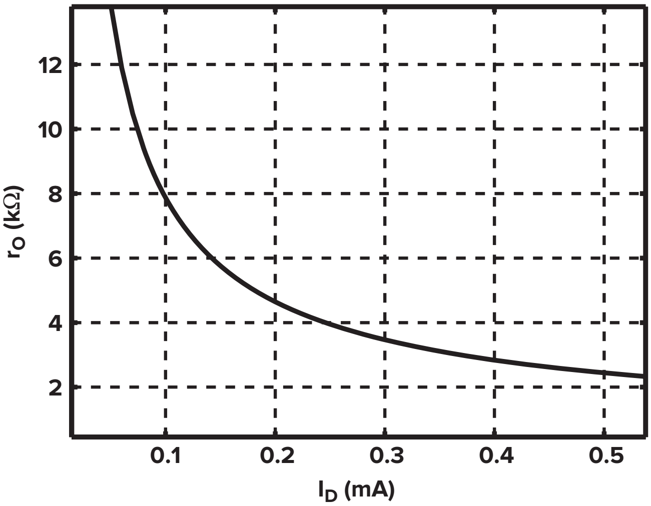


Figure 11.13 Output resistance of a 5-μm/40-nm NMOS device as a function of drain current.

▶ **Example 11.5**

Determine the output resistance of M_1 in Example 11.4.

Solution

The reference transistor in Example 11.4 carries a current of 100 μA, exhibiting an output resistance of 8 kΩ. Since both the width and the drain current are scaled up by a factor of 10, the output resistance drops by the same factor, falling to 800 Ω.

◀

11.4.2 Design for Given g_m and I_D

In many analog circuits, a given transistor must provide sufficient transconductance while consuming minimal power. We thus begin with a specified transconductance, g_{m1}, and an upper limit for the drain bias current, I_{D1}, seek the corresponding values of W/L and $V_{GS} - V_{TH}$. In this section, we assume that $g_{m1} = 10$ mS and $I_{D1} = 1$ mA. Of course, our first task is to determine whether g_{m1} can be obtained at all with $I_D \leq I_{D1}$. The maximum g_m occurs in the subthreshold region (if W/L is large) and is given by $I_D/(\xi V_T)$, where $\xi \approx 1.5$ (Chapter 2). For example, if $I_D = 1$ mA, then g_m cannot exceed 26 mS at the room temperature.

Since in our example, $g_{m1} < I_{D1}/(\xi V_T)$, we can proceed to design the transistor. The reader is encouraged to first read Example 11.3 carefully.

Step 1 Using simulations, we plot g_m as a function of I_D for a reference transistor, e.g., with $W_{REF}/L_{min} = 5$ μm/40 nm (Fig. 11.14).

Step 2 We identify the point (I_{D1}, g_{m1}) on the g_m-I_D plane and draw a line through the origin and this point, obtaining the intersection at $(I_{D,REF}, g_{m,REF}) = (240$ μA, 2.4 mS) and a corresponding overdrive.

Step 3 We multiply W_{REF} by $g_{m1}/g_{m,REF} = 4.2$ so as to travel on the straight line to point (I_{D1}, g_{m1}) while maintaining the same overdrive (Example 11.3). This completes the design of the transistor.

The above procedure elicits two questions. First, does the straight line passing through the origin and (I_{D1}, g_{m1}) always intersect the g_m-I_D plot? If we consider a square-law device in strong inversion, then $g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$ has a slope of infinity at the origin, guaranteeing an intersection point. In the

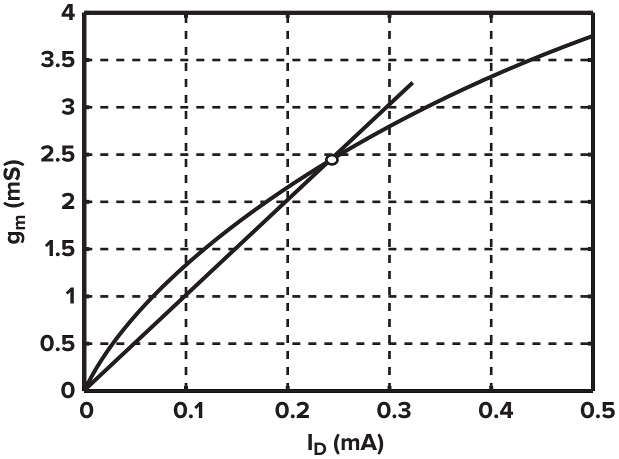


Figure 11.14 Transconductance as a function of I_D for $W/L = 5\ \mu$m/40 nm.

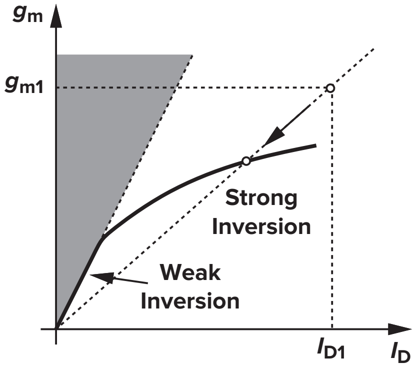


Figure 11.15 Unachievable g_m region.

subthreshold region, on the other hand, $g_m \propto I_D$ (Fig. 11.15), which means that the (I_D, g_m) combinations in the gray region are not achievable.

The second question is, what if $(V_{GS} - V_{TH})_{REF}$ is excessively large? As stipulated in Table 11.1, we must then increase W further, but by what factor? Suppose, as shown in Fig. 11.16, an overdrive of $(V_{GS} - V_{TH})_2 < (V_{GS} - V_{TH})_{REF}$ is desired. We then find the corresponding current, I_{D2}, and transconductance, g_{m2}, on the g_m-I_D plane. Next, we draw a line through the origin and the point (I_{D2}, g_{m2}) and continue to $I_D = I_{D1}$, i.e., we multiply W_{REF} by I_{D1}/I_{D2}. The resulting width guarantees an overdrive of $(V_{GS} - V_{TH})_2$ at a drain current of I_{D1} and provides a transconductance of *at least* g_{m1}. The new transconductance, g'_{m1}, is inevitably greater because the width has been increased beyond $g_{m1}/g_{m,REF}$ $(= I_{D1}/I_{D,REF})$.

11.4.3 Design for Given g_m and $V_{DS,min}$

In some designs, the transconductance is dictated by some performance requirements (voltage gain, noise, etc.) and the minimum V_{DS} by the voltage headroom—with no explicit specification of I_D. Of course, each circuit eventually faces a power budget and hence an upper bound on its bias current(s).

The design procedure for obtaining a transconductance of g_{m1} at $V_{DS,min}$ in this case is as follows.

Step 1 We use simulations to plot the g_m as a function of $V_{GS} - V_{TH}$ for the reference transistor (Fig. 11.17). Now, we select $(V_{GS} - V_{TH})_1 = V_{DS,min}$ and obtain the corresponding transconductance, $g_{m,REF}$. In this case, it is helpful to plot I_D on the same plane and find $I_{D,REF}$ at $(V_{GS} - V_{TH})_1$.

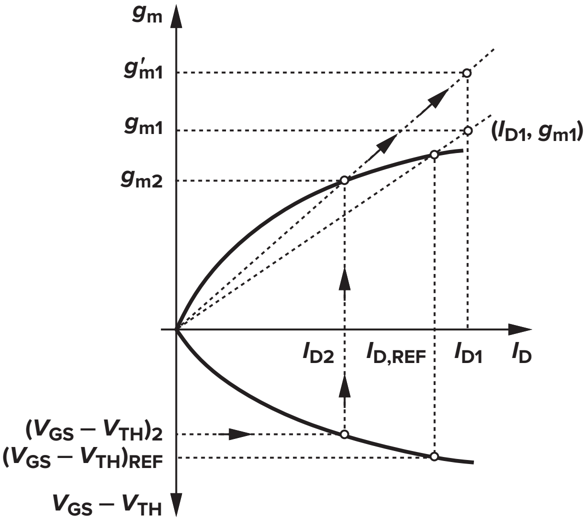


Figure 11.16 Modification of transistor design for a lower overdrive.

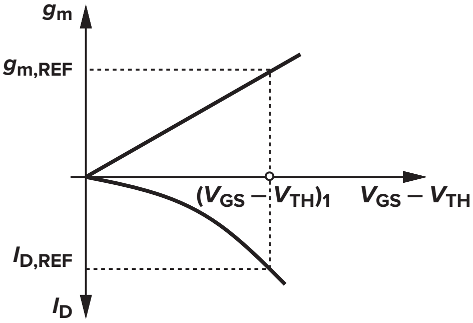


Figure 11.17 Calculation of $g_{m,REF}$ for a given overdrive.

Step 2 To reach the required transconductance, g_{m1}, we scale the transistor width up by a factor of $g_{m1}/g_{m,REF}$. Note that I_D scales by the same factor.

These two steps complete the design, but what if the resulting I_D is excessively large? We can return to Case II in Sec. 11.4.2 and redesign for a given g_m and I_D. The device is now wider and has a smaller transconductance.

As can be seen from our procedures in this section, we have portrayed the overdrive voltage (or $V_{D,sat}$) as an indispensable dimension in our device design. This is because today's low supply voltages have made the problem of headroom more severe than ever.

11.4.4 Design for a Given g_m

Our approach has assumed that the drain current and the overdrive voltage are specified and the other device parameters must be determined. Since power consumption and voltage headroom prove critical in today's analog design, this assumption holds in most cases. However, suppose a design problem specifies only the transconductance, and we wish to compute the remaining parameters. How do we select the transistor's drain current, overdrive voltage, and dimensions?

Two scenarios must be envisaged. (1) We select a certain W/L and raise I_D until we obtain the desired transconductance, g_{m1}. In this case, the required I_D, and hence the power consumption, may be excessive. More important, the overdrive voltage may be unacceptably large, leaving little headroom for voltage swings. (2) We select a reasonable value for I_D (perhaps according to a power budget) and increase W/L to obtain g_{m1}. In this case, however, we may *not* be able to reach g_{m1}; increasing W/L (and hence decreasing V_{GS}) eventually drives the device into the subthreshold region, where g_m cannot exceed $I_D/(\xi V_T)$. This means that the selected current is insufficient, i.e., we should always briefly check to see if the upper limit given by $I_D/(\xi V_T)$ can be met with the current budget.

The above scenarios indicate the need for a systematic approach to the selection of device parameters when only g_{m1} is known. To this end, we return to the concept of the reference device and, using simulations, construct two plots for it. Shown in Fig. 11.18, the two represent g_m and $V_{GS} - V_{TH}$ as a function of I_D.[5] We begin by selecting a reasonable value for $V_{GS} - V_{TH}$, e.g., 200 mV, which points to $I_{D,REF}$ and $g_{m,REF}$. Now, we scale the width and the drain current by a factor of $g_{m1}/g_{m,REF}$.

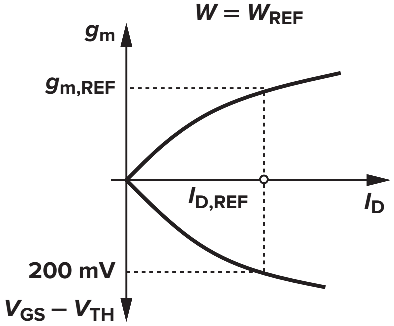


Figure 11.18 Translation of overdrive to $g_{m,REF}$.

What if the foregoing method yields an unacceptably high I_D? We can choose a smaller overdrive, e.g., 150 mV, and repeat the earlier steps.

11.4.5 Choice of Channel Length

If the selection of I_D, $V_{GS} - V_{TH}$, and g_m does not yield a sufficiently high r_O, we must increase the length of the transistor. Of course, to maintain the same drain current, overdrive voltage, and g_m, the *width* must also be increased proportionally. However, such scaling of the length and the width is not straightforward because, as the *drawn* length is increased from L_{min} to, say, $2L_{min}$, the effective length rises from $L_{min} - 2L_D$ to $2L_{min} - 2L_D$, i.e., by a factor of less than 2. For this reason, we must use simulations to construct the I_D-V_{DS}, g_m and r_O characteristics for several channel lengths, e.g., 60 nm, 80 nm, and 100 nm (drawn values).

11.5 ■ Op Amp Design Examples

In this section, we wish to repeat the op amp design examples studied in Chapter 9 in 40-nm technology. We target the following typical specifications:

- Differential Output Voltage Swing = 1 V_{pp}
- Power Consumption = 2 mW
- Voltage Gain = 500
- Supply Voltage = 1 V

[5] Here, V_{DS} is constant and approximately equal to $V_{DD}/2$. In nanometer technologies, different V_{DS} values alter these characteristics to some extent.

The single-ended output swing of 0.5 V_{pp} is small enough to make telescopic or folded-cascode op amps a plausible choice. We therefore explore these topologies before deciding whether a two-stage op amp is necessary.

A few notes about our transistor sizing methodology are warranted. We wish to begin with the minimum allowable width and length unless otherwise dictated by current, transconductance, $V_{D,sat}$, output resistance, or other requirements. Interestingly, in the designs pursued in this chapter, all transistor widths are greater than the minimum value. Also, for simplicity, we may scale the drawn W and L by the same factor even though the effective W/L does not remain exactly constant.

11.5.1 Telescopic Op Amp

Can a telescopic-cascode op amp topology meet the above specifications? In this section, we explore this possibility. It may not, but we will learn a great deal. Consider the circuit shown in Fig. 11.19. Of the total supply current of 2 mA, we allow 50 μA for I_{REF1}, 50 μA for I_{REF2}, and 0.95 mA for each branch of the differential pair. We must now allocate the transistor drain-source voltages so as to accommodate a single-ended peak-to-peak output swing of 0.5 V; i.e., we must distribute the remaining 0.5 V over M_9, $M_{1,2}$, $M_{3,4}$, $M_{5,6}$, and $M_{7,8}$. Let us allow a V_{DS} of 100 mV for each—even though the PMOS devices have a lower mobility. With the bias currents and overdrives known, we can determine the W/L's by examining the transistor I/V characteristics.

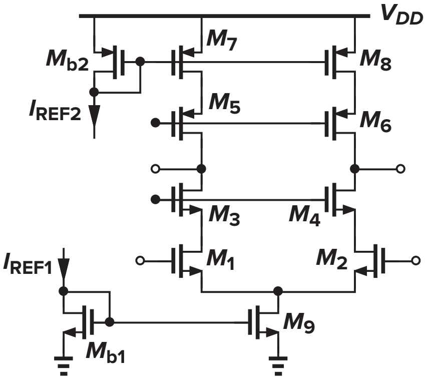


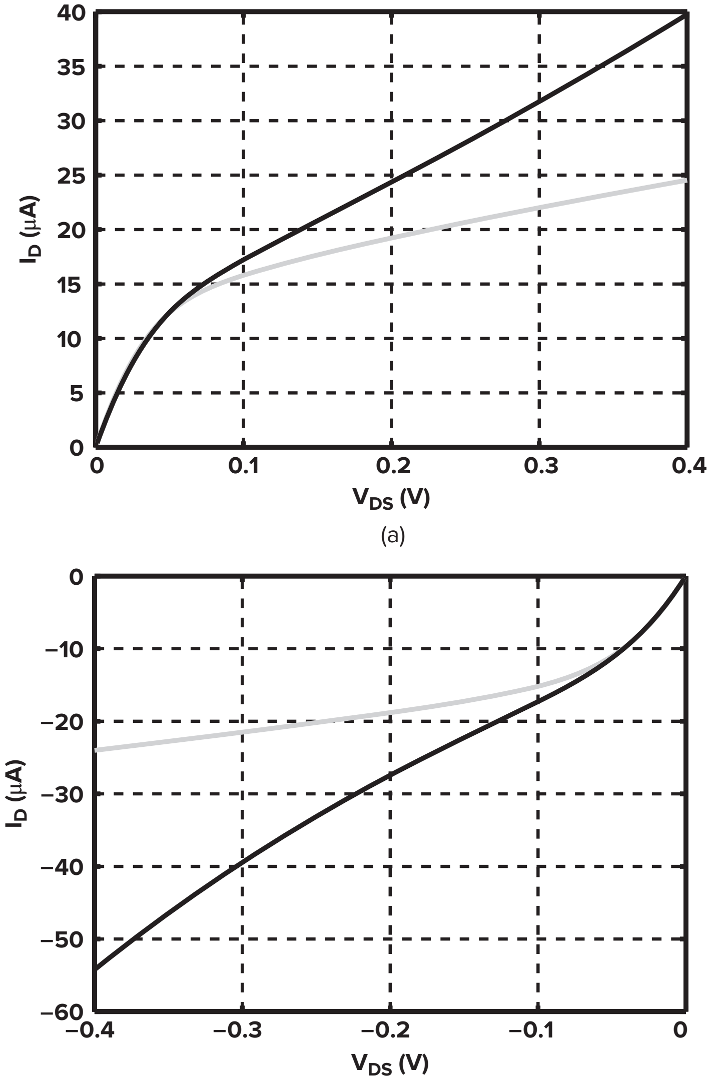
Figure 11.19 Telescopic-cascode op amp.

Before delving into details, however, we should pause and think about the feasibility of the design, specifically in terms of the required voltage gain. We make three observations: (1) for $L = 40$ nm, the intrinsic gain, $g_m r_O$, of NMOS devices is around 7 to 10 and that of PMOS devices around 5 to 7, (2) for reasonable device dimensions, it is difficult to raise $g_m r_O$ beyond 10 for PFETs (unless we allow longer lengths and hence lower speeds), and (3) if we approximate g_m as $2I_D/(V_{GS} - V_{TH}) = 2 \times 0.95\ \text{mA}/100\ \text{mV} = 19$ mS, we estimate $r_O \approx 530\ \Omega$ from $g_m r_O \approx 10$.

Let us now apply the foregoing values to the telescopic topology in Fig. 11.19. If $g_{m1,2} \approx 19$ mS, then for the gain, $G_m R_{out}$, to reach 500, the op amp output impedance must exceed 26 kΩ. equal to $(g_{m5,6} r_{O5,6}) r_{O7,8}$, pointing to a serious limitation. However, with $g_{m3,4} r_{O3,4} \approx 10$ and $r_{O7,8} \approx 530\ \Omega$ (from the third observation above), we have $(g_{m3,4} r_{O3,4}) r_{O1,2} \approx 5.3$ kΩ, obtaining a voltage gain of only about 100 even if the PMOS devices have $\lambda = 0$! This fivefold deficit makes the telescopic arrangement impractical for a gain of 500.

Out of curiosity, we still continue with the design and see what performance we can achieve. To this end, we use simulations to construct the I/V characteristics of NMOS and PMOS devices with $L = 40$ nm and 80 nm, predicting that the minimum length exhibits an unacceptably low r_O and $g_m r_O$. The simulation parameters must also ensure that the devices remain in saturation for $|V_{DS}| \geq 100$ mV. Given that the threshold and the overdrive elude a clear definition in nanometer technologies, we must adjust V_{GS} in simulations to ensure saturation.

The results are plotted in Fig. 11.20(a) for $(W/L)_N = 5\ \mu$m/40 nm and 10 μm/80 nm with $V_{GS} =$ 300 mV, and in Fig. 11.20(b) for $(W/L)_P = 5\ \mu$m/40 nm and 5 μm/80 nm with $V_{GS} = -400$ mV.[6] We should make some remarks. First, it is difficult to distinguish between triode and saturation regions, especially for PFETs. In fact, the 40-nm PMOS device behaves almost as a resistor and displays a



(a)

(b)

Figure 11.20 I_D-V_{DS} characteristics for (a) an NMOS device with $V_{GS} = 300$ mV and $W/L = 5\ \mu$m/40 nm (black plot) or 10 μm/80 nm (gray plot), and (b) a PMOS device with $V_{GS} = -400$ mV and $W/L = 5\ \mu$m/40 nm (black plot) or 5 μm/80 nm (gray plot).

[6] The width of the PMOS device is not scaled here to reveal the small change in I_D as the drawn L is doubled. This occurs because V_{TH} *falls* as L increases from its minimum value (Chapter 17).

decreasing output impedance as $|V_{DS}|$ approaches 400 mV.[7] For the other three characteristics, we can roughly identify a "knee" point beyond which the slope falls considerably. The gate-source voltages have been chosen to place this point below $|V_{DS}| = 100$ mV.

Second, at $V_{DS} = 100$ mV, the 10-μm/80-nm NMOS transistor provides an output resistance of 22.8 kΩ and a drain current of 16 μA. If scaled up to carry 950 μA with the same terminal voltages, the device exhibits an r_O of 385 Ω! Similarly, the 5-μm/80-nm PMOS transistor has an r_O of 18.45 kΩ at $V_{DS} = -100$ mV with $I_D = 15$ μA, thus offering $r_O = 290$ Ω if scaled up to carry 950 μA. These very low r_O values are quite discouraging, but we will continue to explore.

We now scale the NMOS and PMOS device widths to accommodate a drain current of 950 μA with $V_{GS,N} = 300$ mV, $V_{GS,P} = -400$ mV, and $|V_{DS}| = 100$ mV. The resulting design is shown in Fig. 11.21.[8] As a general principle, we prefer to use minimum-length devices in the signal path so as to maximize their speed (or at least minimize their capacitances for a given g_m). It is surprising to see such large widths in 40-nm technology for a drain current of 950 μA, an inevitable outcome of confining $|V_{DS}|$ to 100 mV.

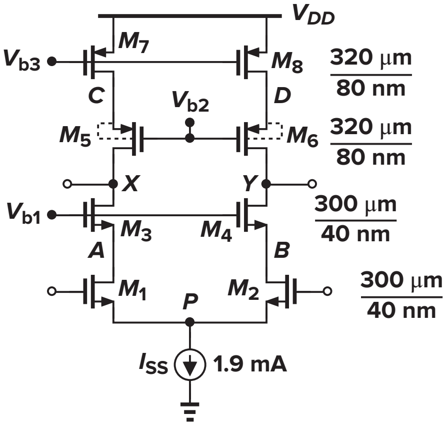


Figure 11.21 First design of telescopic-cascode op amp.

The bias voltages are tentatively chosen as follows: (a) the input common-mode level, $V_{CM,in}$, is equal to 100 mV for the tail current source plus $V_{GS1,2}(= 300$ mV), (b) V_{b1} is equal to $V_{D1,2}$ $(= 200$ mV) plus $V_{GS3,4}(= 300$ mV), (c) V_{b2} is equal to $V_{DD} - |V_{DS7,8}| - |V_{GS5,6}|$, and (d) V_{b3} is equal to $V_{DD} - |V_{GS7,8}|$. Upon simulating the circuit with these values, the reader may encounter a very low or high output common-mode level. This effect arises from the absence of common-mode feedback and hence the departure of $|I_{D7,8}|$ from 1.9 mA/2. We avoid this issue by a slight adjustment of V_{b3} for now.

We perform a dc sweep simulation of the differential input voltage, V_{in}, and examine the voltages at various nodes in Fig. 11.21 to ensure that the transistors are "healthy." Plotted in Fig. 11.22, the drain voltages of M_1 and M_2 are around 220 mV in the middle of the range. Similarly, the drain voltages of M_7 and M_8 are close to the targeted value.

Next, we study the output behavior, depicted in Fig. 11.22 by V_X and V_Y. The slope of each single-ended output is approximately equal to 15 in the vicinity of $V_{in} = 0$, yielding a differential gain of 30, far below our target. Can this design deliver a single-ended peak-to-peak swing of 0.5 V? We note that the characterisitic becomes very nonlinear as each output rises toward 0.7 V. In fact, around this output level, the slope gives a differential gain of about 6.4.

[7] As explained in Chapter 17, the fall in the output impedance arises from drain-induced barrier lowering (DIBL).

[8] The bulks of M_5 and M_6 are tied to their respective sources so as to remove body effect. While not essential, this arrangement reduces $|V_{GS5,6}|$, providing more comfortable margins in the design.

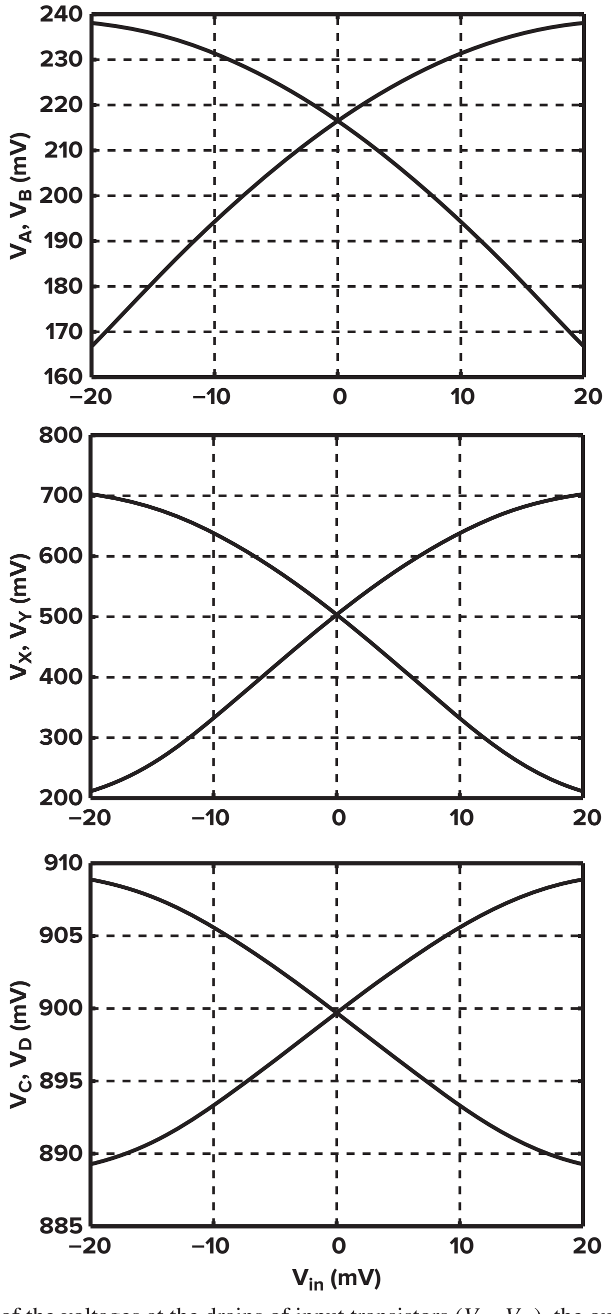


Figure 11.22 Behavior of the voltages at the drains of input transistors (V_A, V_B), the output nodes (V_X, V_Y), and the drains of PMOS current sources (V_C, V_D).

Example 11.6

The slope of the characteristics in Fig. 11.22 predicts a differential gain of 3 from the input to nodes A and B. Explain the reason for such a high gain at the cascode nodes.

Solution

Recall from Chapter 3 that the impedance seen at the source of a cascode device is roughly equal to the impedance seen at its drain divided by its $g_m r_O$. Due to the low $g_m r_O$, the impedance seen at A and B is quite a lot higher than $1/g_{m3,4}$, leading to a large gain.

Let us raise the gain by increasing $(W/L)_{3,4}$ to 600 μm/80 nm. Plotted in Fig. 11.23, the characteristics exhibit a gain of about 54 but still a limited output swing.

Bias Circuit The op amp of Fig. 11.21 relies on the proper choice of I_{SS}, V_{b1}, V_{b2}, and V_{b3}. We must therefore design a circuit to generate these bias quantities. We recognize that I_{SS} and V_{b3} must be established by current mirror action (why?) and V_{b2} by low-voltage cascode biasing. The bias voltage V_{b1} requires a different approach.

We begin with $I_{SS} = 1.9$ mA, choosing a channel length of 40 nm and, scaling from Fig. 11.20, a width of 600 μm for $V_{DS} = 100$ mV. Utilizing a reference budget current of 25 μA, we arrive at the arrangement shown in Fig. 11.24(a), where W_{12} is scaled down from W_{11} by a factor of 1.9 mA/25 μA. Since M_{11} operates with a V_{DS} of 100 mV, we insert R_1 in series with the drain of M_{12} and select its value such that $V_{DS12} = V_{GS12} - V_{R1} = 100$ mV.

The above bias design is still sensitive to the CM level sensed by M_1 and M_2 because $V_{DS11} = V_{CM,in} - V_{GS1,2}$, whereas $V_{DS12} = V_{GS1,2} - V_{R1}$. In other words, we must ensure that the drain voltage of M_{12} tracks $V_{CM,in}$. This can be accomplished as shown in Fig. 11.24(b), where R_1 is replaced by a differential pair driven by V_{in1} and V_{in2}. With proper scaling of the widths, we now have $V_{GS13,14} = V_{GS1,2}$, and hence $V_{DS12} = V_{DS11}$.

Next, we deal with the generation of V_{b1} in Fig. 11.21. This voltage must be equal to $V_{GS3,4} + V_{DS1,2} + V_P$, where $V_{DS1,2} = 100$ mV. Since V_{b1} is higher than V_P by $V_{GS3,4} + V_{DS1,2}$, we surmise that a diode-connected device in series with a drain-source voltage added to V_P can produce V_{b1}. Illustrated in Fig. 11.25, the idea is to match V_{GS15} to $V_{GS3,4}$ and V_{DS16} to $V_{DS1,2}$. The bias current I_b must be much less than I_{SS} so as to negligibly affect the power budget. We select $I_b = 15$ μA, and hence $(W/L)_{15,16} = 10$ μm/80 nm.[9] It is important to observe how V_{b1} tracks $V_{CM,in}$: if $V_{CM,in}$ goes up, so do V_P and, consequently, V_{b1}, thus keeping $V_{DS1,2}$ constant. That is, M_{15} and M_{16} operate as level shifters. If V_{b1} were *constant*, a rise in $V_{CM,in}$ would inevitably reduce $V_{DS1,2}$ and the gain.

In order to generate V_{b3} and V_{b2}, we construct a low-voltage cascode bias network as shown in Fig. 11.26. Here, transistors M_{17} and M_{18} are scaled down from $M_{7,8}$ and $M_{5,6}$, respectively, ensuring that $V_{DS17} = V_{DS7,8}$. To create $V_{b2} = V_{DD} - |V_{DS7,8}| - |V_{GS5,6}|$, we again employ a diode-connected device, M_{20}, in series with a V_{DS} (produced by M_{19}).

We should emphasize that the very narrow voltage margins dictated by the low supply make this design sensitive to mismatches between the bias branches and the core of the circuit. For example, a mismatch between V_{GS18} and $V_{GS5,6}$ can leave less $|V_{DS}|$ for $M_{7,8}$, pushing these two current sources below the knee point. Also, note that we still have a few ideal current sources, which would be copied from a bandgap reference (Chapter 12).

Common-Mode Feedback With various mismatches present in the above op amp design, the PMOS currents in Fig. 11.21 are not exactly equal to $I_{SS}/2$, forcing the output CM level toward V_{DD} or ground

[9] With the values chosen here, $V_{DS16} < V_{DS1,2}$ because $V_{GS16} > 300$ mV (why?). For this reason, some adjustment in simulations is necessary.

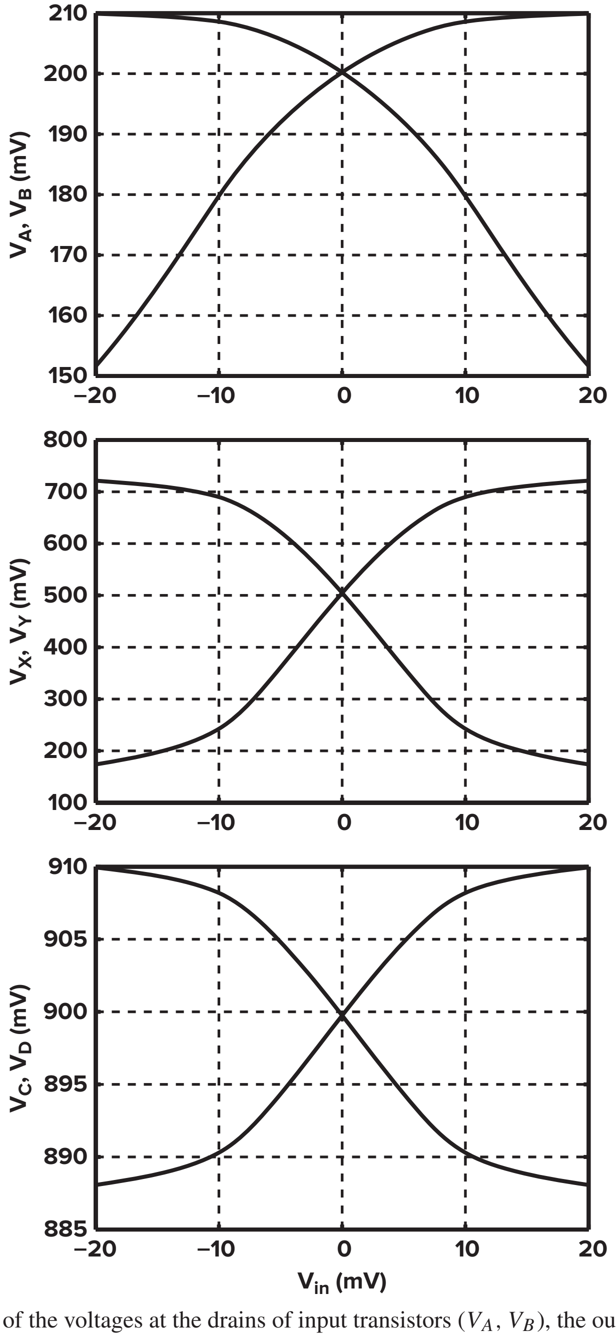


Figure 11.23 Behavior of the voltages at the drains of input transistors (V_A, V_B), the output nodes (V_X, V_Y), and the drains of PMOS current sources (V_C, V_D).

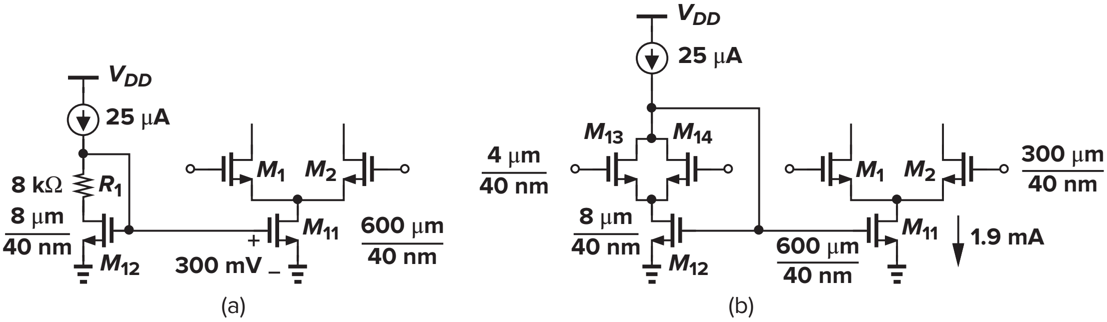


Figure 11.24 (a) Simple and (b) more accurate biasing for tail current source.

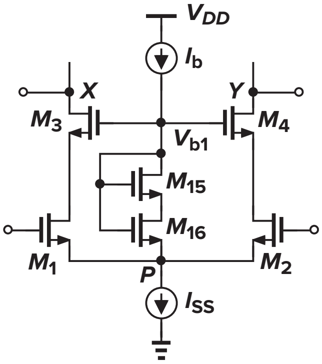


Figure 11.25 Generation of cascode gate bias voltage.

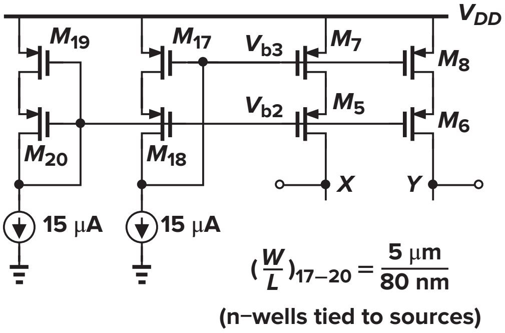


Figure 11.26 Generation of bias for PMOS cascode current sources.

and hence requiring CMFB. We must sense the output CM level, V_{CM}, and feed the result back to the NMOS or PMOS current sources.

Recall from Chapter 9 that the CM level can be sensed by resistors, triode transistors, or source followers. The high output impedance of the op amp dictates very large resistors,[10] and the tight voltage margins demand precise CM control and preclude triode devices. The only solution, source followers, however, cannot measure the CM level across a wide output swing. As shown in Fig. 11.27(a), if V_X (or V_Y) falls (in response to differential signals), I_1 (or I_2) eventually collapses, disabling the source follower. But, is it possible to complement the NMOS followers by PMOS counterparts? Consider the arrangement depicted in Fig. 11.27(b), where PMOS followers M_{23} and M_{24} also sense the output CM

[10] In addition to occupying a significant area, large resistors also degrade the CM loop stability, as explained later.

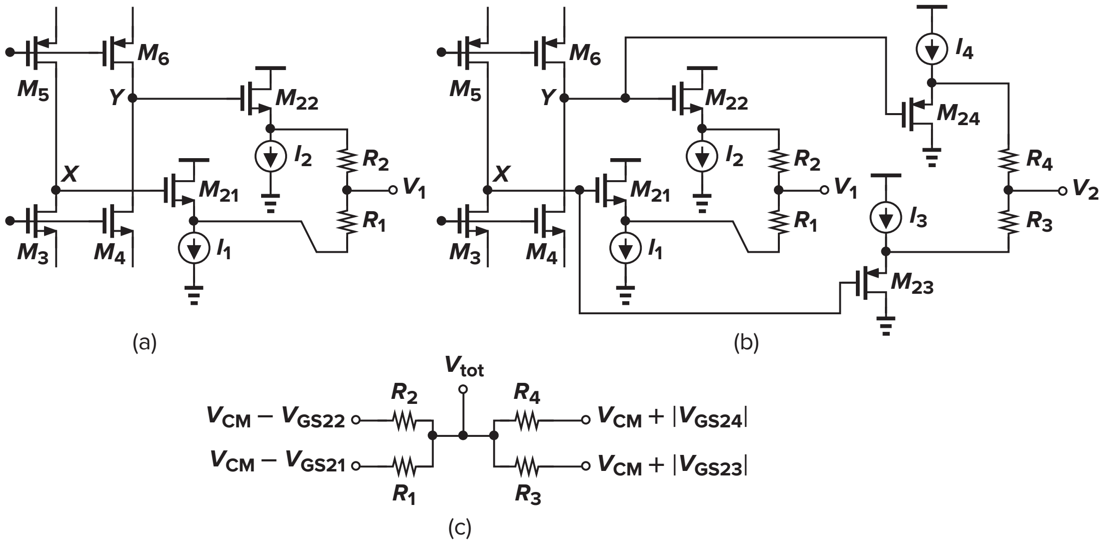


Figure 11.27 (a) CM level reconstruction using NMOS source followers, (b) CM level reconstruction using complementary source followers, and (c) combining network.

level and drive R_3 and R_4, respectively. We recognize that V_1 is lower than V_{CM} by $V_{GS21,22}$ and V_2 is higher than V_{CM} by $|V_{GS23,24}|$:

$$V_1 = V_{CM} - V_{GS21,22} \tag{11.10}$$

$$V_2 = V_{CM} + |V_{GS23,24}| \tag{11.11}$$

We therefore surmise that a linear combination of V_1 and V_2 can remove the V_{GS} terms, yielding a value in proportion to V_{CM}. That is, if

$$\alpha V_1 + \beta V_2 = (\alpha + \beta)V_{CM} - \alpha V_{GS21,22} + \beta |V_{GS23,24}| \tag{11.12}$$

then we must choose $\alpha V_{GS21,22} = \beta |V_{GS23,24}|$, obtaining $\alpha V_1 + \beta V_2 = (\alpha + \beta)V_{CM}$. We also choose $\alpha + \beta = 1$ so that the reconstructed value is equal to the op amp output CM level.

The weighting factors, α and β, can readily be implemented by R_1–R_4 in Fig. 11.27(b). In fact, if V_1 and V_2 are *shorted*, the weighted sum of V_1 and V_2 is produced. With the aid of the equivalent circuit in Fig. 11.27(c), the reader can show that

$$V_{tot} = V_{CM} + \frac{R_N |V_{GS23,24}| - R_P V_{GS21,22}}{R_N + R_P} \tag{11.13}$$

where $R_N = R_1 = R_2$ and $R_P = R_3 = R_4$. We therefore choose $R_N/R_P = V_{GS21,22}/|V_{GS23,24}|$.

In order to evaluate the feasibility of the above idea, we first run a dc sweep in simulations and examine the behavior of V_{tot}. We select a bias current of 10 μA (slightly exceeding the power budget), $W/L = 10$ μm/40 nm for all of the source followers, and $R_N = R_P = 20$ kΩ. The 10-μA bias current sources are also implemented as transistors (with $W/L = 10$ μm/40 nm) to ensure a realistic behavior as V_X and V_Y approach V_{DD} or ground.[11] Figure 11.28 plots the outputs, their actual common-mode level, defined as $(V_X + V_Y)/2$, and the reconstructed counterpart, V_{tot}. We note that V_{tot} closely follows the CM level of V_X and V_Y.

[11] Ideal current sources would allow the source voltages to exceed the supply rails.

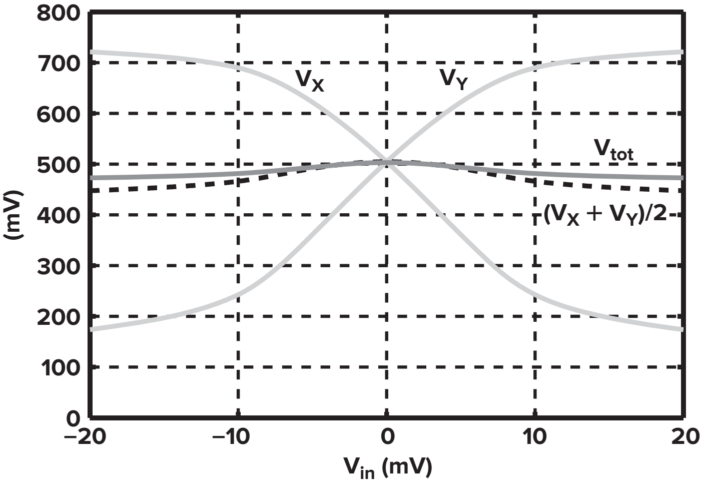


Figure 11.28 Actual CM level, $(V_X + V_Y)/2$, and reconstructed CM level, V_{tot}, of cascode op amp as a function of input differential voltage.

In the next test, let us close the CM feedback loop: we compare V_{tot} to a reference, amplify the error, and return the result to control I_{SS}. To this end, we design the error amplifier as a five-transistor OTA with $W/L = 5$ μm/80 nm for all transistors, a tail current of 20 μA, and a voltage gain of 10. The output of this amplifier controls a fraction of the main tail current, I_1 (Fig. 11.29). For example, if we expect 20% mismatch between the PMOS current sources in the op amp and the tail current source, we choose $I_1 \approx 0.2I_{SS}$. Figure 11.29 depicts the result, where the OTA's input and output connections are chosen so as to establish *negative* feedback around the loop.

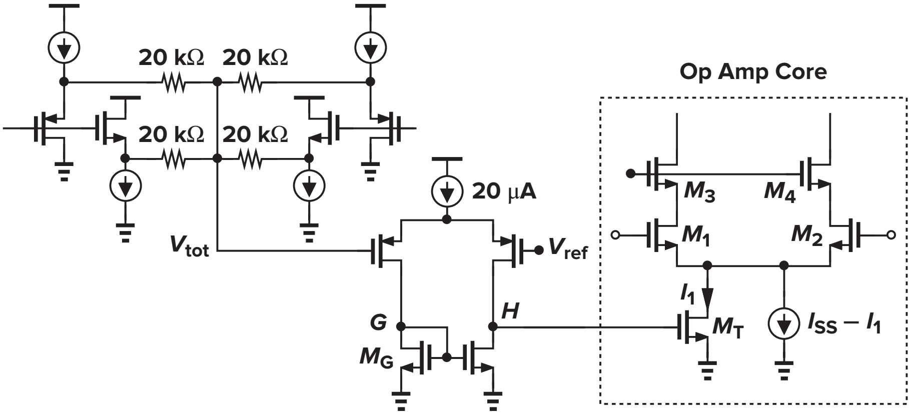


Figure 11.29 CMFB loop around telescopic op amp.

▶ **Example 11.7**

Explain why the OTA in Fig. 11.29 employs PMOS (rather than NMOS) input devices.

Solution

The choice is dictated by two considerations. First, these transistors must sense the CM level while leaving sufficient V_{DS} for their tail current source. With $V_{tot} \approx V_{DD}/2$ in this case, there is no particular preference for NMOS or PMOS devices. Second, the output of the OTA should have a nominal dc value compatible with that required by M_T.

Since $V_H = V_G$ (in the absence of mismatches), and since V_G is equal to the gate-source voltage of a diode-connected NMOS transistor, we expect that M_T nominally copies the bias current of M_G (with a multiplication factor).

Figure 11.30 shows the closed-loop dc sweep results with $V_{ref} = 0.5$ V.

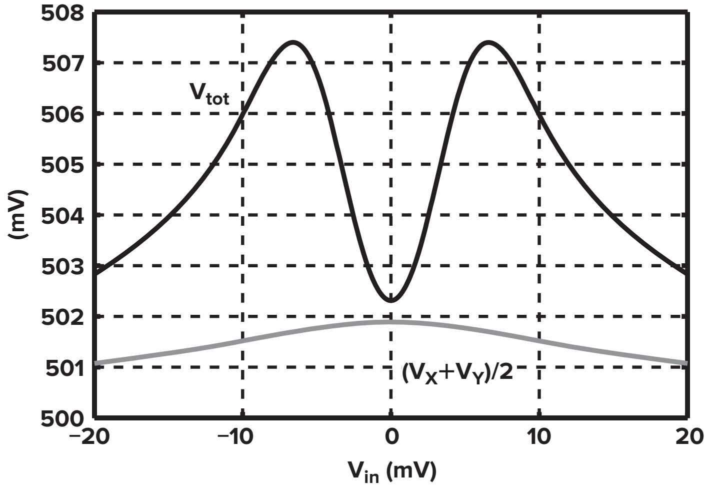


Figure 11.30 Closed-loop behavior of actual CM level, $(V_X + V_Y)/2$, and reconstructed CM level, V_{tot}, of cascode op amp as a function of input differential voltage.

By virtue of feedback, the CM variation is greatly reduced as V_X and V_Y reach high or low values. Next, we create a 10% mismatch between the PMOS current sources (M_7 and M_8 in Fig. 11.21) and $I_{SS}/2 = 950$ μA and repeat the dc sweep. Figure 11.31 depicts the variations, indicating that CMFB suppresses the mismatch by adjusting I_1.

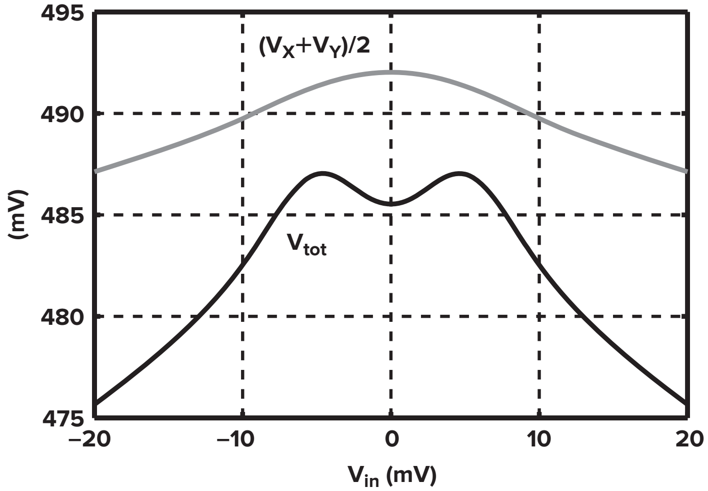


Figure 11.31 Closed-loop behavior of actual CM level, $(V_X + V_Y)/2$, and reconstructed CM level, V_{tot}, of cascode op amp in the presence of mismatch between tail and PMOS current sources.

CMFB Stability We must investigate the stability of the CM loop. This is accomplished by placing the overall op amp in its intended feedback system, applying differential pulses at the input, and examining the differential and common-mode behavior of the output. Figure 11.32(a) shows a feedback topology for a nominal closed-loop gain of 2, and Fig. 11.32(b) depicts a more detailed diagram highlighting the CM feedback loop.[12]

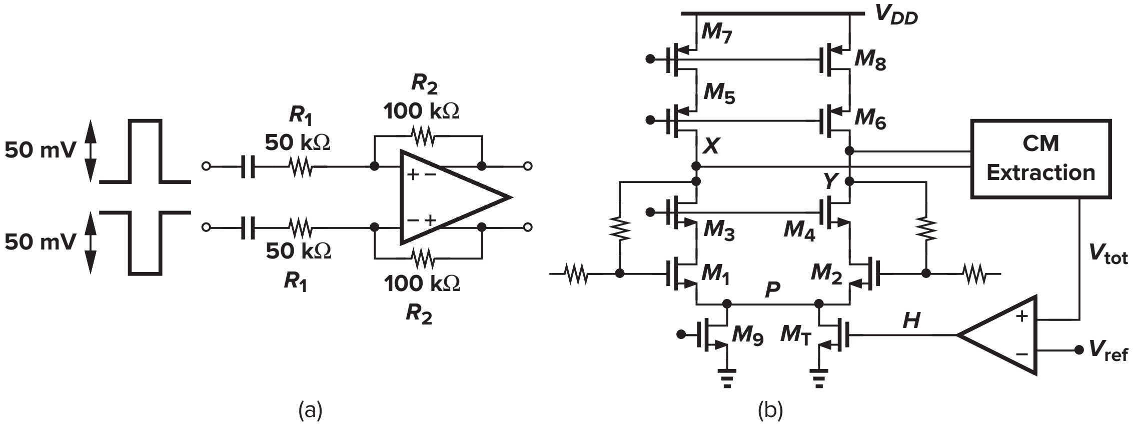


Figure 11.32 (a) Closed-loop amplifier for transient analysis, and (b) detailed view showing the CMFB loop.

Plotted in Fig. 11.33 are the output waveforms in response to an input step, revealing common-mode instability. As evident from Fig. 11.32(b), the CM loop contains a pole at the input of the error amplifier, one at node H, one at node P, one at the sources of the NMOS cascode devices, and one at the main outputs. The loop therefore demands compensation.

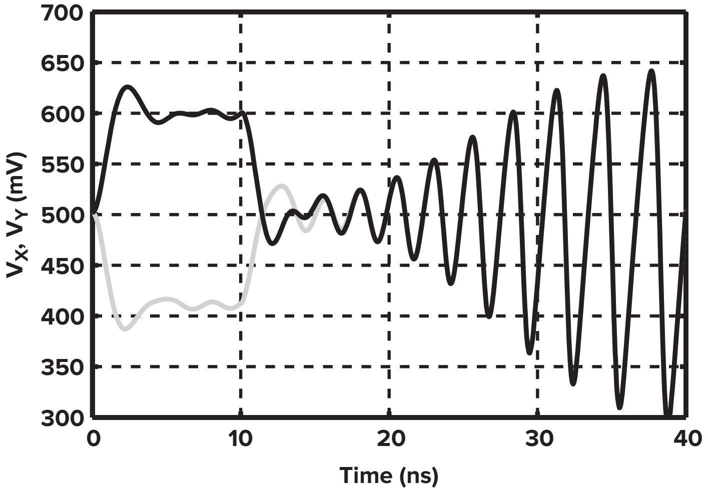


Figure 11.33 Transient response revealing CM loop instability.

▶ Example 11.8

We wish to study the CM loop frequency response and obtain the phase margin. Should the *differential* feedback be present when the CM loop is broken? In other words, which of the two topologies in Fig. 11.34(a) should be used to determine the CM loop transmission?

[12] Since a telescopic-cascode op amp does not easily lend itself to equal input and output CM levels, two 1-μA constant current sources (not shown) are tied from the inputs of the op amp to ground, shifting the input CM level down by 100 mV.

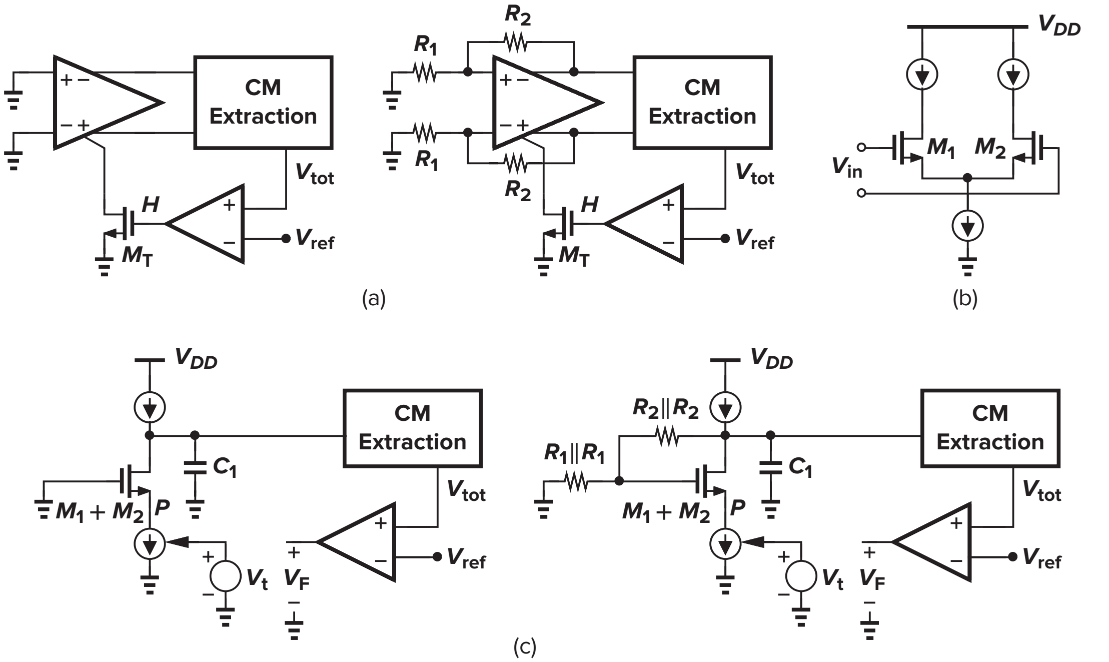


Figure 11.34

Solution

Common-mode feedback must ultimately behave well with differential feedback present. This is because the actual environment in which CMFB must be stable incorporates differential feedback. As an example, consider the simple op amp shown in Fig. 11.34(b). For CM analysis, the two sides can be merged into one, yielding the two possible scenarios depicted in Fig. 11.34(c) if the differential feedback is absent or present. Obtained as $-V_F/V_t$, the CM loop transmissions derived for these two cases are not necessarily the same. For example, if the capacitance at the drain, C_1, is considered, the pole associated with this node assumes different values in the two topologies. Thus, we must maintain differential feedback while studying CM stability.

Let us break the CM loop in Fig. 11.32(b) at node H as shown in Fig. 11.35. Here, the error amplifier drives a dummy device, M_d, identical to M_T so as to see the loading effect of the latter. Plotted in

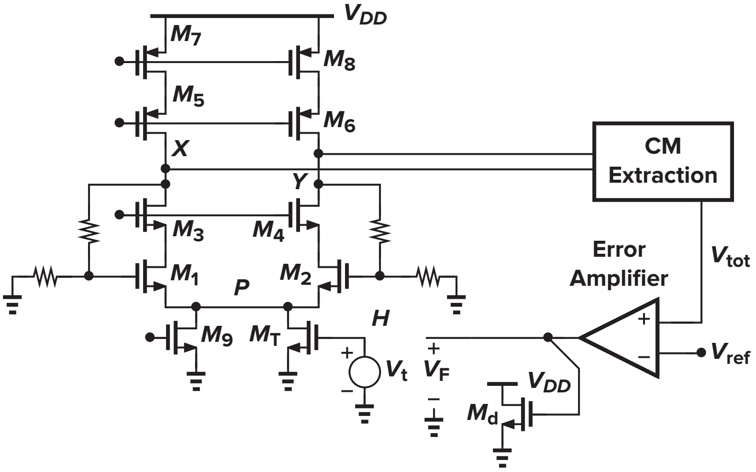
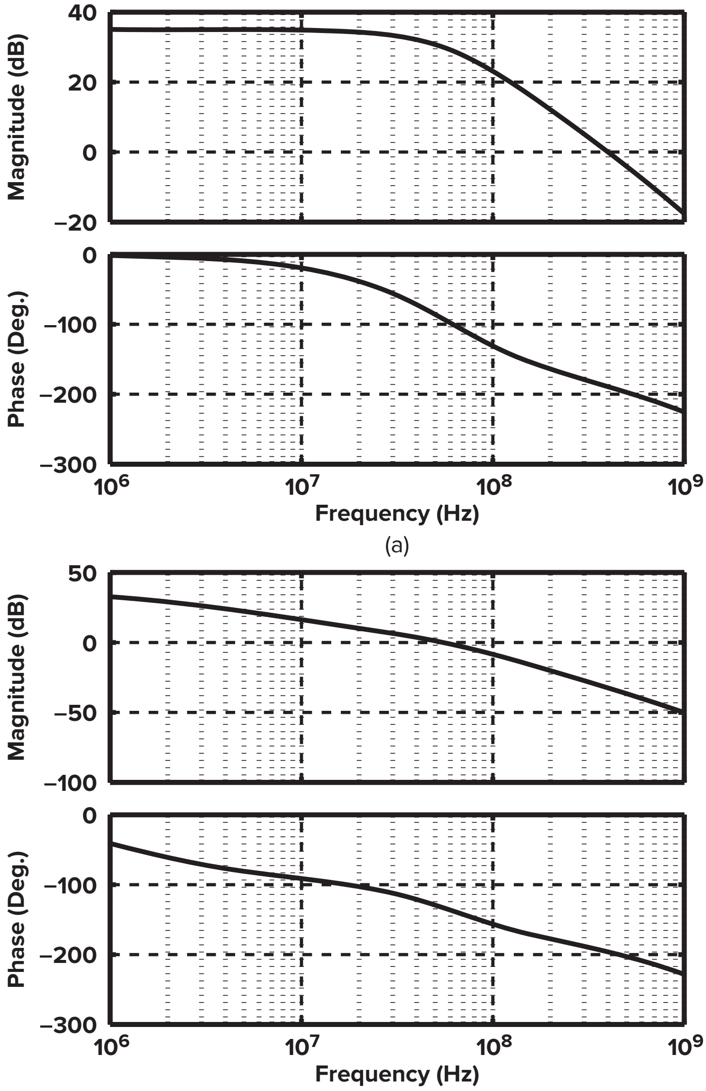


Figure 11.35 Measurement of CMFB loop transmission.

Fig. 11.36(a) are the magnitude and phase of $-V_F/V_t$ as a function of frequency, revealing a phase of $-190°$ at the unity-gain frequency. We seek a convenient node for compensation. Unfortunately, the error amplifier in Fig. 11.29 does not provide signal inversion from V_{tot} to H and hence cannot employ Miller compensation.



(b)

Figure 11.36 CMFB loop transmission (a) before and (b) after compensation.

Can we compensate the CM loop by adding capacitance from high-impedance nodes X and Y to ground? Yes, but this also affects the differential response. Instead, we tie a 3-pF capacitor from the error amplifier output to ground, obtaining the response shown in Fig. 11.36(b) and a phase margin of about 50°. The closed-loop pulse response depicted in Fig. 11.37(a) implies that the common-mode feedback loop is now properly compensated and the CM level incurs little ringing.

Differential Compensation Why do V_X and V_Y in Fig. 11.37(a) exhibit *differential* ringing? This is because the pole formed by the large resistors in the feedback network of Fig. 11.32(a) and the input

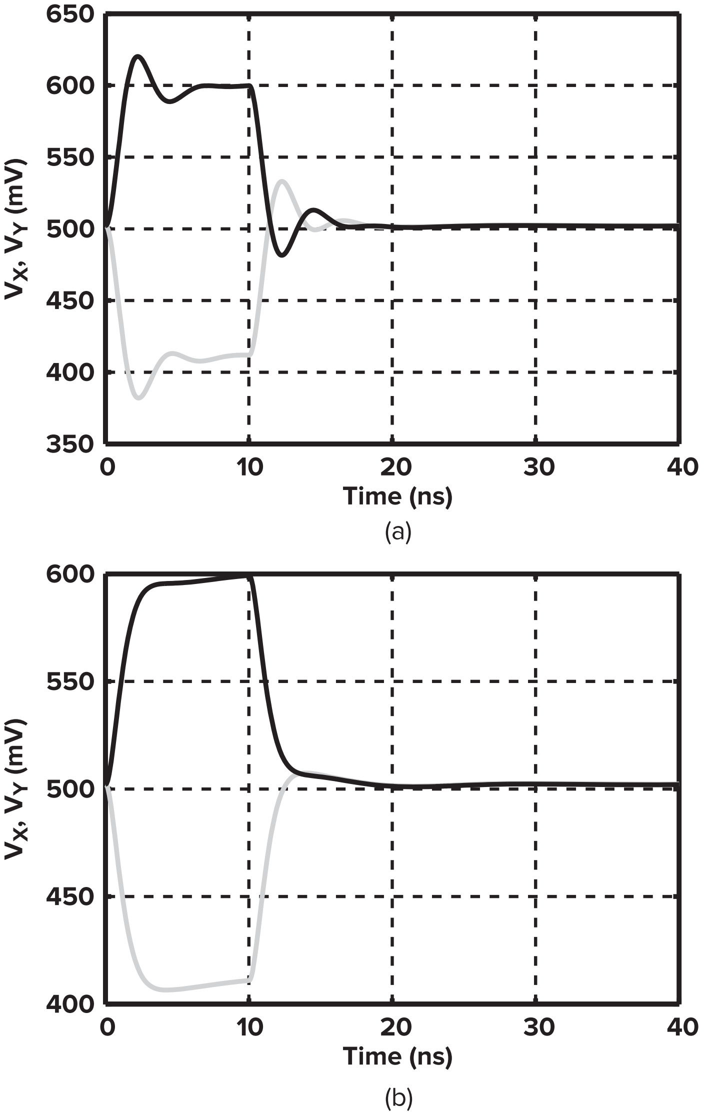


Figure 11.37 Transient response with (a) CMFB loop compensation and (b) additional differential compensation.

capacitance of the op amp is located at a low frequency, degrading the phase margin (of differential feedback). To compensate the differential signal path, we connect two 7-fF capacitors from the outputs of the op amp to its inputs (in parallel with the feedback resistors) so as to create Miller multiplication. Shown in Fig. 11.37(b), the resulting response is now well behaved. This pole-zero cancellation technique is studied in Problem 11.14.

▶ **Example 11.9**

Explain what design modifications are necessary if the op amp drives a significant load capacitance, C_L.

Solution

The load capacitance lowers the magnitude of the pole at X (and Y), increasing the differential signal path phase margin (Chapter 10) while decreasing the CM loop phase margin. For this reason, the capacitance tied to the error amplifier output must be increased, or the pole at X and Y must become the dominant pole for the CM loop as well.

Design Summary In this section, we have attempted to design a telescopic-cascode op amp for a voltage gain of 500 and a differential output swing of 1 $\mathrm{V_{pp}}$. Neither specification could be met with a 1-V supply, but we have established the steps that one must complete in order to arrive at the final design. Specifically, we have dealt with the following general principles:

1. Allocation of V_{DS} and I_D to transistors according to required swings and power dissipation, respectively
2. Characterization and scaling of MOSFETs for allowable V_{DS} and desired current level
3. Quick estimate of the achievable voltage gain
4. Use of dc sweep to study bias conditions and nonlinearity
5. Design of bias circuitry using current mirrors and low-voltage cascodes
6. Common-mode feedback design and compensation
7. Use of closed-loop transient analysis to study CM and differential stability

As seen in subsequent sections, these principles provide a systematic approach to the design of op amps.

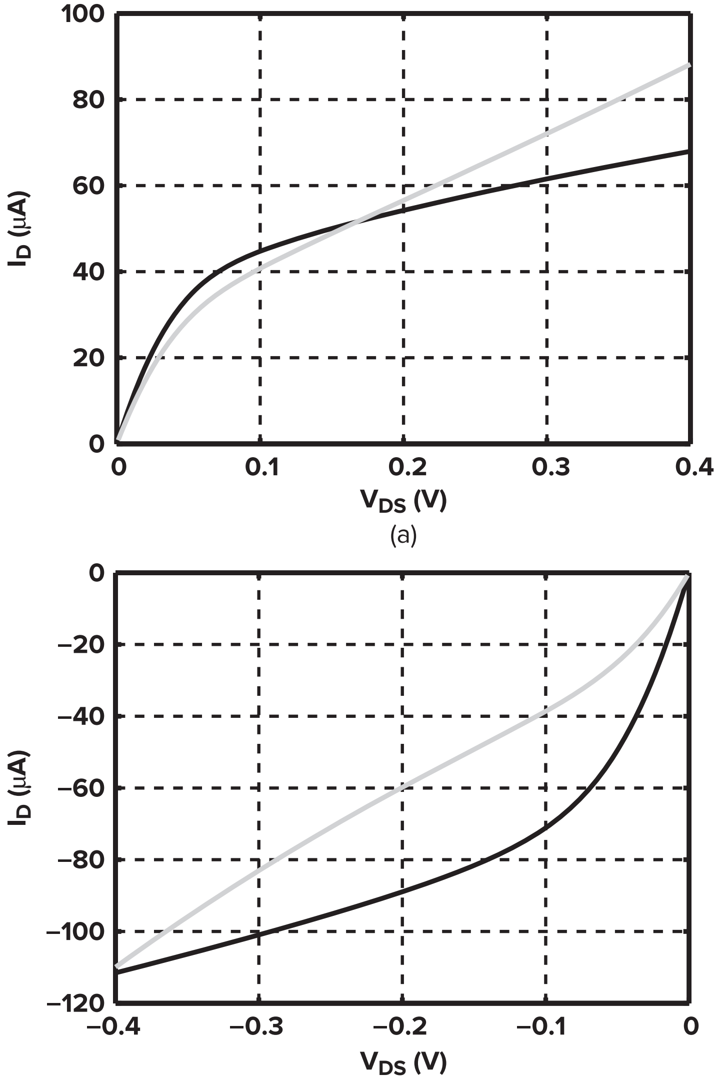
The next natural candidate for our op amp design is the folded cascade. However, our gain calculations for the telescopic cascode roughly apply here as well, predicting that it is extremely difficult to achieve a gain of 500. For this reason, we do not pursue the folded-cascode topology for these specifications.

11.5.2 Two-Stage Op Amp

Both the relatively high voltage gain and the 1-$\mathrm{V_{pp}}$ swing point to a two-stage configuration as a feasible candidate. We note that a gain of 500 dictates the use of cascoding in the first stage, encouraging us to utilize the telescopic design from the previous section (Fig. 11.21). However, two points must be borne in mind. First, the previous design exhausts the power budget, leaving none for the second stage. Second, with a gain of about 50 in the first stage, the gain of the second stage can be around 10. Thus, the single-ended peak-to-peak swing at the outputs of the first stage can be as small as 50 mV, allowing us to redesign the cascode for greater V_{DS}'s and hence more robust operation.

We must first partition the power budget between the two stages, a task requiring speed and/or noise specifications. We split the power equally here; further optimization could be pursued after one round of complete design. With about 100 μA reserved for the bias network, we allocate 1.9 mA/4 $=$ 475 μA to each branch of transistors in the first and second stages.

First-Stage Design The telescopic-cascode configuration must accommodate a single-ended swing of 50 $\mathrm{mV_{pp}}$, allowing 0.95 V for the sum of five V_{DS}'s. With some margin, we choose $V_{DS,N} = 150$ mV and $V_{DS,P} = 200$ mV and simulate our reference transistors ($W/L = 5$ μm/40 nm and 10 μm/80 nm), seeking acceptable knee voltages. Shown in Fig. 11.38 for $V_{GS,N} = 350$ mV and $V_{GS,P} = -450$ mV, the characteristics exhibit substantially higher drain currents than those in Sec. 11.5.1. It is interesting to note that, as a result of velocity saturation, the knee voltage has not increased by 50 mV. The width of the NMOS transistors in the signal path must be scaled by a factor of 450 μA/50 μA for either $L = 40$ nm or $L = 80$ nm. Similarly, the width of the PMOS device must be scaled by a factor of 450 μA/90 μA for $L = 80$ nm. We also choose for the tail current device $W = (900\ \mu\text{A}/50\ \mu\text{A}) \times 10\ \mu$m and $L = 80$ nm. The cascode-stage devices are thus much narrower than those used in the previous section. The first-stage design is shown in Fig. 11.39(a) and its simulated behavior in Fig. 11.39(b), revealing a gain of about 50. The biasing of this stage is similar to that described in Sec. 11.5.1.



(a)

(b)

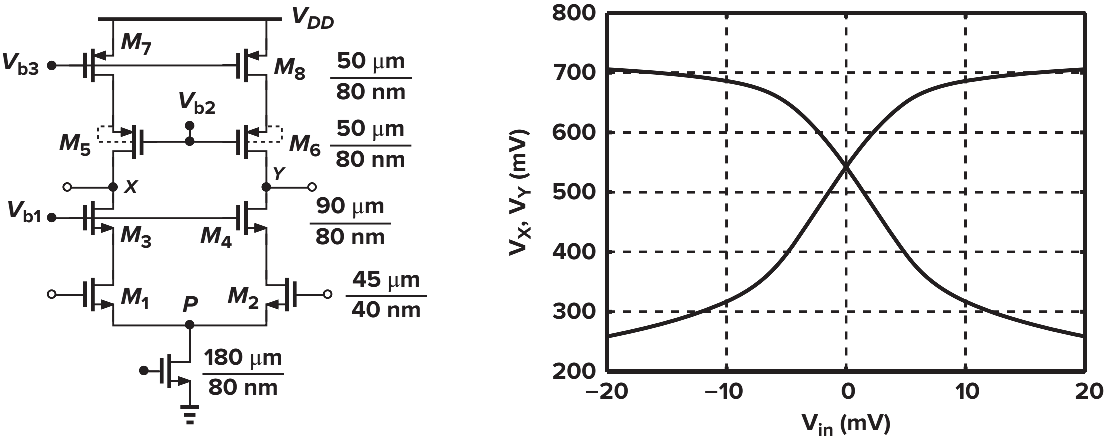
Figure 11.38 I_D-V_{DS} characteristics for (a) an NMOS device with $V_{GS} = 350$ mV and $W/L = 5$ μm/40 nm (gray plot) or 10 μm/80 nm (black plot), and (b) a PMOS device with $V_{GS} = -450$ mV and $W/L = 5$ μm/40 nm (gray plot) or 10 μm/80 nm (black plot).

▶ Example 11.10

In order to determine, with the aid of simulation, the small-signal resistance seen at node X in Fig. 11.39(a), a student sets the input signal to zero, ties a unit ac current source from this node to ground, and measures the resulting voltage. Explain why this test overestimates the resistance.

Solution

The voltage developed at X causes a current to flow through r_{O3} to the drain of M_1 and through r_{O1} to the source of M_2. In other words, since M_1 is degenerated by M_2, the resistance at X is overestimated. To avoid this error, a large



(a) (b)

Figure 11.39 (a) First stage design, and (b) its input-output characteristics.

capacitance must be tied from the source of M_1 to ground to create a short circuit at the test frequency. Alternatively, we can attach the ac current source between X and Y, measure the resistance, and divide the result by two.

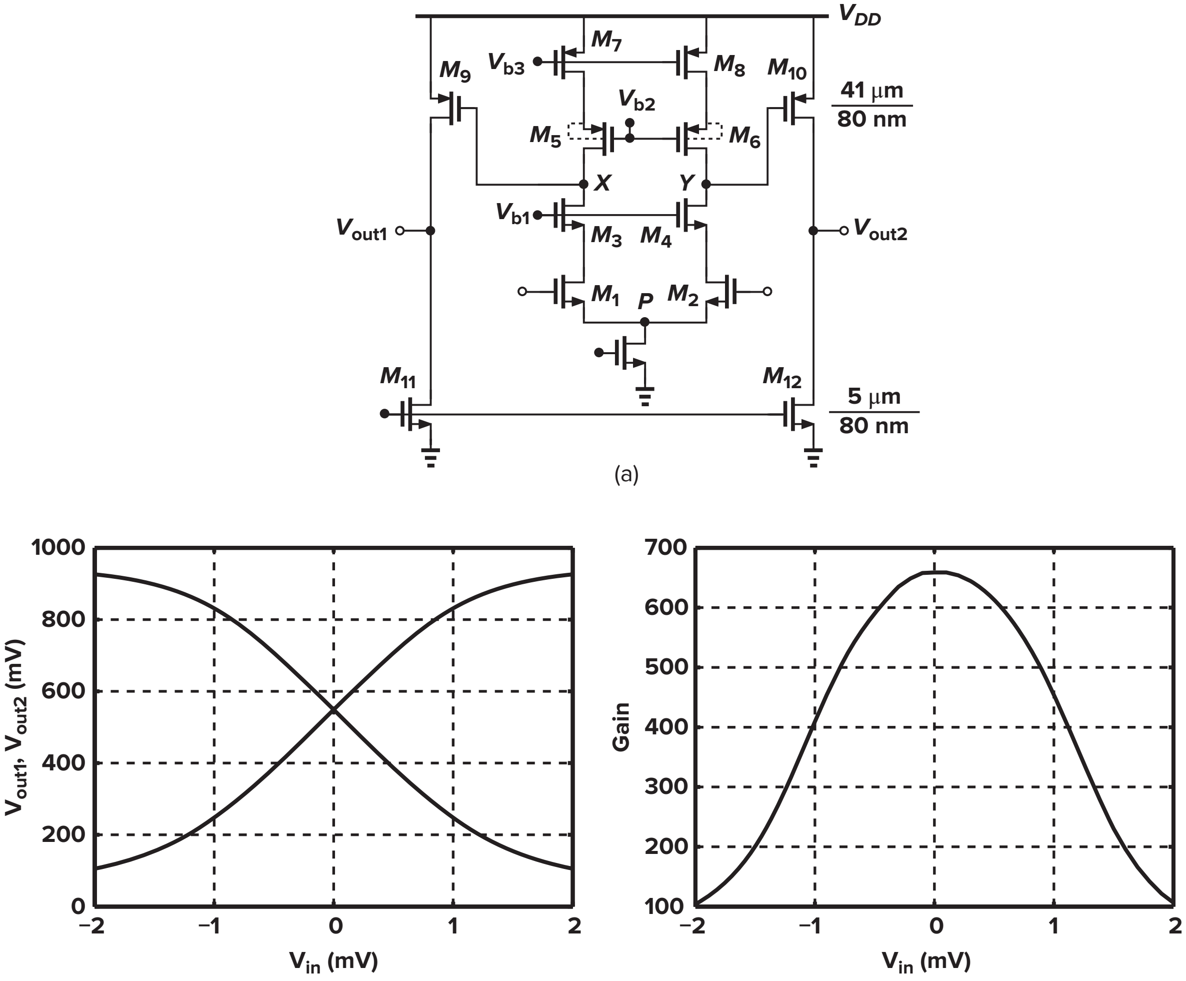
Second-Stage Design The second stage must provide a voltage gain of about 10, dictating channel lengths greater than 40 nm for both NMOS and PMOS devices. Do we use an NFET or a PFET for the input of the second stage? The need for gain may point to an NFET due to its higher $g_m r_O$, but we must examine the situation more closely. Bearing in mind that the output CM level of the first stage is around 0.55 V, let us consider a transistor having $W/L = 10$ μm/80 nm and determine its $g_m r_O$ if it is an NFET with $V_{GS} \approx 0.55$ V or a PFET with $|V_{GS}| \approx 0.45$ V. Using simulations, we obtain $(g_m r_O)_N = 12.8$ and $r_{ON} = 1.86$ kΩ at $V_{DS} = 0.5$ V and $I_D = 900$ μA, and $(g_m r_O)_P = 17.5$ and $r_{OP} = 9.75$ kΩ at $|V_{DS}| = 0.5$ V and $|I_D| = 110$ μA. We therefore select the PFET and scale its width to (450 μA/110 μA) $\times$ 10 μm $\approx$ 41 μm to accommodate the nominal bias current. With $W/L = 41$ μm/80 nm, such a device exhibits an output resistance of 2.38 kΩ. The drain of the PFET is tied to an NMOS current source.

The NMOS current source output resistance must not lower the gain of the second stage, $|A_{v2}|$, below 10. Writing $|A_{v2}| = g_{mP}(r_{OP}||r_{ON}) \geq 10$, we have $r_{ON} \geq 1.33 r_{OP} = 3.0$ kΩ at $I_D = 475$ μA. If the 10-μm/80-nm NFET considered above with $r_O = 1.86$ kΩ and $I_D = 900$ μA is scaled down by a factor of 2, it yields $r_{ON} = 3.72$ kΩ, which is close to the desired value.

Figure 11.40(a) shows the op amp developed thus far, and Fig. 11.40(b) plots the input-output characteristics. In order to determine the maximum output swing that the op amp can handle, we plot the slope of the differential characteristic in Fig. 11.40(c), noting that the differential output cannot exceed 450 mV if the gain must not drop below 500. To resolve this issue, we double the width and length of the output NFETs, raising the gain and arriving at the results depicted in Fig. 11.41. Now, the single-ended swing reaches 530 mV for a minimum gain of 500. Of course, the gain *variation* (nonlinearity) is unabated, posing difficulties in some applications.

Common-Mode Feedback As explained in Chapter 9, two-stage op amps generally require CMFB for both stages. For the first stage, we can utilize the CMFB scheme illustrated in Fig. 11.29. We therefore focus on CMFB for the second stage.

The second stage can also incorporate the method of Fig. 11.29 and control the NMOS current sources. However, the lower output impedance here allows the use of resistors for direct sensing of the CM level,



(a)

(b)

(c)

Figure 11.40 (a) Two-stage op amp design, (b) its input-output characteristics, and (c) its gain variation.

simplifying the design. Consider the topology depicted in Fig. 11.42(a), where R_1 and R_2 ($\approx$ 30 kΩ) reconstruct the CM level at node G, applying the result to the gates of M_{11} and M_{12}. Under equilibrium, the resistors draw no current, establishing an output CM voltage equal to $V_{GS11,12}$. This voltage varies by about 50 mV with PVT, a value that can be tolerated in this design. Note that this CMFB loop is stable.

What if $V_{GS11,12}$ is not close to the desired output CM level? As shown in Fig. 11.42(b), if we inject a current I_B into node G, the output CM level is shifted by $I_B R_1/2 (= I_B R_2/2)$. For example, a shift of 100 mV requires a current of (100 mV/30 kΩ) $\times$ 2 $=$ 6.7 μA. A positive I_B shifts the CM level down and vice versa.

Frequency Compensation The two-stage op amp designed above contains several poles and most likely demands compensation. Recall from Chapter 10 that the first nondominant pole of two-stage op amps typically arises from the output node and hence depends on the load capacitance, C_L. The stability

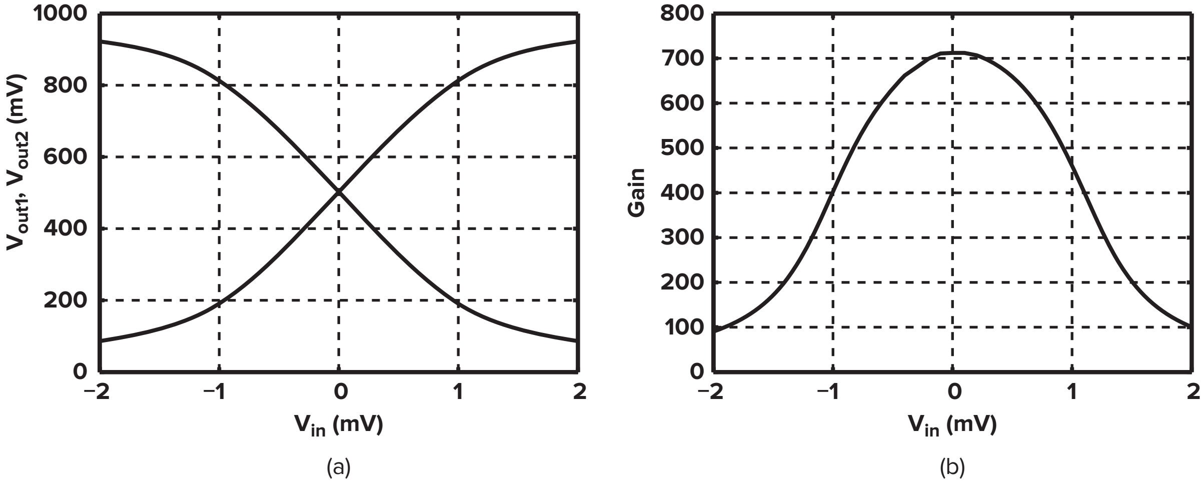


Figure 11.41 (a) Input-output characteristics and (b) gain variation of modified two-stage op amp with $(W/L)_{11,12} = 10\ \mu$m/0.16 μm.

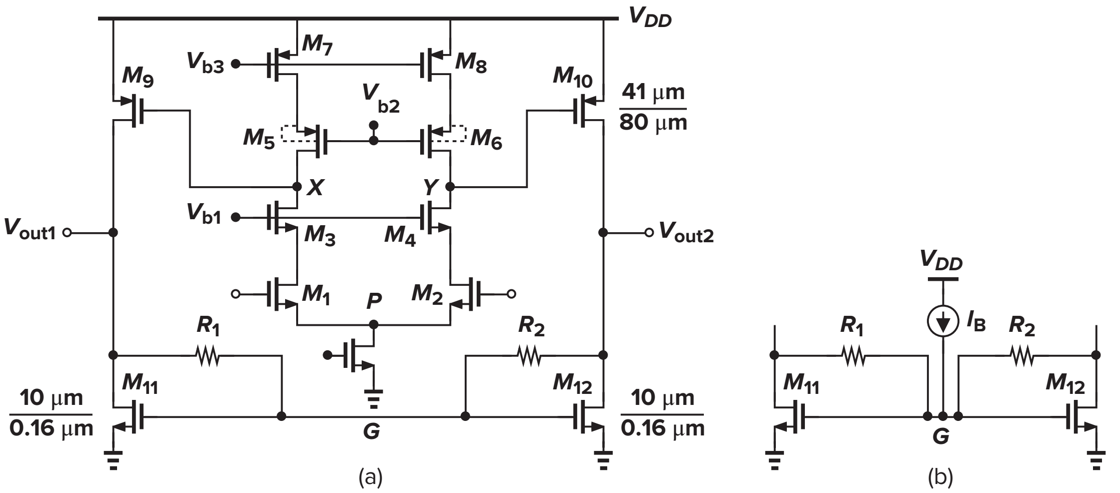


Figure 11.42 (a) Simple common-mode feedback around the second stage; (b) injection of current to shift CM level.

analysis must therefore assume a value of C_L, which itself is given by the environment in which the op amp is used. Let us choose a single-ended load capacitance of 1 pF in this example, obtaining an output pole frequency of around 90 MHz. We begin the study with the open-loop op amp, bearing in mind that the feedback network may add its own effects and eventually require further modifications.

Plotted in Fig. 11.43 are the open-loop (differential) magnitude and phase response of the circuit, revealing a low-frequency gain of 57 dB ($\approx$ 700), a unity-gain frequency of 3.2 GHz, and a phase margin of about -8°. This bandwidth appears very impressive, but we also note that the phase reaches -120° at 240 MHz. In other words, after compensation for 60° phase margin, the unity-gain bandwidth can drop by a factor of 13!

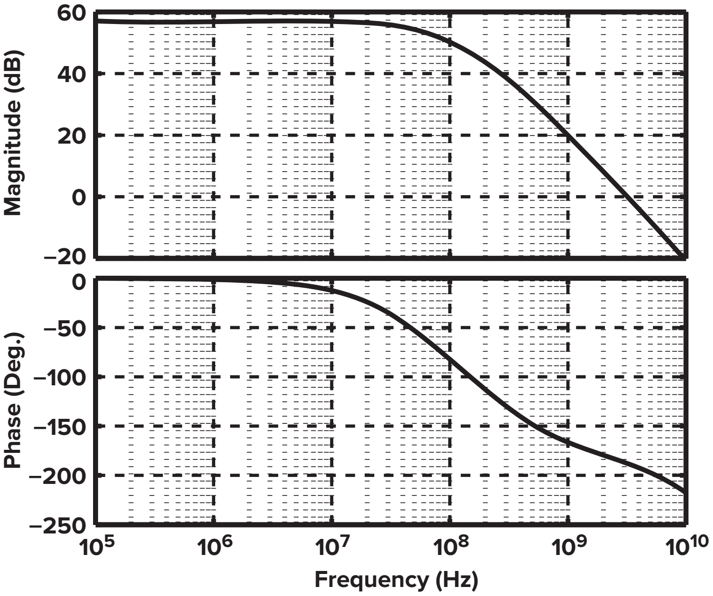


Figure 11.43 Frequency response of open-loop op amp.

▶ Example 11.11

The above results are rather curious: the output pole is located around 90 MHz, suggesting a phase of about $-135°$ at this frequency, but the actual phase is around $-85°$ at this frequency. Explain the reason for this behavior.

Solution

In the above design, we cannot say that the phase reaches $-135°$ at the second pole because the poles are not widely spaced. In fact, the pole at X in Fig. 11.42 is around 95 MHz. The pole at X and the output pole produce at 90 MHz a phase shift of $-\tan^{-1}(90/95) - \tan^{-1}(90/90) \approx -88°$.

We express the phase shift at 240 MHz due to these two poles as $-\tan^{-1}(240\text{ MHz}/95\text{ MHz}) - \tan^{-1}(240\text{ MHz}/90\text{ MHz}) = -138°$. Why does this result disagree with the simulated value of $-120°$? This is because the gate-drain capacitance of the output PMOS transistor creates some pole splitting, raising the output pole beyond 90 MHz and lowering the pole at X below 95 MHz.

◀

In order to compensate the op amp, we begin at 240 MHz and 0 dB on the magnitude plot in Fig. 11.43 and draw a straight line toward the y-axis with a slope of -20 dB/dec. The frequency at which this line intersects the magnitude plot is roughly equal to 240 MHz/700 = 344 kHz (why?), yielding the desired value for the dominant pole.

Which node should produce the dominant pole: X or the output node? We prefer the former for two reasons, namely, a smaller compensation capacitance due to Miller multiplication and pole splitting; neither of these benefits accrues if the dominant pole is established at the output.

With the output resistance of 8 kΩ seen at node X and a voltage gain of about 12 provided by the output stage, we choose a Miller compensation capacitor, C_C, equal to 4.5 pF so as to create a 344-kHz pole at this node. Figure 11.44 shows the resulting open-loop frequency response, confirming that the dominant pole is now located around 340 kHz. Unfortunately, the gain crossover occurs at 350 MHz and the phase margin is only 18° because the zero introduced by C_C, $\omega_z = g_{m10}/C_C$, is as low as 250 MHz. As explained in Chapter 10, we can insert a resistor, R_z, in series with C_C so as to move the zero to the second pole, ω_{p2}. The second pole can be roughly estimated from Fig. 11.44 as the frequency at

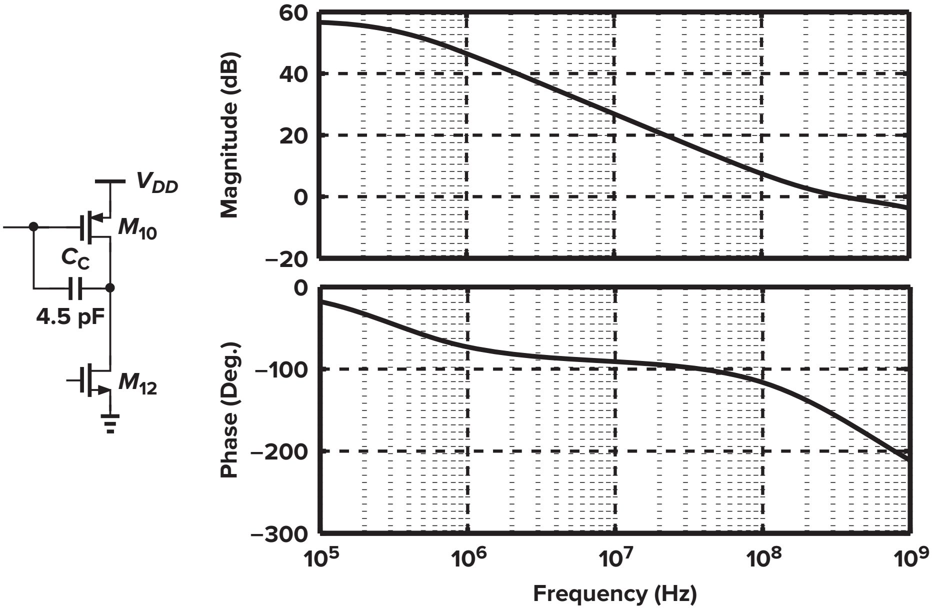


Figure 11.44 Frequency response of compensated open-loop op amp with $C_C = 4.5$ pF.

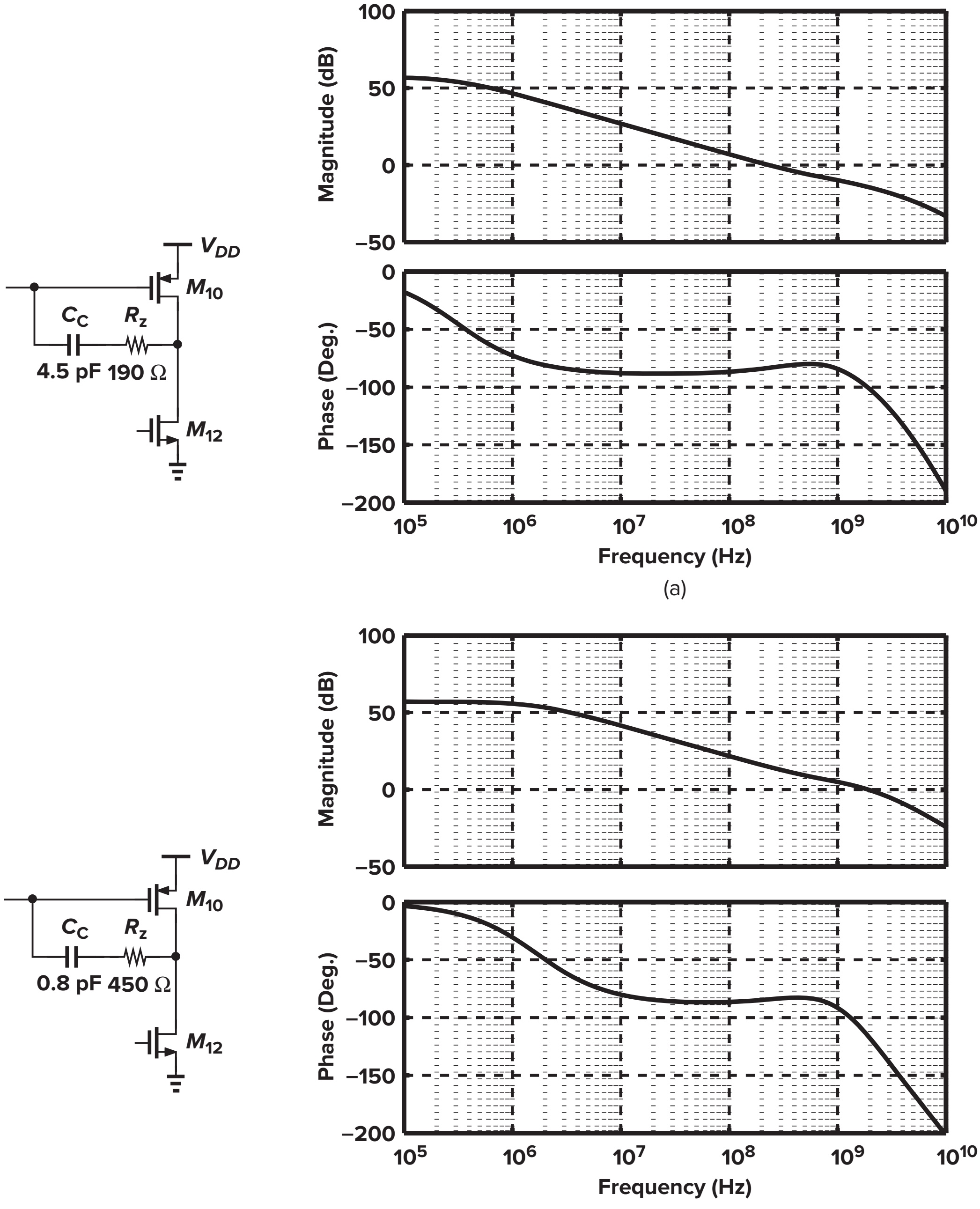
which $\angle H$ reaches -135° and is equal to 185 MHz. Selecting R_z according to $(\omega_{p2}C_C)^{-1} = 190\ \Omega$, we observe the response depicted in Fig. 11.45(a). The phase margin rises to 96° because of the pole-zero cancellation.

The phase margin revealed by Fig. 11.45(a) suggests that the compensation capacitor can be smaller and the unity-gain bandwidth larger. By some iteration, we choose $C_C = 0.8$ pF and $R_z = 450\ \Omega$, arriving at the response shown in Fig. 11.45(b). Remarkably, the op amp now exhibits a unity-gain bandwidth of 1.9 GHz with a phase margin of 65°.

Closed-Loop Behavior We now configure the op amp as a closed-loop amplifier having a nominal gain of 2 and a load capacitance of 1 pF [Fig. 11.46(a)]. The small-signal transient response appears as shown in Fig. 11.46(b), exhibiting significant ringing. Why does this happen despite the 65° phase margin obtained above? This is due to the large resistance values used in the feedback network. We draw the single-ended equivalent as in Fig. 11.47(a) for the loop transmission calculation, observing that an open-loop pole around $[2\pi(100\ \text{k}\Omega||50\ \text{k}\Omega)C_{in}]^{-1} \approx 95$ MHz is formed at the input of the op amp.

In order to improve the closed-loop stability, we can reduce R_1 and R_2 in Fig. 11.47(a) to, say, 25 kΩ and 50 kΩ, respectively, before the open-loop gain falls appreciably, but this remedy only doubles the input pole frequency. Alternatively, we can increase the resistance in series with the compensation capacitors from 450 Ω to 1500 Ω, arriving at the response shown in Fig. 11.47(b). The circuit now settles much faster.

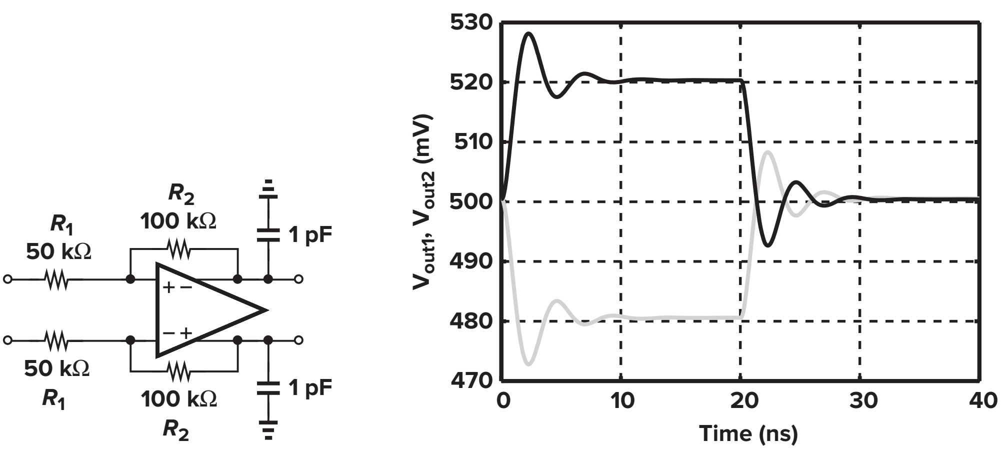
We conclude this section with two remarks. First, the op amp has been compensated for unity-gain feedback, whereas the topology of Fig. 11.46 operates with a feedback factor of $50\ \text{k}\Omega/150\ \text{k}\Omega = 1/3$. Thus, the compensation capacitance can be reduced to lower the phase margin to around 60°. Second, the design has assumed the “typical-NMOS, typical-PMOS” (TT) corner of the process, a temperature of 27°C, and a constant supply of 1 V. In practice, we must account for other corners (e.g., SS or FF), the required temperature range (e.g., 0°C to 75°C), and supply variations (e.g., by ±5%). To meet the specifications under all of these conditions, the design must often be more conservative in terms of gain, swings, and power consumption than that presented here.



(a)

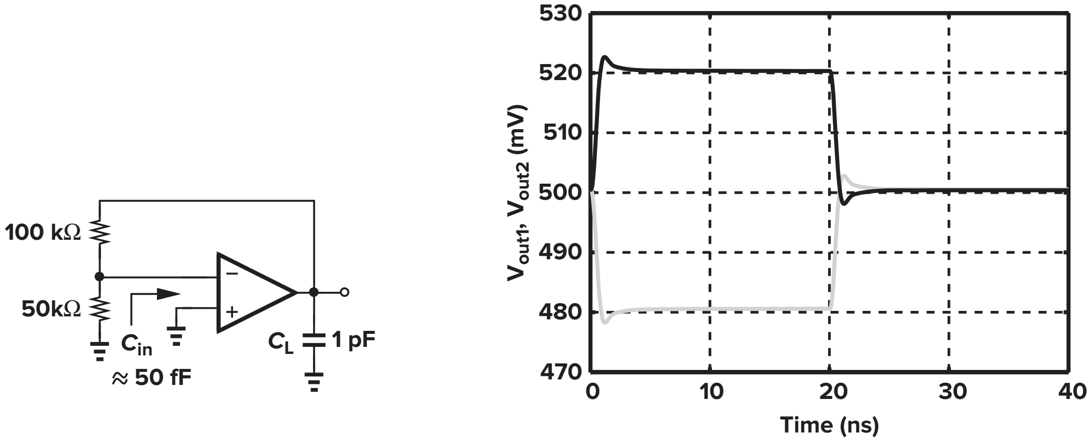
(b)

Figure 11.45 Frequency response of open-loop amp with (a) $C_C = 4.5$ pF and $R_Z = 190\ \Omega$, and (b) $C_C = 0.8$ pF and $R_Z = 450\ \Omega$.



(a) (b)

Figure 11.46 (a) Closed-loop amplifier and (b) its step response.



(a) (b)

Figure 11.47 (a) Equivalent circuit of closed-loop amplifier and (b) step response with $R_z = 1500\ \Omega$.

11.6 ■ High-Speed Amplifier

Some applications require an amplifier with fast settling and accurate gain. For example, "pipelined" ADC architectures can tolerate but a small gain error in their constituent amplifiers. In this section, we design a differential amplifier according to the following specifications:

- Voltage gain $= 4$
- Gain error $\leq 1\%$
- Differential output swing $= 1\ \mathrm{V}_{pp}$
- Load capacitance $= 1$ pF
- Step response settling time to 0.5% accuracy $= 5$ nS
- $V_{DD} = 1$ V

As illustrated in Fig. 11.48, the settling time, t_s, is defined as the time necessary for the output to reach within 0.5% of its final value. Our objective is to minimize the power consumption of the circuit.

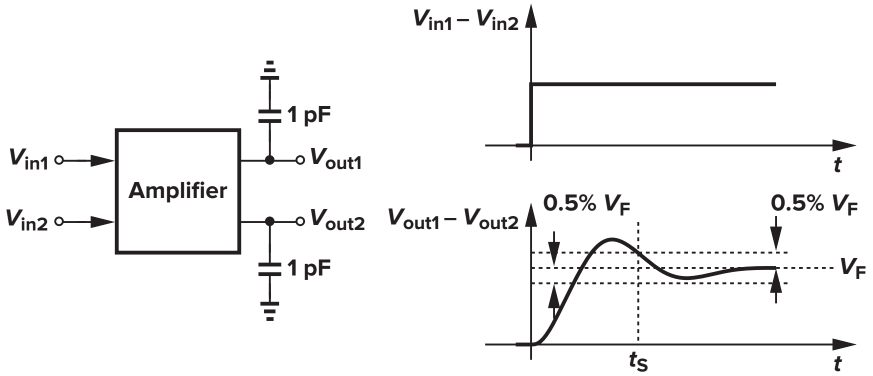


Figure 11.48 Definition of settling time.

11.6.1 General Considerations

Precision Issues With a myriad of amplifier topologies, where do we begin? In this case, the specifications readily narrow down our choices. The maximum tolerable gain error of 1% indicates a *closed-loop* configuration so that the gain can be defined as the ratio of two passive component values and remain relatively independent of PVT. We must therefore design an amplifier whose open-loop gain is high enough to yield a closed-loop gain error of less than 1%. This observation along with the required output swing of 1 V_{pp} calls for a two-stage op amp.

We have now arrived at the feedback arrangement shown in Fig. 11.49, where the closed-loop gain is given by

$$\frac{V_{out}}{V_{in}} = -\frac{R_2}{R_1}\frac{1}{1+(1+\frac{R_2}{R_1})\frac{1}{A_0}} \tag{11.14}$$

$$\approx -\frac{R_2}{R_1}\left[1-\left(1+\frac{R_2}{R_1}\right)\frac{1}{A_0}\right] \tag{11.15}$$

We choose $R_2/R_1 = 4$ and ensure that the gain error falls below 1%:

$$\left(1+\frac{R_2}{R_1}\right)\frac{1}{A_0} \le 0.01 \tag{11.16}$$

thereby obtaining $A_0 \ge 500$. This calculation neglects the loading of the feedback network on the op amp.

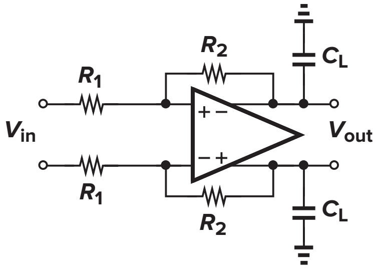


Figure 11.49 Closed-loop amplifier with resistive feedback.

▶ **Example 11.12**

Determine the closed-loop output impedance and bandwidth of the above topology in terms of the open-loop op amp characteristics.

Solution

Drawing the half-circuit equivalent for loop gain calculation as shown in Fig. 11.50 and applying a test signal, V_t, we observe that the feedback network senses V_{out} and returns a fraction equal to $\beta = [R_1/(R_1 + R_2)]V_{out}$ to the input. The loop gain is therefore equal to $\beta A_0 = A_0 R_1/(R_1 + R_2) \approx A_0/5 \approx 100$, indicating that the output resistance falls by a factor of 100 as a result of feedback. The bandwidth rises by the same factor.

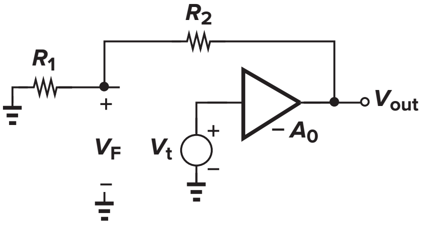


Figure 11.50

◀

The use of a resistive feedback network poses a difficulty: as explained in Sec. 11.5.2, if R_1 and R_2 are large enough not to reduce the open-loop gain of the op amp, then they form a significant pole with the input capacitance, degrading the phase margin. We therefore consider *capacitive* feedback instead, configuring the circuit as shown in Fig. 11.51(a). The closed-loop gain is now approximately equal to C_1/C_2, or more accurately (Chapter 13):

$$\frac{V_{out}}{V_{in}} \approx -\frac{C_1}{C_2}\left(1 - \frac{C_1 + C_2 + C_{in}}{C_2}\frac{1}{A_0}\right) \tag{11.17}$$

where C_{in} denotes the (single-ended) input capacitance of the op amp. Drawing the single-ended counterpart for the loop transmission calculation [Fig. 11.51(b)], we observe that C_1 and C_2 do *not* contribute additional poles because $(C_1 + C_{in})C_2/(C_1 + C_{in} + C_2)$ simply appears in parallel with C_L. (As explained in Chapter 13, the capacitors also allow sampling and discrete-time operation.)

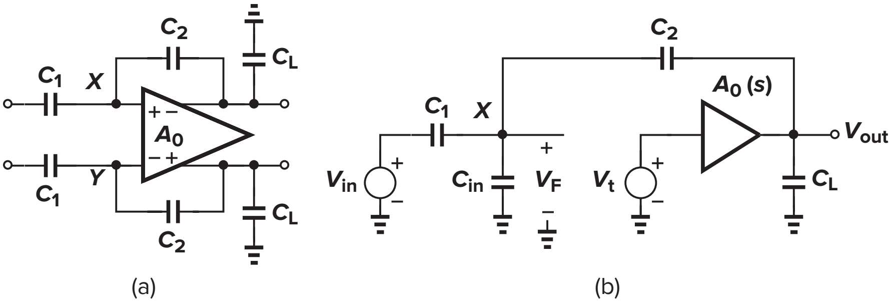


Figure 11.51 (a) Closed-loop amplifier using capacitive feedback, and (b) its simplified equivalent for loop gain calculation.

▶ **Example 11.13**

Circuits incorporating capacitive "coupling" typically exhibit a high-pass response. Is that the case for the amplifier of Fig. 11.51(a)?

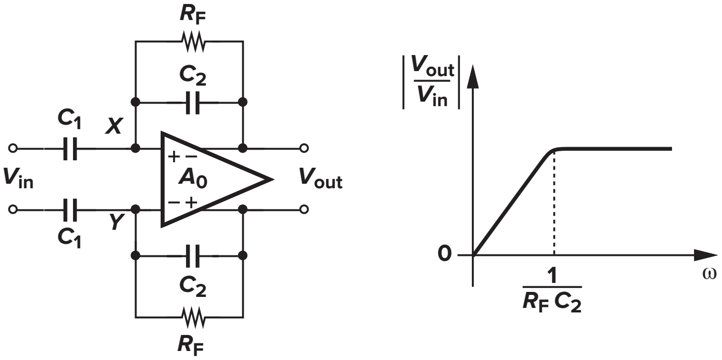
Solution

No, it is not. Since there is no resistive path to X and Y, the time constant at these nodes is infinite (if leakage currents are neglected), yielding a frequency response that extends to nearly $f = 0$. The reader can consider the frequency response of a simple capacitive divider as an example to appreciate this property.

The circuit of Fig. 11.51(a) provides no bias for the op amp inputs, i.e., the dc levels at X and Y are not defined and can assume any value. (In the presence of gate leakage at the inputs, these nodes charge to V_{DD} or discharge to ground.) Illustrated in Fig. 11.52, a simple remedy is to add two feedback resistors so that the input and output dc levels become equal. However, the finite time constant at X and Y leads to a high-pass response; if $A_0 = \infty$, then

$$\frac{V_{out}}{V_{in}}(s) = -\frac{R_F || \dfrac{1}{C_2 s}}{\dfrac{1}{C_1 s}} \tag{11.18}$$

$$= -\frac{R_F C_1 s}{R_F C_2 s + 1} \tag{11.19}$$



(a) (b)

Figure 11.52 Addition of feedback resistors to define input dc levels and the resulting transfer function.

The corner frequency, $1/(2\pi R_F C_2)$, must therefore be chosen less than the minimum input frequency of interest, a condition that is not practical in all applications. As explained in Chapter 13, R_F can be replaced with a switch, but we proceed here assuming that $R_F C_2$ is sufficiently large. In other words, we assume that the circuit reduces to that in Fig. 11.51(a) for the frequencies of interest.

Equation (11.17) indicates that the capacitive-feedback amplifier's gain error also depends on C_{in}. For example, if $C_{in} \approx (C_1 + C_2)/5$, then A_0 must be 20% higher than the value dictated by Eq. (11.16). We can choose $C_1 + C_2 \gg C_{in}$, but at the cost of settling speed.

Speed Issues The amplifier must settle to 0.5% accuracy in 5 ns. Let us first assume a linear, first-order circuit and write the step response as

$$V_{out}(t) = V_0 \left(1 - \exp\frac{-t}{\tau}\right) u(t) \tag{11.20}$$

The time necessary for V_{out} to reach $0.995V_0$ is $t_s = -\tau \ln 0.005 = 5.3\tau$; i.e., τ must be no more than 0.94 ns. Thus, the closed-loop amplifier must achieve a −3-dB bandwidth of at least $1/(2\pi \times 0.94 \text{ ns}) \approx$ 170 MHz.

If the op amp in Fig. 11.51(a) is modeled simply by a dependent current source, $G_m V_{in}$, and an output resistance, R_{out}, then the closed-loop time constant is given by (Chapter 13)

$$\tau = \frac{C_L(C_1 + C_{in}) + C_L C_2 + C_2(C_1 + C_{in})}{G_m C_2} \tag{11.21}$$

where $G_m R_{out}$ is assumed much greater than unity. This expression can be rewritten as

$$\tau = \left(\frac{C_1 + C_2 + C_{in}}{C_2}\right) \frac{C_L + \dfrac{C_2(C_{in} + C_1)}{C_2 + C_{in} + C_1}}{G_m} \tag{11.22}$$

suggesting that the op amp sees the series combination of C_2 and $C_1 + C_{in}$ in parallel with C_L, and its G_m is reduced by the feedback factor, $C_2/(C_1 + C_2 + C_{in})$ (Fig. 11.53) (Chapter 13).

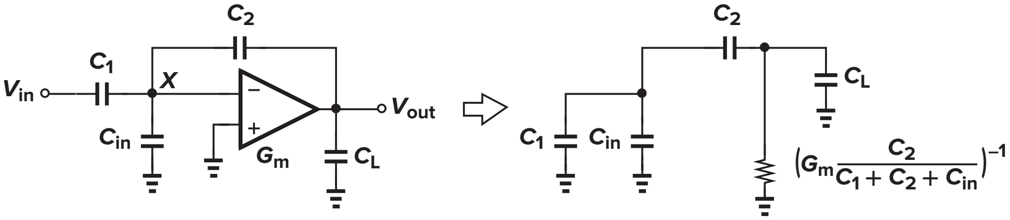


Figure 11.53 Representation of closed-loop time constant by an equivalent network.

The foregoing model is not accurate for a two-stage op amp because the internal pole (at the output of the first stage) inevitably affects the response. Let us improve our approximation by considering a frequency-compensated two-stage op amp. Recall that if the loop gain falls to 1 at the second pole, ω_{p2}, the phase margin is about 45° for unity-gain feedback.

How do we compensate the op amp for a closed-loop gain of 4 (rather than 1)? In this case, $|\beta H|$ (rather than $|H|$) must fall to 0 dB at ω_{p2} (i.e., the circuit is not compensated for unity-gain feedback). As illustrated in Fig. 11.54(a), we begin at $\omega = \omega_{p2}$ and draw a line with a slope of −20 dB/decade toward the y-axis, seeking its intercept with the plot of $|\beta H|$. We calculate the location of the compensated dominant pole, ω'_{p1}, as follows. Between ω'_{p1} and ω_{p2}, we can approximate the compensated $\beta H(s)$ as $\beta A_0/(1 + s/\omega'_{p1})$; we set its magnitude to 1 at ω_{p2}: $|\beta A_0/(1 + j\omega_{p2}/\omega'_{p1})| = 1$. It follows that

$$\omega'_{p1} \approx \sqrt{\frac{\omega_{p2}^2}{\beta^2 A_0^2 - 1}} \tag{11.23}$$

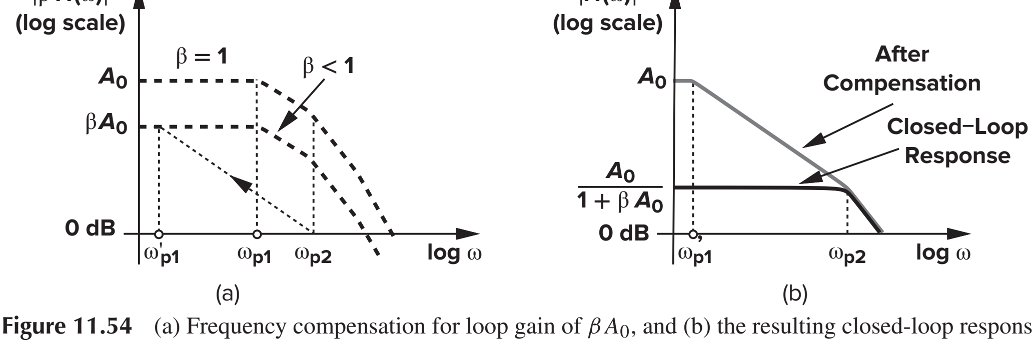


Figure 11.54 (a) Frequency compensation for loop gain of βA_0, and (b) the resulting closed-loop response.

and

$$\omega'_{p1} \approx \frac{\omega_{p2}}{\beta A_0} \tag{11.24}$$

As expected, ω'_{p1} must be chosen lower in magnitude if β increases, i.e., if the feedback becomes stronger.

Example 11.14

We compensate an op amp for $\beta = 1/5$ and PM $= 45°$. Plot the open-loop frequency response of the op amp, H.

Solution

On a logarithmic scale, the plot of H is obtained if we shift the plot of $|\beta H|$ up by $-20 \log \beta$. As shown in Fig. 11.55, H begins to fall at ω'_{p1} and reaches a value of approximately $1/\beta$ at ω_{p2}.

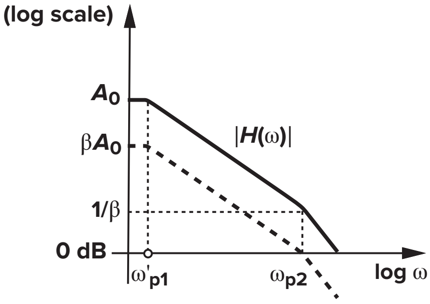


Figure 11.55

Let us now construct the closed-loop frequency response with this choice of ω'_{p1}. To this end, we first plot the magnitude of the loop transmission, $|\beta H|$, for $\beta = 1$ and after compensation [Fig. 11.54(b)]. The closed-loop response begins at $A_0/(1 + \beta A_0)$ at low frequencies and begins to roll off at $\omega \approx \omega_{p2}$. From another perspective, since the ratio of the open-loop and closed-loop gains at ω'_{p1} is approximately equal to βA_0 and since the open-loop gain falls at 20 dB/dec (i.e., in proportion to ω), the two responses intersect at $\omega \approx \beta A_0 \omega'_{p1} \approx \omega_{p2}$. We therefore choose this bandwidth equal to $2\pi(170\text{ MHz})/125 = 2\pi(1.36\text{ MHz})$.

In summary, the closed-loop gain and settling speed requirements have translated to a dominant pole at 1.36 MHz and a second pole at 170 MHz. The open-loop gain must fall from 500 at low frequencies to 4 at the second pole. These values assume a phase margin of 45° and must be eventually revisited.

11.6.2 Op Amp Design

Based on the foregoing calculations, we seek a two-stage op amp with an open-loop gain of 500, a dominant pole at 1.36 MHz, a second pole at 170 MHz, and a differential output swing of 1 V_{pp}. We thus return to the prototype designed in Sec. 11.5.2 and see whether it can serve our purpose. Most of the specifications of that op amp are the same as those needed here. But, since the compensation can be relaxed to suit a feedback factor of 1/5, the dominant pole of the op amp need not be as low as 344 kHz. Equation (11.24) suggests that, if the feedback factor is reduced from 1 to β, then the dominant pole can increase by roughly a factor of $1/\beta$. We therefore expect that the compensation capacitor leading to the response in Fig. 11.45(a) can be lowered from 4.5 pF to 0.9 pF, raising the dominant pole frequency from 340 kHz to 1.7 MHz. For the feedforward zero to cancel the second pole, R_2 must rise by the same factor, reaching 950 Ω.

As observed in Sec. 11.5.2, the zero-pole cancellation creates a greater phase margin, allowing us to reduce C_C from 4.5 pF to 0.8 pF. However, in the present design, the feedback network capacitors also load the output stage, lowering the *nondominant* pole, ω_{p2}. Since this effect has not been included in our calculations, we resist the temptation to reduce C_C for now and proceed to study the closed-loop behavior.

11.6.3 Closed-Loop Small-Signal Performance

Figure 11.56 shows the overall op amp and its closed-loop environment. For a nominal gain of 4, we choose $C_1 = 1$ pF and $C_2 = 0.25$ pF. With $C_{in} \approx 50$ fF, Eq. (11.17) predicts a gain error of less than 1% if $A_0 > 520$. This gain is slightly greater than that achieved by the op amp at its peak output swings (Fig. 11.41), but we deal with this issue later. For transient studies, R_F must be large enough not to cause significant "droop" during time scales of interest. Specifically, for a settling time of 5 ns, we select $R_F C_2 > 10\ \mu$s so as to confine the discharge of the capacitors to well below 1%; i.e., $R_F = 40$ MΩ. (This extremely large value implies that a switched-capacitor solution is more practical.)

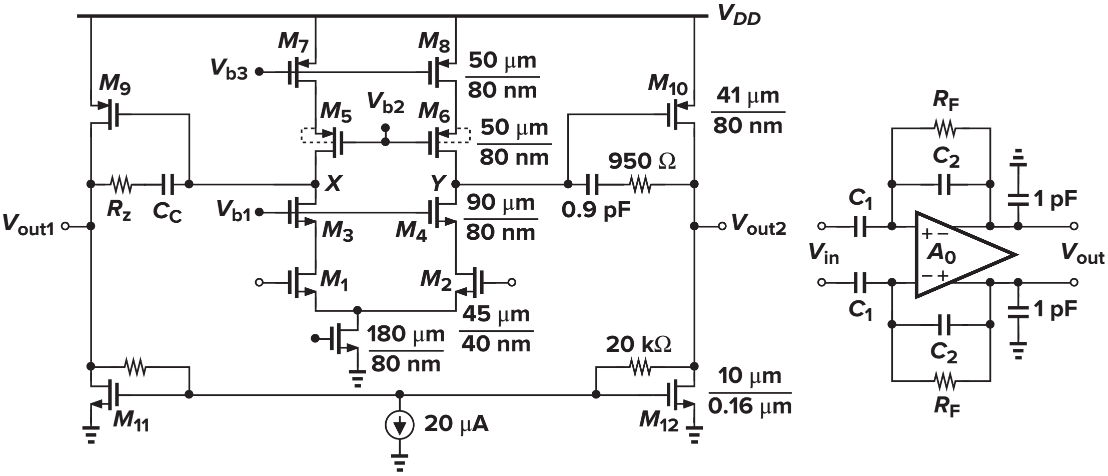


Figure 11.56 Overall compensated two-stage op amp and the closed-loop environment.

Let us apply a small step to the above circuit and examine the output behavior. With a differential input step of 25 mV, we expect an output around 99 mV (for 1% gain error). Depicted in Fig. 11.57(a) is the differential output waveform and in Fig. 11.57(b) a close-up showing the fine settling. We observe a final value equal to 98.82 mV, a result of insufficient open-loop gain.

How do we increase the gain? If we raise the length (and hence the width) of the first-stage input transistors in Fig. 11.56, C_{in} also increases, counteracting A_0 in Eq. (11.17). Instead, we double the (drawn) width and length of the NMOS cascode transistors, obtaining the output shown in Fig. 11.58. Now, the gain error is less than 1%.

▶ **Example 11.15**

Is it possible to increase the length of the PMOS devices in the first stage to raise the gain?

Solution

Designed to provide a much higher impedance than its NMOS counterpart, the PMOS cascode structures have a weak effect on the gain of the first stage. The NMOS cascode devices, on the other hand, directly determine the voltage gain (why?).

◀

Let us turn our attention to the settling behavior of the amplifier. If the output reaches 99.1 mV at $t = \infty$, how do we define the settling time to 1% precision? We must find the time at which

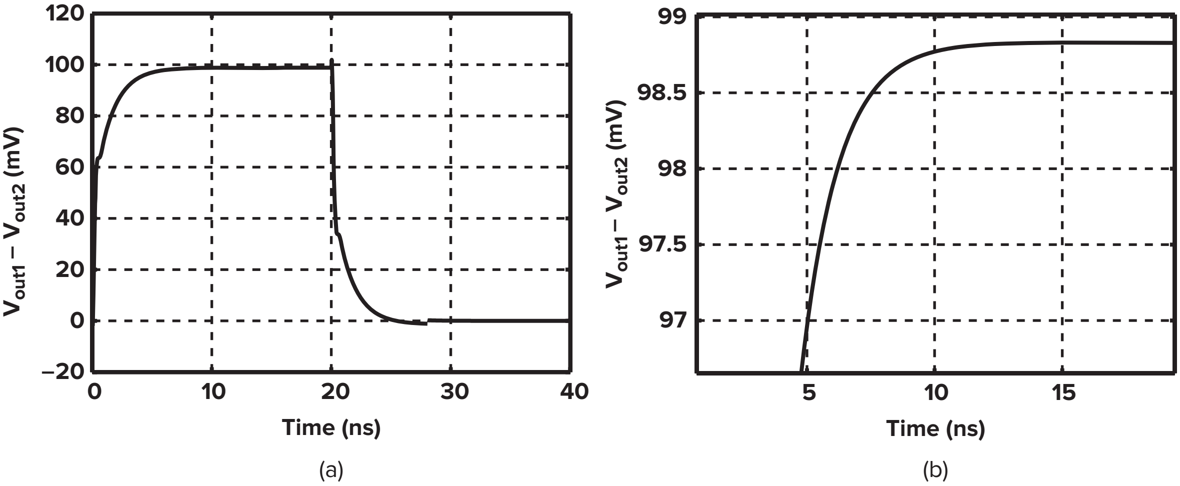


Figure 11.57 (a) Closed-loop step response and (b) close-up showing settling to 1% accuracy.

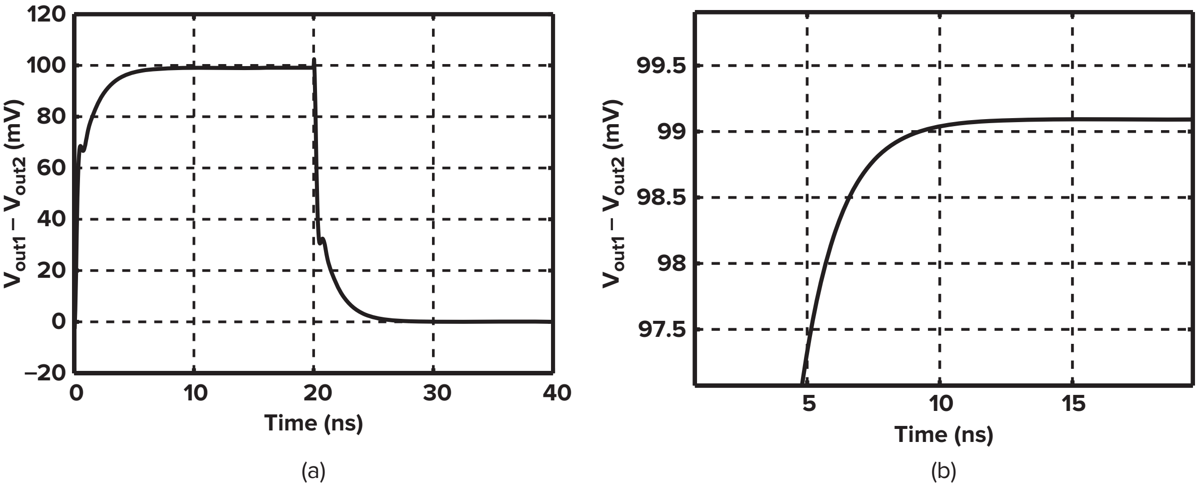


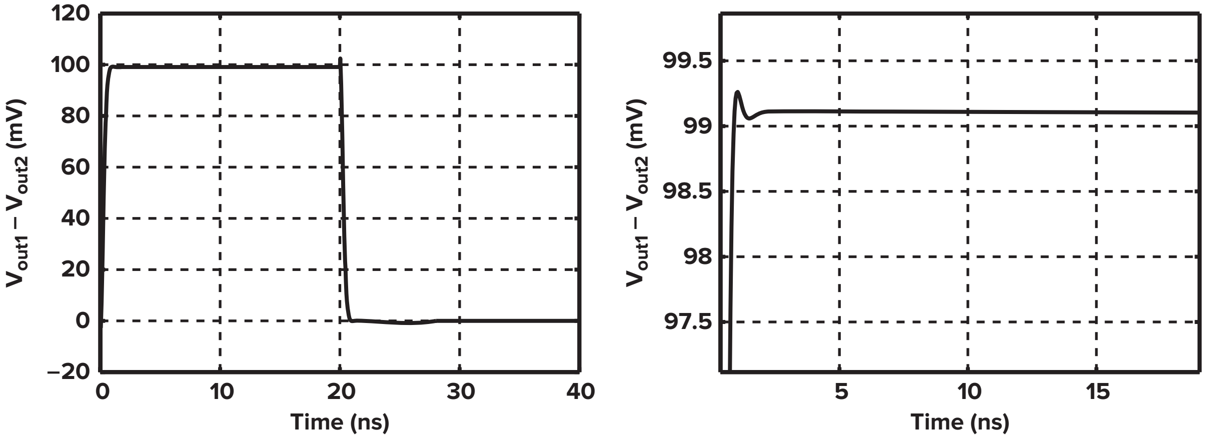
Figure 11.58 (a) Closed-loop step response and (b) close-up showing settling to 1% accuracy for $(W/L)_{3,4} = 180\ \mu\text{m}/0.16\ \mu\text{m}$.

$V_{out} = 99.1\text{ mV} \pm 0.01 \times 99.1\text{ mV} \approx 99.1\text{ mV} \pm 1\text{ mV}$. From the waveform in Fig. 11.58(b), we obtain $t_s \approx 5.8$ ns.

In order to improve the amplifier's speed, we recognize from Fig. 11.58(a) that the circuit is "overcompensated," i.e., the output appears overdamped. We can therefore return to our choice of C_C and R_z and adjust them more aggressively. We adjust these two values, patiently explore the design space, and examine the trends in the output behavior. With $C_C = 0.3$ pF and $R_z = 700\ \Omega$, we observe the settling shown in Fig. 11.59. The settling time drops to 800 ps, a remarkable improvement. Note that R_z is *reduced* in this case, thus moving the zero to *higher* frequencies.

11.6.4 Op Amp Scaling

If the settling time is so much shorter than the required value, can we trade speed for power dissipation? As explained in Chapter 9, a straightforward approach is to perform "linear scaling." We begin with the



(a) (b)

Figure 11.59 (a) Closed-loop step response and (b) close-up showing settling to 1% accuracy.

response shown in Fig. 11.60(a) and scale down all transistor widths and bias currents by a factor of α, thereby reducing the power by the same factor while retaining the voltage gain and the headroom. But how about C_C and R_z? We make four observations. (1) With the load capacitance fixed, the output pole (before compensation) is scaled down by a factor of α [Fig. 11.60(b)] because the output resistance of the second stage is scaled up by this factor (why?). (2) To maintain the same phase margin, the dominant

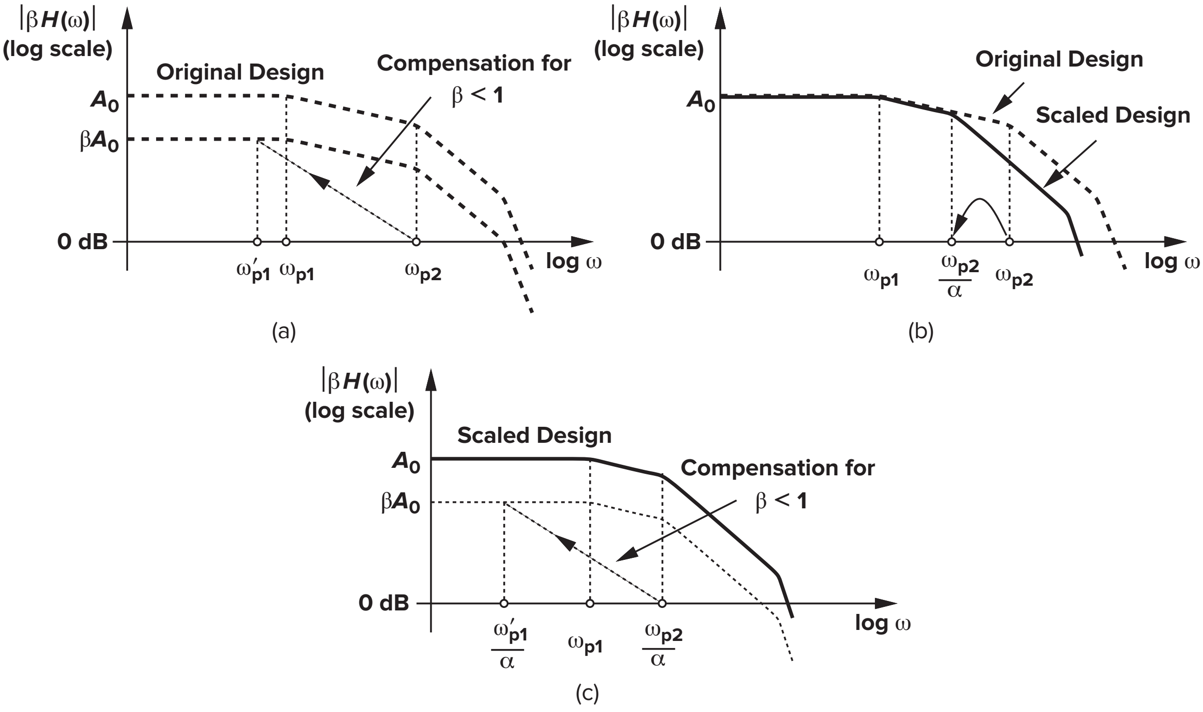
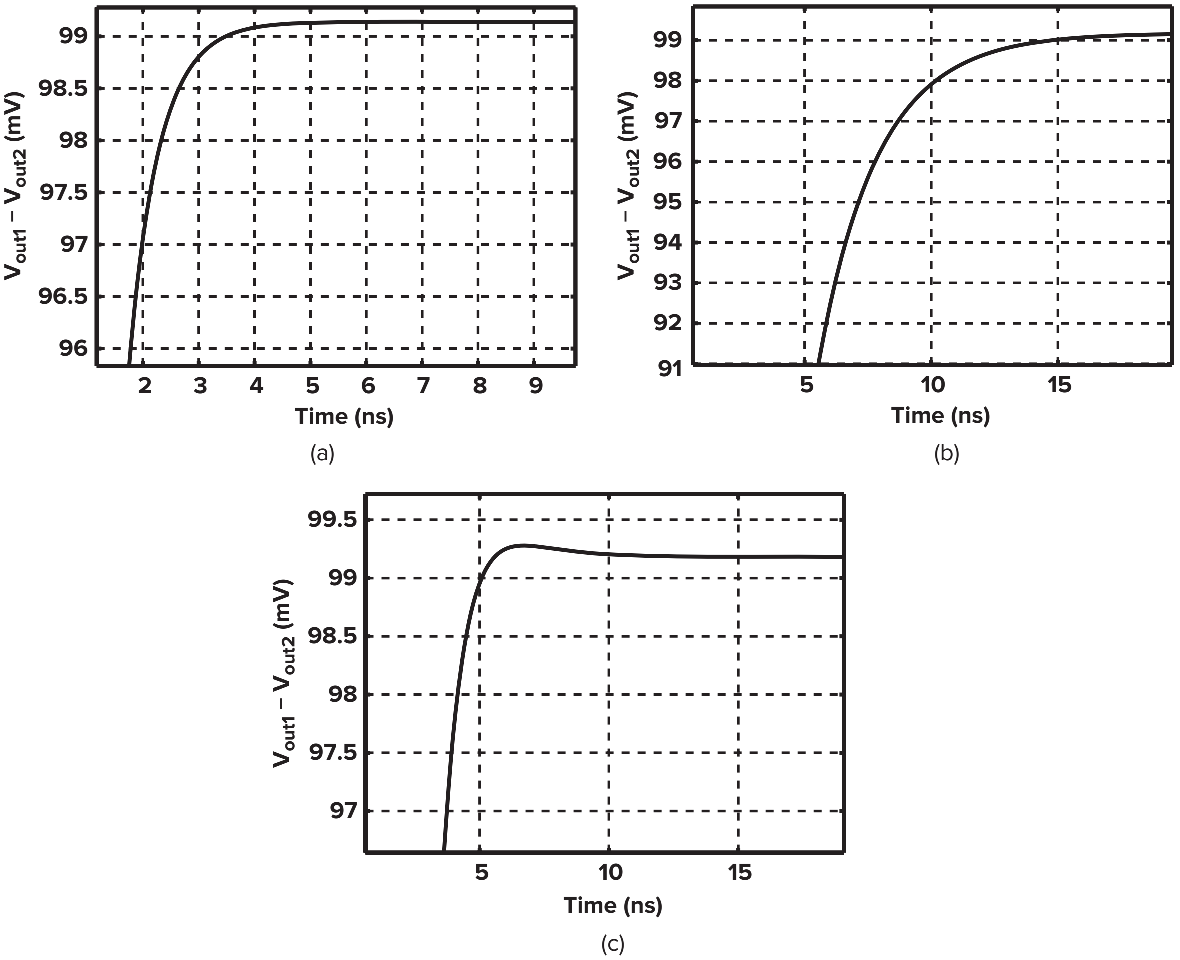


Figure 11.60 (a) Original op amp reponse and frequency compensation, (b) scaled op amp response, and (c) compensation of scaled op amp.

pole after compensation must also be scaled down by this factor [Fig. 11.60(c)]. (3) Since the output impedance of the first stage is multiplied by α, C_C should remain at its original value. (4) To place the zero introduced by R_z atop ω_{p2}/α, we multiply R_z by α (why?).

Let us try $\alpha = 2$ and examine the results. Figure 11.61(a) plots the output waveform, revealing the same final values as before and an overdamped response with $t_s \approx 2.5$ ns. We can then try scaling by another factor of 4 (i.e., $\alpha = 8$ with respect to the original design), observing the heavily overdamped transient shown in Fig. 11.61(b). Now, we adjust C_C and R_z manually to optimize the speed. With $C_C = 0.15$ pF and $R_z = 9$ kΩ, the step response appears as in Fig. 11.61(c), exhibiting $t_s \approx 4.5$ ns.



(a)

(b)

(c)

Figure 11.61 (a) Step response of op amp with scaling factor of 2, (b) step response of op amp with scaling factor of 8, and (c) modified design with $C_C = 0.15$ pF and $R_z = 9$ kΩ.

It is remarkable that linear scaling along with some adjustment of C_C and R_z affords an eightfold reduction in power (and area of the transistors). This scaling method entails minimal redesign effort because it does not alter the circuit's gain and swing values. Of course, the scaling gives rise to longer settling and higher noise (and offset). Figure 11.62 shows the scaled op amp design.[13]

[13] In practice, we do not rely on such precise values as 113 μA and 5.1 μm but round them to 115 μA and 5 μm, respectively.

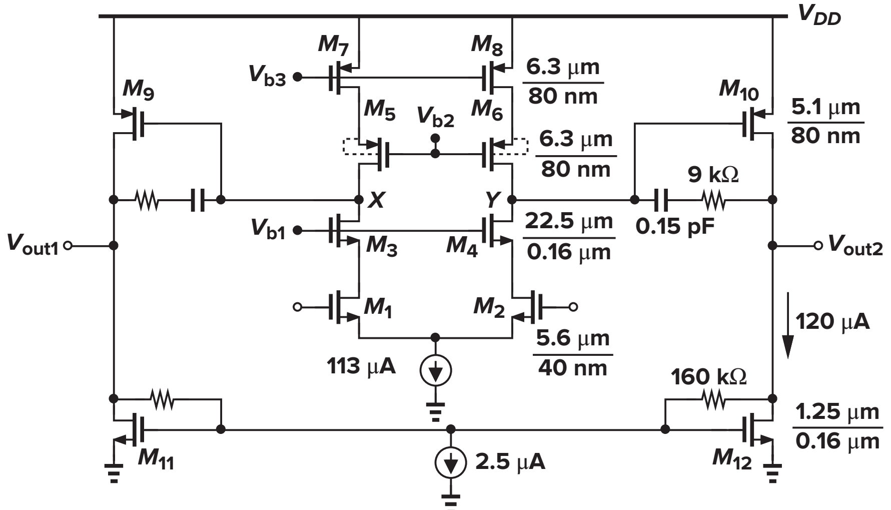


Figure 11.62 Scaled op amp design.

11.6.5 Large-Signal Behavior

The amplifier's ultimate test is with large output swings (1 $V_{pp,\text{diff}}$). Under this condition, the open-loop gain may drop as some transistors sustain less V_{DS}, and the speed may suffer as slewing may occur. In the previous simulations, the differential output begins from zero, jumps to some value, and returns to zero. For large-signal tests, however, V_{out} must swing from −0.5 V to +0.5 V, which can be accomplished by setting the initial differential conditions at the op amp inputs such that $V_{out} = -0.5$ V at $t = 0$. The result is shown in Fig. 11.63.

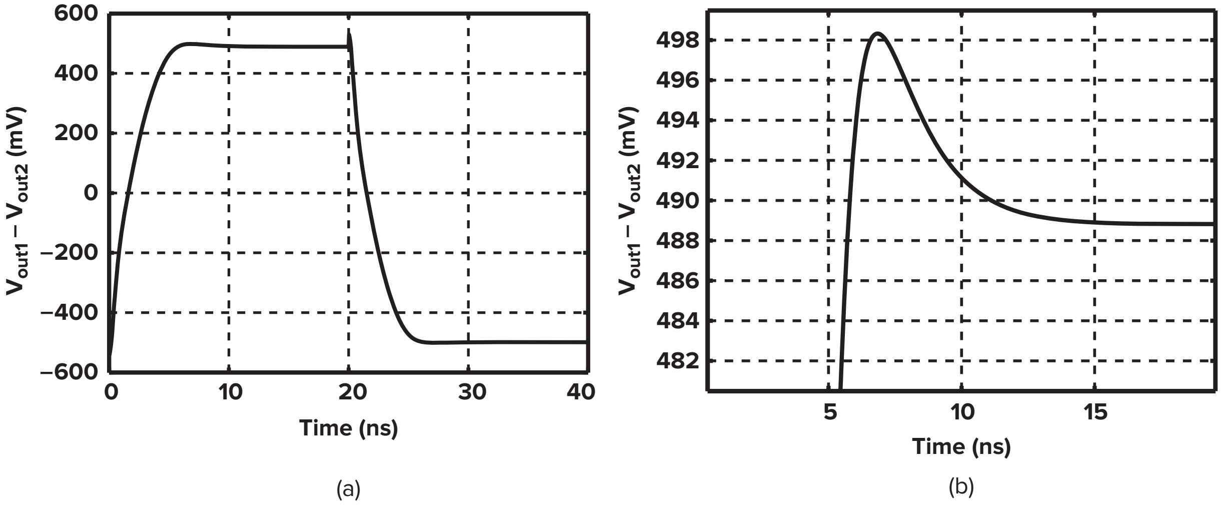


Figure 11.63 (a) Large-signal response of closed-loop amplifier, and (b) close-up of (a) showing settling to 1% accuracy.

We make two observations, First, the total change in V_{out} from $t \approx 20$ ns to $t \approx 40$ ns is equal to 987.4 mV, about 2.6 mV less than the allowed value for 1% gain error. Second, the settling to 1% from the final value is about 6 ns.

Let us first deal with the insufficient gain. We can measure the voltage gain of each stage under these conditions by dividing its differential output swing by its differential input swing (after the voltages have settled). We obtain $A_v = 39.5$ for the first stage and 10.2 for the second. (In small-signal operation, these values are equal to 46.3 and 11.2, respectively.) The open-loop gain has thus dropped from 518 to 403. To raise the gain, we double W and L of the NMOS cascode transistors ($W/L = 45\ \mu\text{m}/0.32\ \mu\text{m}$) in the first stage and the NMOS current sources in the second, arriving at the output shown in Fig. 11.64. The gain error is now less than 1%, but the settling has become longer because the pole associated with the source of the NMOS cascode transistors significantly degrades the phase margin.

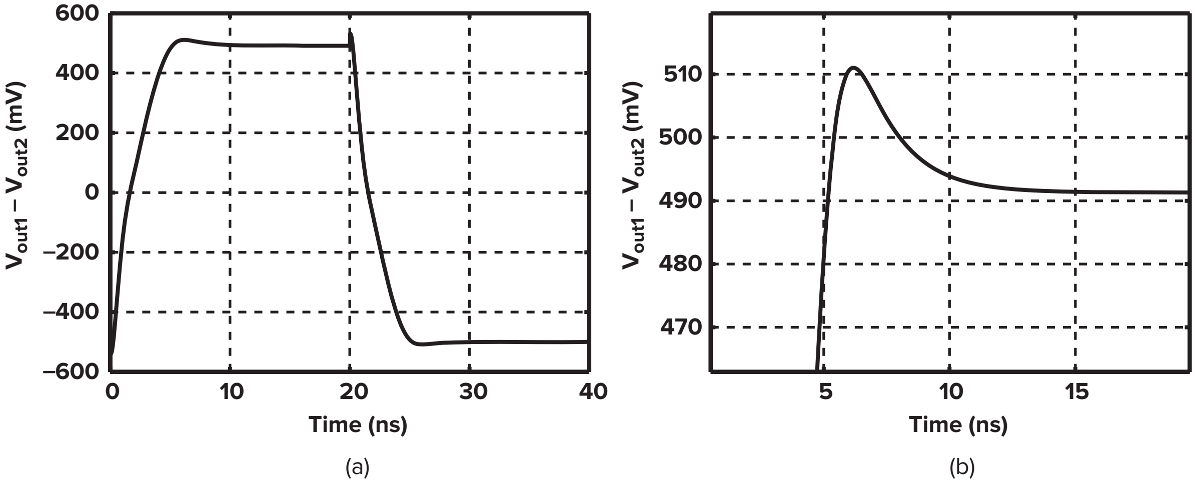


Figure 11.64 (a) Large-signal response of closed-loop amplifier, and (b) close-up of (a) showing settling to 1% accuracy.

▶ Example 11.16

Estimate the above pole frequency.

Solution

With $C_{ox} \approx 15\ \text{fF}/\mu\text{m}^2$, the gate-source capacitance of the cascode NMOS transistors amounts to $(2/3)(45\ \mu\text{m} \times 0.32\ \mu\text{m}) \times 15\ \text{fF}/\mu\text{m}^2 \approx 144$ fF. (Our calculation is sloppy because the effective length is less than 0.32 μm and the overlap capacitance is neglected.) To this value we must add the S/D junction capacitances and the gate-drain capacitance of the input transistors, obtaining roughly 200 fF. To estimate the transconductance of the cascode devices, we assume that they operate in weak inversion and write $g_m \approx I_D/(\xi V_T) \approx 56.5\ \mu\text{A}/(1.5 \times 26\ \text{mV}) = 1/(690\ \Omega)$. The pole frequency is thus around 1.15 GHz, contributing substantial phase shift at the open-loop unity-gain frequency.

◀

To address the settling issue, we consider cascode compensation (Chapter 10). In fact, we can combine both methods and, with some iteration, reach the design shown in Fig. 11.65(a). Depicted in Fig. 11.65(b), takes less than 5 ns.[14] This performance is achieved with a power consumption of 370 μW.

[14] In this case, the 1% margin to the final value is equal to 490 mV ± 1 V/100 because the total swing is from −0.5 V to +0.5 V.

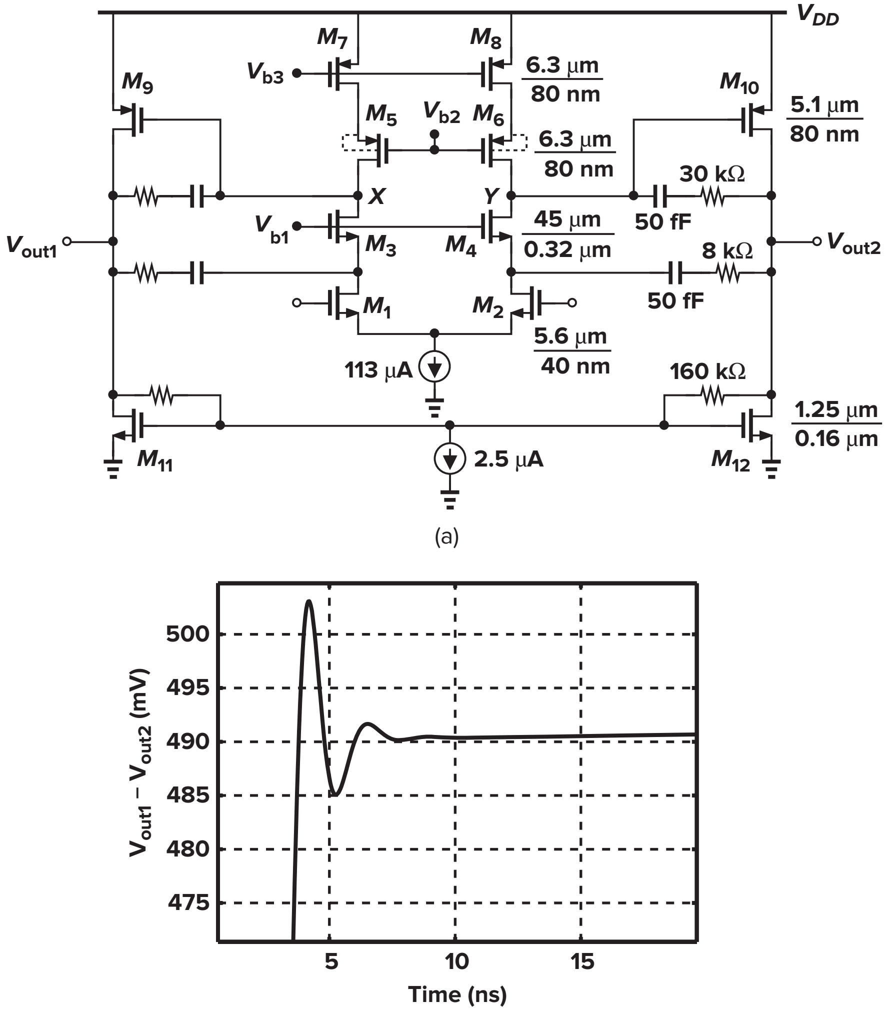


Figure 11.65 (a) Final op amp design, and (b) its closed-loop large-signal step response.

11.7 ■ Summary

This chapter has portrayed to the reader how the analog designer's mind works. We have seen that the design proceeds methodically, assuming an almost arbitrary power budget, and first strives to meet the voltage swing and gain requirements (the toughest issues today). With a reasonable design in hand, we then aggressively reduce the power by linear scaling, paying attention to parameters that cannot be scaled (e.g., the load capacitance) and bearing in mind that speed, noise, and offset degrade. Our efforts exemplify three steps in good analog design. (1) We closely examine the circuit's behavior and understand the root cause of the undesired phenomena. (2) We adjust only the circuit parameters that relate to the root cause—we do not blindly play with any random device. (3) We continue to explore various circuit techniques and new ideas, sometimes reaching a dead end but many a time improving the performance. The reader can see that we optimize the circuits "by hand" rather than automate the task using tools found in simulators. High-performance analog design requires human intelligence.

Problems

11.1. Consider the characteristics shown in Fig. 11.2. Estimate a λ value for $V_{GS} - V_{TH} = 350$ mV based on the slope from $V_{DS} = 0.2$ V to 1 V. [Hint: express the ratio of two currents at V_{DS1} and V_{DS2} as $(1 + \lambda V_{DS1})/(1 + \lambda V_{DS2})$]. Repeat this calculation for $V_{GS} - V_{TH} = 200$ mV, 250 mV, and 300 mV. What trend do you observe?

11.2. Explain why g_m falls in Fig. 11.6 as $V_{GS} - V_{TH}$ exceeds 0.5 V.

11.3. Suppose a hypothetical transistor exhibits a transconductance given by $g_m = \beta(V_{GS} - V_{TH})^2$.
(a) Find an expression for I_D as a function of $V_{GS} - V_{TH}$.
(b) Find two other expressions for g_m.

11.4. Sketch the plots in Fig. 11.7 for the device introduced in the previous problem.

11.5. We wish to bias a transistor with $L = 40$ nm at $I_D = 0.25$ mA. Referring to Fig. 11.13, determine which case yields a higher output impedance: $W = 5\ \mu$m or $W = 10\ \mu$m.

11.6. Explain what happens to the unachievable region in Fig. 11.15 if ξ falls from 1.5 to 1.0. Assume that the behavior in strong inversion does not change.

11.7. Modeling the thermal noise of M_6 in Fig. 11.21 by a voltage source in series with its gate, determine the gain that it sees as it reaches node Y. Use the exact expression for the gain of a degenerated CS stage. Compare this result with the thermal noise contributed by M_8.

11.8. Consider the arrangement shown in Fig. 11.24(b). How high can the input CM level be for M_{13} and M_{14} to remain in saturation? Does I_{D11}/I_{D12} increase or deacrease beyond this point?

11.9. Suppose a closed-loop amplifier exhibits ringing at a frequency f_1 in its step response [as in Fig. 11.46(b)]. Does this provide any information about the phase response of the open-loop circuit?

11.10. A two-stage op amp contains a nondominant pole, ω_{p2}, at the output and is compensated for PM = 45° so that $|\beta H|$ drops to unity at ω_{p2}. Assume that the dominant pole is much lower than ω_{p2}.
(a) Estimate the degradation in PM if the load capacitance seen at the output is doubled.
(b) How should the compensation be modified to ensure that PM = 45° again?

11.11. Estimate the closed-loop time constant in Fig. 11.57 and see if it agrees with the open-loop dominant frequency of 1.7 MHz.

11.12. Suppose that in Fig. 11.60, we scale an op amp *up* by a factor of α. If the load capacitance is constant, how much bandwidth improvement can be achieved?

11.13. Modeling the op amp in Fig. 11.51(a) by a voltage-dependent current source equal to $G_m V_{XY}$ and an output resistance equal to R_{out}, calculate the zero of the closed-loop transfer function. (Hint: the output voltage is equal to zero at the zero frequency.)

11.14. Consider the situation illustrated in Fig. 11.47(a). Suppose we place a capacitor, C_F, in parallel with the feedback resistor. Prove that C_F introduces a zero in the loop transmission and determine its value so as to cancel the pole created by C_{in}.

CHAPTER

12

Bandgap References

Analog circuits incorporate voltage and current references extensively. Such references are dc quantities that exhibit little dependence on supply and process parameters and a *well-defined* dependence on the temperature. For example, the bias current of a differential pair must be generated according to a reference, for it affects the voltage gain and noise of the circuit. We have also seen the need for precise voltages to define common-mode levels in op amps. Moreover, in systems such as A/D and D/A converters, a reference is required to define the input or output full-scale range.

In this chapter, we deal with the design of reference generators in CMOS technology, focusing on well-established "bandgap" techniques. First, we study supply-independent biasing and the problem of start-up. Next, we describe temperature-independent references and examine issues such as the effect of offset voltages. Finally, we present constant-G_m biasing and study an example of state-of-the-art bandgap references.

12.1 ■ General Considerations

As mentioned above, the objective of reference generation is to establish a dc voltage or current that is independent of the supply and process and has a well-defined behavior with temperature. In most applications, the required temperature dependence assumes one of three forms: (1) proportional to absolute temperature (PTAT); (2) constant-G_m behavior, i.e., such that the transconductance of certain transistors remains constant; (3) temperature independent. We can therefore divide the task into two design problems: supply-independent biasing and definition of the temperature variation.

In addition to supply, process, and temperature variability, several other parameters of reference generators may be critical as well. These include output impedance, output noise, and power dissipation. We return to these issues later in this chapter.

12.2 ■ Supply-Independent Biasing

Our use of bias currents and current mirrors in previous chapters has implicitly assumed that a "golden" reference current is available. As shown in Fig. 12.1(a), if I_{REF} does not vary with V_{DD}, and channel-length modulation of M_2 and M_3 is neglected, then I_{D2} and I_{D3} remain independent of the supply voltage. The question then is—How do we generate I_{REF}?

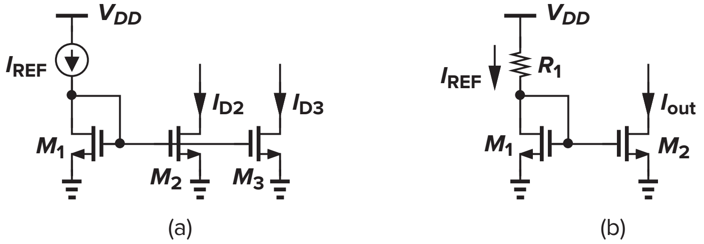


Figure 12.1 Current mirror biasing using (a) an ideal current source and (b) a resistor.

As an approximation of a current source, we tie a resistor from V_{DD} to the gate of M_1 [Fig. 12.1(b)]. However, the output current of this circuit is quite sensitive to V_{DD}:

$$\Delta I_{out} = \frac{\Delta V_{DD}}{R_1 + 1/g_{m1}} \cdot \frac{(W/L)_2}{(W/L)_1} \tag{12.1}$$

In order to arrive at a less sensitive solution, we postulate that the circuit must bias *itself*, i.e., I_{REF} must be somehow derived from I_{out}. The idea is that if I_{out} is to be ultimately independent of V_{DD}, then I_{REF} can be a replica of I_{out}. Figure 12.2 illustrates an implementation where M_3 and M_4 copy I_{out}, thereby defining I_{REF}. In essence, I_{REF} is "bootstrapped" to I_{out}. With the sizes chosen here, we have $I_{out} = KI_{REF}$ if channel-length modulation is neglected. Note that, since each diode-connected device feeds from a current source, I_{out} and I_{REF} are relatively independent of V_{DD}.

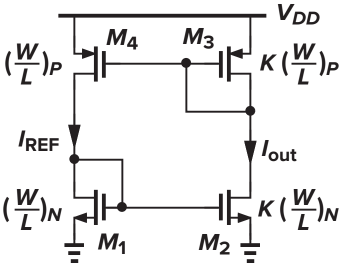


Figure 12.2 Simple circuit to establish supply-independent currents.

Since I_{out} and I_{REF} in Fig. 12.2 display little dependence on V_{DD}, their magnitude is set by other parameters. How do we calculate these currents? Interestingly, if M_1–M_4 operate in saturation and $\lambda \approx 0$, then the circuit is governed by only one equation, $I_{out} = KI_{REF}$, and hence can support *any* current level! For example, if we initially force I_{REF} to be 10 μA, the resulting I_{out} of $K \times 10\ \mu$A "circulates" around the loop, sustaining these current levels in the left and right branches indefinitely.

To uniquely define the currents, we add another constraint to the circuit, e.g., as shown in Fig. 12.3(a). Here, resistor R_S decreases the current of M_2 while the PMOS devices require that $I_{out} = I_{REF}$ because they have identical dimensions and thresholds. We can write $V_{GS1} = V_{GS2} + I_{D2}R_S$, or

$$\sqrt{\frac{2I_{out}}{\mu_n C_{ox}(W/L)_N}} + V_{TH1} = \sqrt{\frac{2I_{out}}{\mu_n C_{ox}K(W/L)_N}} + V_{TH2} + I_{out}R_S \tag{12.2}$$

Neglecting body effect, we have

$$\sqrt{\frac{2I_{out}}{\mu_n C_{ox}(W/L)_N}}\left(1 - \frac{1}{\sqrt{K}}\right) = I_{out}R_S \tag{12.3}$$

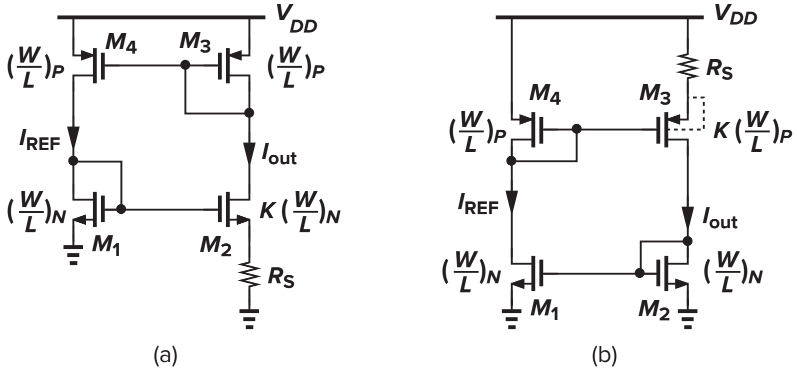


Figure 12.3 (a) Addition of R_S to define the currents; (b) alternative implementation eliminating body effect.

and hence

$$I_{out} = \frac{2}{\mu_n C_{ox}(W/L)_N} \cdot \frac{1}{R_S^2}\left(1 - \frac{1}{\sqrt{K}}\right)^2 \tag{12.4}$$

As expected, the current is independent of the supply voltage (but still a function of process and temperature).

The assumption $V_{TH1} = V_{TH2}$ introduces some error in the foregoing calculations because the sources of M_1 and M_2 are at different voltages. Shown in Fig. 12.3(b) is to place the resistor in the source of M_3 while eliminating body effect by tying the source and bulk of each PMOS transistor. Another solution is described in Problem 12.1.

The circuits of Figs. 12.3(a) and (b) exhibit little supply dependence if channel-length modulation is negligible. For this reason, relatively long channels are used for all of the transistors in the circuit. This also helps reduce their flicker noise.

▶ Example 12.1

Assuming $\lambda \neq 0$ in Fig. 12.3(a), estimate the change in I_{out} for a small change ΔV_{DD} in the supply voltage.

Solution

Simplifying the circuit as depicted in Fig. 12.4, where $R_1 = r_{O1}\|(1/g_{m1})$ and $R_3 = r_{O3}\|(1/g_{m3})$, we calculate the "gain" from V_{DD} to I_{out}. The small-signal gate-source voltage of M_4 equals $-I_{out}R_3$, and the current through r_{O4} is $(V_{DD} - V_X)/r_{O4}$. Thus,

$$\frac{V_{DD} - V_X}{r_{O4}} + I_{out}R_3 g_{m4} = \frac{V_X}{R_1} \tag{12.5}$$

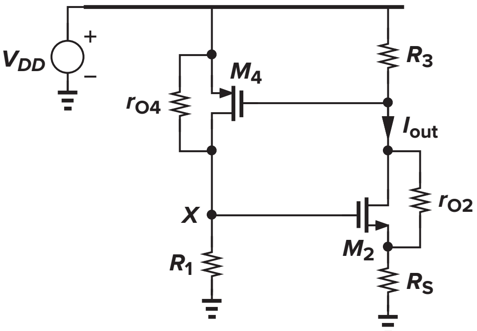


Figure 12.4

If we denote the equivalent transconductance of M_2 and R_S by $G_{m2} = I_{out}/V_X$, then

$$\frac{I_{out}}{V_{DD}} = \frac{1}{r_{O4}}\left[\frac{1}{G_{m2}(r_{O4}\|R_1)} - g_{m4}R_3\right]^{-1} \tag{12.6}$$

Note from Chapter 3 that

$$G_{m2} = \frac{g_{m2}r_{O2}}{R_S + r_{O2} + (g_{m2}+g_{mb2})R_S r_{O2}} \tag{12.7}$$

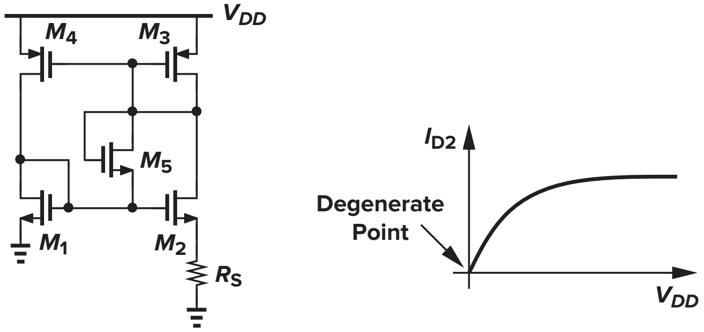
Interestingly, the sensitivity vanishes if $r_{O4} = \infty$.

In some applications, the sensitivity given by (12.6) is prohibitively large. Also, owing to various capacitive paths, the supply sensitivity of the circuit rises at high frequencies. For these reasons, the supply voltage of the core is often derived from a locally-generated, less sensitive voltage. We return to this point in Sec. 12.8.

An important issue in supply-independent biasing is the existence of "degenerate" bias points. In the circuit of Fig. 12.3(a), for example, if all of the transistors carry zero current when the supply is turned on, they may remain off indefinitely because the loop can support a zero current in both branches. This condition is not predicted by (12.4) because in manipulating (12.3), we divided both sides by $\sqrt{I_{out}}$, tacitly assuming that $I_{out} \neq 0$. In other words, the circuit can settle in one of *two* different operating conditions.

Called the "start-up" problem, the above issue is resolved by adding a mechanism that drives the circuit out of the degenerate bias point when the supply is turned on. Shown in Fig. 12.5(a) is a simple example, where the diode-connected device M_5 provides a current path from V_{DD} through M_3 and M_1 to ground upon start-up. Thus, M_3 and M_1, and hence M_2 and M_4, cannot remain off. Of course, this technique is practical only if $V_{TH1} + V_{TH5} + |V_{TH3}| < V_{DD}$ and $V_{GS1} + V_{TH5} + |V_{GS3}| > V_{DD}$, the latter to ensure that M_5 remains off after start-up. Another start-up circuit is analyzed in Problem 12.2.

The problem of start-up generally requires careful analysis and simulation. The supply voltage must be ramped from zero in a dc sweep simulation (such that parasitic capacitances do not cause false start-up) as well as in a transient simulation and the behavior of the circuit examined for each supply voltage. Figure 12.5(b) depicts an example of the observed behavior in the presence of the start-up circuit. In complex implementations, more than one degenerate point may exist.



(a) (b)

Figure 12.5 (a) Addition of start-up device to the circuit of Fig. 12.3(a), and (b) illustration of degenerate point.

12.3 ■ Temperature-Independent References

Reference voltages or currents that exhibit little dependence on temperature prove essential in many analog circuits. It is interesting to note that, since most process parameters vary with temperature, if a reference is temperature-independent, then it is usually process-independent as well.

How do we generate a quantity that remains constant with temperature? We postulate that if two quantities having opposite temperature coefficients (TCs) are added with proper weighting, the result displays a zero TC. For example, for two voltages V_1 and V_2 that vary in opposite directions with temperature, we choose α_1 and α_2 such that $\alpha_1 \partial V_1/\partial T + \alpha_2 \partial V_2/\partial T = 0$, obtaining a reference voltage, $V_{REF} = \alpha_1 V_1 + \alpha_2 V_2$, with zero TC.

We must now identify two voltages that have positive and negative TCs. Among various device parameters in semiconductor technologies, the characteristics of bipolar transistors have proven the most reproducible and well-defined quantities that can provide positive and negative TCs. Even though many parameters of MOS devices have been considered for the task of reference generation [1, 2], bipolar operation still forms the core of such circuits.

12.3.1 Negative-TC Voltage

The base-emitter voltage of bipolar transistors or, more generally, the forward voltage of a *pn*-junction diode exhibits a negative TC. We first obtain the expression for the TC in terms of readily-available quantities.

For a bipolar device, we can write $I_C = I_S \exp(V_{BE}/V_T)$, where $V_T = kT/q$. The saturation current I_S is proportional to $\mu k T n_i^2$, where μ denotes the mobility of minority carriers and n_i is the intrinsic carrier concentration of silicon. The temperature dependence of these quantities is represented as $\mu \propto \mu_0 T^m$, where $m \approx -3/2$, and $n_i^2 \propto T^3 \exp[-E_g/(kT)]$, where $E_g \approx 1.12$ eV is the bandgap energy of silicon. Thus,

$$I_S = bT^{4+m} \exp \frac{-E_g}{kT} \tag{12.8}$$

where b is a proportionality factor. Writing $V_{BE} = V_T \ln(I_C/I_S)$, we can now compute the TC of the base-emitter voltage. In taking the derivative of V_{BE} with respect to T, we must know the behavior of I_C as a function of the temperature. To simplify the analysis, we assume for now that I_C is held *constant*. Thus,

$$\frac{\partial V_{BE}}{\partial T} = \frac{\partial V_T}{\partial T} \ln \frac{I_C}{I_S} - \frac{V_T}{I_S} \frac{\partial I_S}{\partial T} \tag{12.9}$$

From (12.8), we have

$$\frac{\partial I_S}{\partial T} = b(4+m)T^{3+m} \exp \frac{-E_g}{kT} + bT^{4+m} \left(\exp \frac{-E_g}{kT} \right) \left(\frac{E_g}{kT^2} \right) \tag{12.10}$$

Therefore,

$$\frac{V_T}{I_S} \frac{\partial I_S}{\partial T} = (4+m)\frac{V_T}{T} + \frac{E_g}{kT^2} V_T \tag{12.11}$$

With the aid of (12.9) and (12.11), we can write

$$\frac{\partial V_{BE}}{\partial T} = \frac{V_T}{T} \ln \frac{I_C}{I_S} - (4+m)\frac{V_T}{T} - \frac{E_g}{kT^2} V_T \tag{12.12}$$

$$= \frac{V_{BE} - (4+m)V_T - E_g/q}{T} \tag{12.13}$$

Equation (12.13) gives the temperature coefficient of the base-emitter voltage at a given temperature T, revealing dependence on the magnitude of V_{BE} itself. With $V_{BE} \approx 750$ mV and $T = 300$ K, we have $\partial V_{BE}/\partial T \approx -1.5$ mV/K.

In old bipolar technologies, where I_C/I_S was relatively small (because the transistors were large), $V_{BE} \approx 700$ mV and $\partial V_{BE}/\partial T \approx -1.9$ mV/K at room temperature. Modern bipolar transistors typically operate at much higher current densities, exhibiting $V_{BE} \approx 800$ mV and hence $\partial V_{BE}/\partial T \approx -1.5$ mV/K at $T = 300$ K.

From (12.13), we note that the temperature coefficient of V_{BE} itself depends on the temperature, creating error in constant reference generation if the positive-TC quantity exhibits a *constant* temperature coefficient.

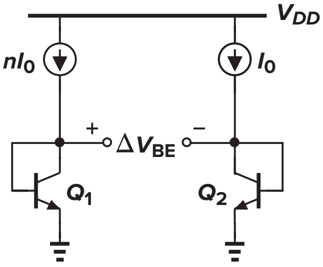


Figure 12.6 Generation of PTAT voltage.

12.3.2 Positive-TC Voltage

It was recognized in 1964 [3] that if two bipolar transistors operate at unequal current densities,[1] then the *difference* between their base-emitter voltages is directly proportional to the absolute temperature. For example, as shown in Fig. 12.6, if two identical transistors ($I_{S1} = I_{S2}$) are biased at collector currents of nI_0 and I_0 and their base currents are negligible, then

$$\Delta V_{BE} = V_{BE1} - V_{BE2} \tag{12.14}$$

$$= V_T \ln \frac{nI_0}{I_{S1}} - V_T \ln \frac{I_0}{I_{S2}} \tag{12.15}$$

$$= V_T \ln n \tag{12.16}$$

Thus, the V_{BE} difference exhibits a positive temperature coefficient:

$$\frac{\partial \Delta V_{BE}}{\partial T} = \frac{k}{q} \ln n \tag{12.17}$$

Interestingly, this TC is independent of the temperature or behavior of the collector currents.[2]

▶ Example 12.2

How must n be chosen to yield a TC of $+1.5$ mV/K so as to cancel the TC of the base-emitter voltage at $T = 300$ K?

Solution

We choose n so that $(k/q)\ln n = 1.5$ mV/K. Since $k/q = V_T/T = 0.087$ mV/K, we have $\ln n \approx 17.2$ and hence $n = 2.95 \times 10^7$!! We must therefore modify the circuit to avoid such a large disparity between the two currents. ◀

[1] Current density is defined as the ratio of the collector current, I_C, and the saturation current, I_S.

[2] Nonidealities in the characteristics of bipolar transistors introduce a small temperature dependence in this TC.

▶ **Example 12.3**

Calculate ΔV_{BE} in the circuit of Fig. 12.7, where Q_2 is formed as the parallel combination of m units, each identical to Q_1.

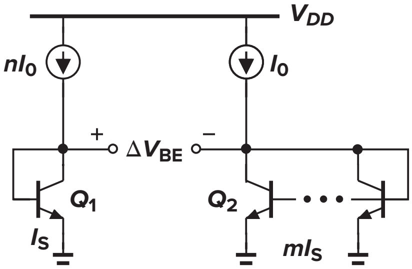


Figure 12.7

Solution

Neglecting base currents, we can write

$$\Delta V_{BE} = V_T \ln \frac{nI_0}{I_S} - V_T \ln \frac{I_0}{mI_S} \tag{12.18}$$

$$= V_T \ln(nm) \tag{12.19}$$

The temperature coefficient is therefore equal to $(k/q)\ln(nm)$. In this circuit, the two transistors' current densities differ by a factor of nm.

◀

12.3.3 Bandgap Reference

With the negative- and positive-TC voltages obtained above, we can now develop a reference that has a nominally zero temperature coefficient. We write $V_{REF} = \alpha_1 V_{BE} + \alpha_2(V_T \ln n)$, where $V_T \ln n$ is the difference between the base-emitter voltages of the two bipolar transistors operating at different current densities. How do we choose α_1 and α_2? Since at room temperature, $\partial V_{BE}/\partial T \approx -1.5$ mV/K whereas $\partial V_T/\partial T \approx +0.087$ mV/K, we may set $\alpha_1 = 1$ and choose $\alpha_2 \ln n$ such that $(\alpha_2 \ln n)(0.087 \text{ mV/K}) = 1.5$ mV/K. That is, $\alpha_2 \ln n \approx 17.2$, indicating that for zero TC

$$V_{REF} \approx V_{BE} + 17.2V_T \tag{12.20}$$

$$\approx 1.25 \text{ V} \tag{12.21}$$

Let us now devise a circuit that adds V_{BE} to $17.2V_T$. First, consider the circuit shown in Fig. 12.8, where the base currents are assumed to be negligible, transistor Q_2 consists of n unit transistors in parallel, and Q_1 is a unit transistor. Suppose we somehow force V_{O1} and V_{O2} to be equal. Then, $V_{BE1} = RI + V_{BE2}$ and $RI = V_{BE1} - V_{BE2} = V_T \ln n$. Thus, $V_{O2} = V_{BE2} + V_T \ln n$, suggesting that V_{O2} can serve as a temperature-independent reference if $\ln n \approx 17.2$ (while V_{O1} and V_{O2} remain equal).

The circuit of Fig. 12.8 requires three modifications to become practical. First, a mechanism must be added to guarantee that $V_{O1} = V_{O2}$. Second, since $\ln n = 17.2$ translates to a prohibitively large n, the term $RI = V_T \ln n$ must be scaled up by a reasonable factor. Third, V_{O2}, which is somehow forced to be equal to V_{O1}, *cannot* become temperature-independent because $V_{O2} \approx V_{BE1} \approx 800$ mV whereas, for temperature independence, we must have $V_{O2} = V_{BE2} + 17.2V_T \approx 1.25$ V. Shown in Fig. 12.9 is an implementation accomplishing all tasks [4]. Here, amplifier A_1 senses V_X and V_Y, driving the top terminals of R_1 and R_2 ($R_1 = R_2$) such that X and Y settle to approximately equal voltages. The

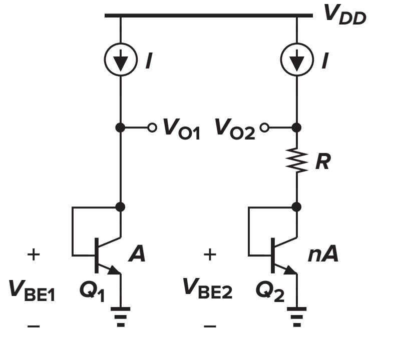


Figure 12.8 Conceptual generation of temperature-independent voltage.

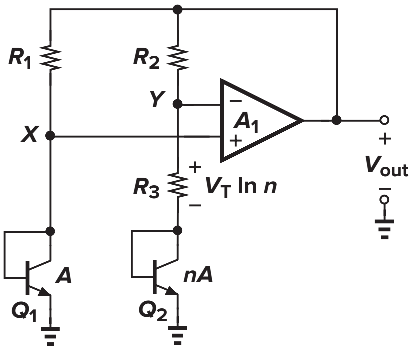


Figure 12.9 Actual implementation of the concept shown in Fig. 12.8.

reference voltage is obtained at the output of the amplifier (rather than at node Y). Following the analysis of Fig. 12.8, we have $V_{BE1} - V_{BE2} = V_T \ln n$, arriving at a current equal to $V_T \ln n / R_3$ through the right branch and hence an output voltage of

$$V_{out} = V_{BE2} + \frac{V_T \ln n}{R_3}(R_3 + R_2) \tag{12.22}$$

$$= V_{BE2} + (V_T \ln n)\left(1 + \frac{R_2}{R_3}\right) \tag{12.23}$$

For a zero TC, we must have $(1 + R_2/R_3)\ln n \approx 17.2$. For example, we may choose $n = 31$ and $R_2/R_3 = 4$. Note that these results do not depend on the TC of the resistors.

It is interesting to understand how the third issue mentioned above is resolved in the topology of Fig. 12.9: we do not attempt to make V_Y ($\approx V_{BE1}$) temperature-independent; rather, we amplify the PTAT voltage drop across R_3 by a factor of $1 + R_2/R_3$ and then add the result to V_{BE2}.

▶ Example 12.4

In Fig. 12.9, R_1 and R_2 are equal and sustain equal voltages, each carrying a current of $(V_T \ln n)/R_3$. We therefore have

$$V_{out} = V_{BE1} + (V_T \ln n)\frac{R_1}{R_3} \tag{12.24}$$

But the second term is *not* equal to $17.2V_T$ if we have already chosen $(V_T \ln n)(1 + R_2/R_3) = 17.2V_T$. Explain this discrepancy.

Solution

The first terms in (12.23) and (12.24) are different. We substitute $V_{BE1} = V_{BE2} + V_T \ln n$ in Eq. (12.13):

$$\frac{\partial V_{BE1}}{\partial T} = \frac{V_{BE2} + V_T \ln n - (4+m)V_T - E_g/q}{T} \tag{12.25}$$

$$= \frac{\partial V_{BE2}}{\partial T} + \frac{k}{q}\ln n \tag{12.26}$$

Thus,

$$\frac{\partial V_{out}}{\partial T} = \frac{\partial V_{BE1}}{\partial T} + \left(\frac{k}{q}\ln n\right)\frac{R_1}{R_3} \tag{12.27}$$

$$= \frac{\partial V_{BE2}}{\partial T} + \left(\frac{k}{q}\ln n\right)\left(1 + \frac{R_1}{R_3}\right) \tag{12.28}$$

which is consistent with (12.23).

◀

The circuit of Fig. 12.9 entails a number of design issues. We consider each one below.

Collector Current Variation The circuit of Fig. 12.9 violates one of our earlier assumptions: the collector currents of Q_1 and Q_2, given by $(V_T \ln n)/R_3$, are proportional to T, whereas $\partial V_{BE}/\partial T \approx -1.5$ mV/K was derived for a *constant* current. What happens to the temperature coefficient of V_{BE} if the collector currents are PTAT? As a first-order iterative solution, let us assume that $I_{C1} = I_{C2} \approx (V_T \ln n)/R_3$. Returning to Eq. (12.9) and including $\partial I_C/\partial T$, we have

$$\frac{\partial V_{BE}}{\partial T} = \frac{\partial V_T}{\partial T}\ln\frac{I_C}{I_S} + V_T\left(\frac{1}{I_C}\frac{\partial I_C}{\partial T} - \frac{1}{I_S}\frac{\partial I_S}{\partial T}\right) \tag{12.29}$$

Since $\partial I_C/\partial T \approx (V_T \ln n)/(R_3 T) = I_C/T$, we can write

$$\frac{\partial V_{BE}}{\partial T} = \frac{\partial V_T}{\partial T}\ln\frac{I_C}{I_S} + \frac{V_T}{T} - \frac{V_T}{I_S}\frac{\partial I_S}{\partial T} \tag{12.30}$$

Equation (12.13) is therefore modified as

$$\frac{\partial V_{BE}}{\partial T} = \frac{V_{BE} - (3+m)V_T - E_g/q}{T} \tag{12.31}$$

indicating that the TC is slightly less negative than -1.5 mV/K. In practice, accurate simulations are necessary to predict the temperature coefficient.

Compatibility with CMOS Technology Our derivation of a temperature-independent voltage relies on the exponential characteristics of bipolar devices for both negative- and positive-TC quantities. We must therefore seek structures in a standard CMOS technology that exhibit such characteristics.

In n-well processes, a *pnp* transistor can be formed as depicted in Fig. 12.10. A p^+ region (the same as the S/D region of PFETs) inside an n-well serves as the emitter and the n-well itself as the base. The p-type substrate acts as the collector and it is inevitably connected to the most negative supply (usually ground). The circuit of Fig. 12.9 can therefore be redrawn as shown in Fig. 12.11.

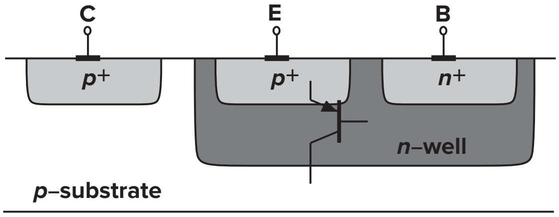


Figure 12.10 Realization of a *pnp* bipolar transistor in CMOS technology.

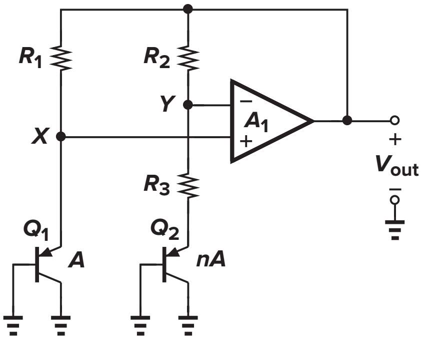


Figure 12.11 Circuit of Fig. 12.9 implemented with *pnp* transistors.

Op Amp Offset and Output Impedance As explained in Chapter 14, owing to asymmetries, op amps suffer from input "offsets," i.e., the output voltage of the op amp is not zero if the input is set to zero. The input offset voltage of the op amp in Fig. 12.9 introduces error in the output voltage. Included in Fig. 12.12, the effect is quantified as $V_{BE1} - V_{OS} \approx V_{BE2} + R_3 I_{C2}$ (if A_1 is large) and $V_{out} = V_{BE2} + (R_3 + R_2)I_{C2}$. Thus,

$$V_{out} = V_{BE2} + (R_3 + R_2)\frac{V_{BE1} - V_{BE2} - V_{OS}}{R_3} \quad (12.32)$$

$$= V_{BE2} + \left(1 + \frac{R_2}{R_3}\right)(V_T \ln n - V_{OS}) \quad (12.33)$$

where we have assumed that $I_{C2} \approx I_{C1}$ despite the offset voltage. The key point here is that V_{OS} is amplified by $1 + R_2/R_3$, introducing error in V_{out}. More important, as explained in Chapter 14, V_{OS} itself varies with temperature, raising the temperature coefficient of the output voltage.

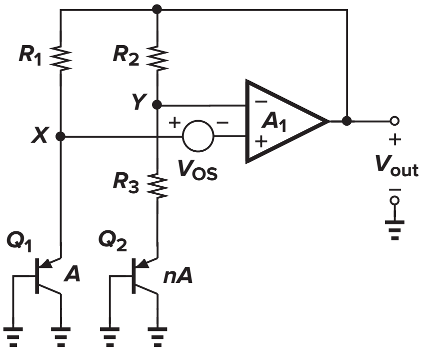


Figure 12.12 Effect of op amp offset on the reference voltage.

▶ Example 12.5

Assuming an ideal op amp, determine the small-signal gain from V_{OS} to V_{out} in Fig. 12.12.

Solution

In the absence of the op amp offset, the two diode-connected bipolar transistors carry equal bias currents, exhibiting a transconductance of g_m. Replacing Q_1 and Q_2 with a small-signal resistance equal to $1/g_m$ and noting that $V_X - V_{OS} \approx V_Y$, we write the following small-signal equation:

$$\frac{1/g_m}{1/g_m + R_1} V_{out} - V_{OS} = \frac{1/g_m + R_3}{1/g_m + R_3 + R_2} V_{out} \tag{12.34}$$

Since $R_1 = R_2$,

$$\frac{V_{out}}{V_{OS}} = -\left[1 + \frac{1}{g_m R_2} + \frac{(1/g_m + R_2)^2}{R_2 R_3}\right] \tag{12.35}$$

If $g_m R_2 \gg 1$, then $V_{out}/V_{OS} \approx -(1 + R_2/R_3)$, agreeing with the results obtained previously. (After all, if $1/g_m \approx 0$, V_{OS} simply sees a noninverting amplifier with a gain of $1 + R_2/R_3$.)

Why does (12.35) not completely agree with the $-V_{OS}(1 + R_2/R_3)$ component in (12.33)? Recall that (12.33) was derived with the assumption that $I_{C1} \approx I_{C2}$ despite the offset voltage. Since $V_X - V_{OS} = V_Y$, we have $I_{C1}R_1 - V_{OS} = I_{C2}R_2$, and hence $I_{C1} = I_{C2} + V_{OS}/R_2$. Let us return to (12.32) and write

$$V_{BE1} - V_{BE2} - V_{OS} = V_T \ln \frac{I_{C1}}{I_{S1}} - V_T \ln \frac{I_{C2}}{I_{S2}} - V_{OS} \tag{12.36}$$

$$= V_T \ln n - V_T \ln \frac{I_{C1}}{I_{C2}} - V_{OS} \tag{12.37}$$

$$= V_T \ln n - V_T \ln \left(1 + \frac{V_{OS}}{R_2 I_{C2}}\right) - V_{OS} \tag{12.38}$$

$$\approx V_T \ln n - V_T \frac{V_{OS}}{R_2 I_{C2}} - V_{OS} \tag{12.39}$$

$$\approx V_T \ln n - \left(1 + \frac{1}{g_m R_2}\right) V_{OS} \tag{12.40}$$

The output offset contribution therefore amounts to $-[1 + 1/(g_m R_2)](1 + R_2/R_3)V_{OS}$, which is approximately the same as (12.35). ◀

Several methods are employed to lower the effect of V_{OS}. First, the op amp incorporates large devices in a carefully chosen topology so as to minimize the offset (Chapter 19). Second, as illustrated in Fig. 12.7, the collector currents of Q_1 and Q_2 can be ratioed by a factor of m such that $\Delta V_{BE} = V_T \ln(mn)$. Third, each branch may use two pn junctions in series to double ΔV_{BE}. Figure 12.13 depicts a realization using the last two techniques. Here, R_1 and R_2 are ratioed by a factor of m, producing $I_1 \approx mI_2$. Neglecting base currents and assuming that A_1 is large, we can now write $V_{BE1} + V_{BE2} - V_{OS} = V_{BE3} + V_{BE4} + R_3 I_2$ and $V_{out} = V_{BE3} + V_{BE4} + (R_3 + R_2)I_2$. It follows that

$$V_{out} = V_{BE3} + V_{BE4} + (R_3 + R_2)\frac{2V_T \ln(mn) - V_{OS}}{R_3} \tag{12.41}$$

$$= 2V_{BE} + \left(1 + \frac{R_2}{R_3}\right)[2V_T \ln(mn) - V_{OS}] \tag{12.42}$$

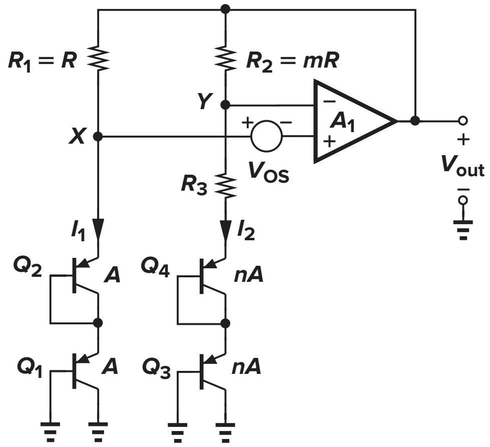


Figure 12.13 Reduction of the effect of op amp offset.

Thus, the effect of the offset voltage is reduced by increasing the first term in the square brackets. The issue, however, is that $V_{out} \approx 2 \times 1.25\text{ V} = 2.5\text{ V}$, a value difficult to generate by the op amp at low supply voltages.

In the circuits studied above, the op amp drives two resistive branches and must therefore provide a low output impedance. Fortunately, it is possible to avoid this issue by a simple modification described below.

The implementation of Fig. 12.13 is not feasible in a standard CMOS technology because the collectors of Q_2 and Q_4 are not grounded. In order to utilize the bipolar structure shown in Fig. 12.10, we modify the series combination of the diodes as illustrated in Fig. 12.14(a), converting one of the diodes to an emitter follower. However, we must ensure that the bias currents of both transistors have the same behavior with temperature. Thus, we bias each transistor by a PMOS current source rather than a resistor [Fig. 12.14(b)]. The overall circuit then assumes the topology shown in Fig. 12.15, where the op amp adjusts the gate voltage of the PMOS devices so as to equalize V_X and V_Y. Interestingly, in this circuit, the op amp experiences no resistive loading, but the mismatch and channel-length modulation of the PMOS devices introduce error at the output (Problem 12.3).

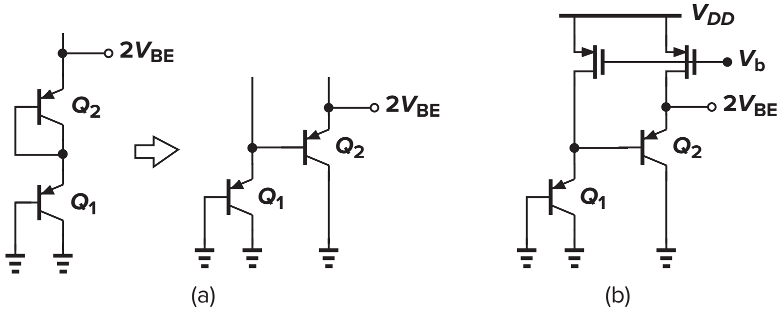


Figure 12.14 (a) Conversion of series diodes to a topology with grounded collectors; (b) circuit of part (a) biased by PMOS current sources.

An important concern in the circuit of Fig. 12.15 is the relatively low current gain of the "native" *pnp* transistors. Since the base currents of Q_2 and Q_4 generate an error in the emitter currents of Q_1 and Q_3, a means of base current cancellation may be necessary (Problem 12.5).

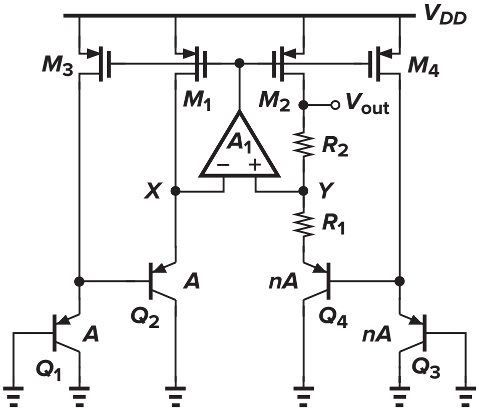


Figure 12.15 Reference generator incorporating two series base-emitter voltages.

Feedback Polarity In the circuit of Fig. 12.9, the feedback signal produced by the op amp returns to both of its inputs. The negative-feedback factor is given by

$$\beta_N = \frac{1/g_{m2} + R_3}{1/g_{m2} + R_3 + R_2} \tag{12.43}$$

and the positive-feedback factor by

$$\beta_P = \frac{1/g_{m1}}{1/g_{m1} + R_1} \tag{12.44}$$

To ensure an overall negative feedback, β_P must be less than β_N, preferably by roughly a factor of two so that the circuit's transient response remains well behaved with large capacitive loads.

Bandgap Reference The voltage generated according to (12.20) is called a "bandgap reference." To understand the origin of this terminology, let us write the output voltage as

$$V_{REF} = V_{BE} + V_T \ln n \tag{12.45}$$

and hence:

$$\frac{\partial V_{REF}}{\partial T} = \frac{\partial V_{BE}}{\partial T} + \frac{V_T}{T} \ln n \tag{12.46}$$

Setting this to zero and substituting for $\partial V_{BE}/\partial T$ from (12.13), we have

$$\frac{V_{BE} - (4+m)V_T - E_g/q}{T} = -\frac{V_T}{T} \ln n \tag{12.47}$$

If $V_T \ln n$ is found from this equation and inserted in (12.45), we obtain

$$V_{REF} = \frac{E_g}{q} + (4+m)V_T \tag{12.48}$$

Thus, the reference voltage exhibiting a nominally-zero TC is given by a few *fundamental* numbers: the bandgap voltage of silicon, E_g/q, the temperature exponent of mobility, m, and the thermal voltage, V_T. The term "bandgap" is used here because as $T \to 0$, $V_{REF} \to E_g/q$.

▶ **Example 12.6**

Prove directly that, as $T \to 0$, $V_{BE} \to E_g/q$, and hence $V_{REF} = V_{BE} + V_T \ln n \to E_g/q$.

Solution

From Eq. (12.8), we have

$$V_{BE} = V_T \ln \frac{I_C}{I_S} \tag{12.49}$$

$$= V_T \left[\ln I_C - \ln b - (4+m)\ln T + \frac{E_g}{kT} \right] \tag{12.50}$$

Thus, $V_{BE} \to E_g/q$ if $T \to 0$ and I_C is constant. ◀

Supply Dependence and Start-Up In the circuit of Fig. 12.9, the output voltage is relatively independent of the supply voltage so long as the open-loop gain of the op amp is sufficiently high. The circuit may require a start-up mechanism because if V_X and V_Y are equal to zero, the input differential pair of the op amp may turn off. Start-up techniques similar to those of Fig. 12.5 can be added to ensure that the op amp turns on when the supply is applied.

The supply rejection of the circuit typically degrades at high frequencies owing to the op amp's rejection properties, often mandating "supply regulation." An example is described in Sec. 12.8.

Curvature Correction If plotted as a function of temperature, bandgap voltages exhibit a finite "curvature," i.e., their TC is typically zero at one temperature and positive or negative at other temperatures (Fig. 12.16). The curvature arises from temperature variation of base-emitter voltages, collector currents, and offset voltages.

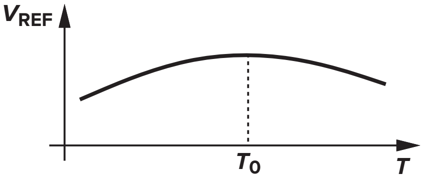


Figure 12.16 Curvature in temperature dependence of a bandgap voltage.

Many curvature correction techniques have been devised to suppress the variation of V_{REF} [5, 6] in bipolar bandgap circuits, but they are seldom used in CMOS counterparts. This is because, due to large offsets and process variations, samples of a bandgap reference display substantially different zero-TC temperatures (Fig. 12.17), making it difficult to correct the curvature reliably.

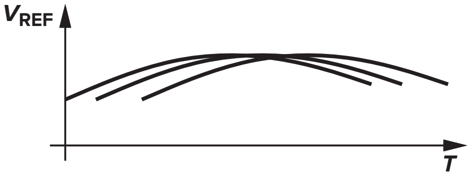


Figure 12.17 Variation of the zero-TC temperature for different samples.

12.4 ■ PTAT Current Generation

In the analysis of bandgap circuits, we noted that the bias currents of the bipolar transistors are in fact proportional to absolute temperature. Useful in many applications, PTAT currents can be generated by a topology such as that shown in Fig. 12.18. Alternatively, we can combine the supply-independent biasing scheme of Fig. 12.2 with a bipolar core, arriving at Fig. 12.19.[3] Assuming for simplicity that M_1-M_2 and M_3-M_4 are identical pairs, we note that for $I_{D1} = I_{D2}$, the circuit must ensure that $V_X = V_Y$. Thus, $I_{D1} = I_{D2} = (V_T \ln n)/R_1$, yielding the same behavior for I_{D5}. In practice, due to mismatches between the transistors and, more important, the temperature coefficient of R_1, the variation of I_{D5} deviates from the ideal equation. For low-voltage operation, the topology of Fig. 12.18 is preferred.

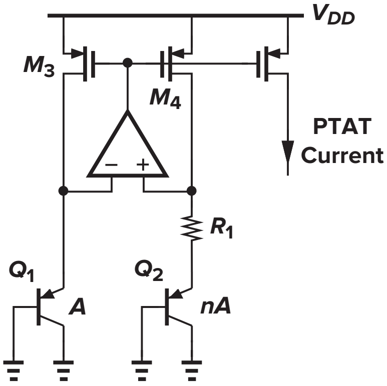


Figure 12.18 Generation of a PTAT current.

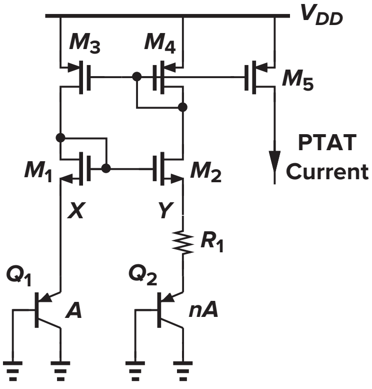


Figure 12.19 Alternative method of generating a PTAT current.

The circuit of Fig. 12.18 can be readily modified to provide a bandgap reference voltage as well. Illustrated in Fig. 12.20, the idea is to add a PTAT voltage $I_{D5}R_2$ to a base-emitter voltage. The output therefore equals

$$V_{REF} = |V_{BE3}| + \frac{R_2}{R_1} V_T \ln n \tag{12.51}$$

[3] The two circuits in Figs. 12.18 and 12.19 exhibit different supply rejections. With a carefully-designed op amp, the former achieves a higher rejection.

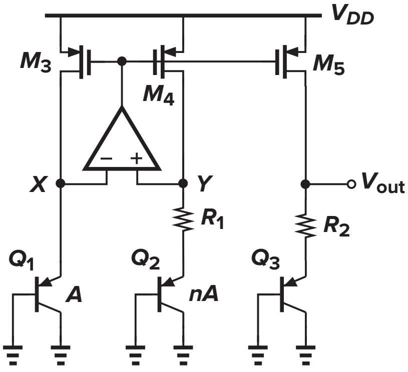


Figure 12.20 Generation of a temperature-independent voltage.

where all of the PMOS transistors are assumed identical. Note that the value of V_{BE3} and hence the size of Q_3 are somewhat arbitrary so long as the sum of the two terms in (12.51) gives a zero TC. In reality, mismatches of the PMOS devices introduce error in V_{out}.

12.5 ■ Constant-G_m Biasing

The transconductance of MOSFETs plays a critical role in analog circuits, determining such performance parameters as noise, small-signal gain, and speed. For this reason, it is often desirable to bias the transistors such that their transconductance does not depend on the temperature, process, or supply voltage.

A simple circuit used to define the transconductance is the supply-independent bias topology of Fig. 12.3. Recall that the bias current is given by

$$I_{out} = \frac{2}{\mu_n C_{ox}(W/L)_N} \frac{1}{R_S^2} \left(1 - \frac{1}{\sqrt{K}}\right)^2 \tag{12.52}$$

Thus, the transconductance of M_1 equals

$$g_{m1} = \sqrt{2\mu_n C_{ox} \left(\frac{W}{L}\right)_N I_{D1}} \tag{12.53}$$

$$= \frac{2}{R_S}\left(1 - \frac{1}{\sqrt{K}}\right) \tag{12.54}$$

a value independent of the supply voltage and MOS device parameters.

In reality, the value of R_S in (12.54) does vary with temperature and process. If the temperature coefficient of the resistor is known, bandgap and PTAT reference generation techniques can be utilized to cancel the temperature dependence. *Process* variations, however, limit the accuracy with which g_{m1} is defined.

In systems where a precise clock frequency is available, the resistor R_S in Fig. 12.3 can be replaced by a switched-capacitor equivalent (Chapter 13) to achieve a somewhat higher accuracy. Depicted in Fig. 12.21, the idea is to establish an average resistance equal to $(C_S f_{CK})^{-1}$ between the source of M_2 and ground, where f_{CK} denotes the clock frequency. Capacitor C_B is added to shunt the high-frequency components resulting from switching to ground. Since the absolute value of capacitors is typically more tightly controlled and since the TC of capacitors is much smaller than that of resistors, this technique provides a higher reproducibility in the bias current and transconductance.

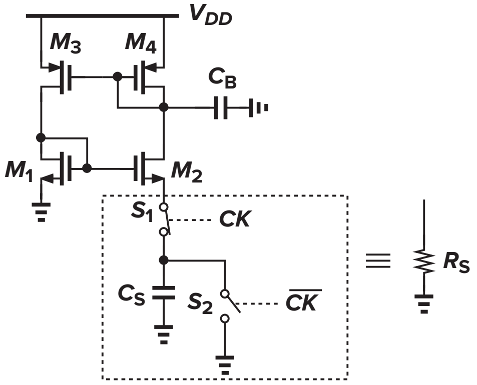


Figure 12.21 Constant-G_m biasing by means of a switched-capacitor "resistor."

The switched-capacitor approach of Fig. 12.21 can be applied to other circuits as well. For example, as shown in Fig. 12.22, a voltage-to-current converter with a relatively high accuracy can be constructed.

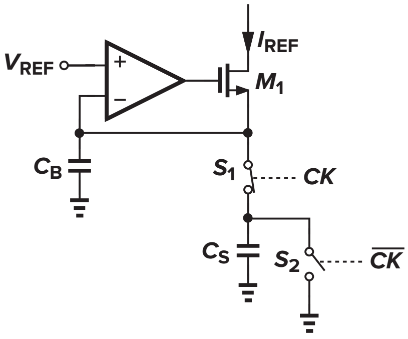


Figure 12.22 Voltage-to-current conversion by means of a switched-capacitor resistor.

12.6 ■ Speed and Noise Issues

Even though reference generators are low-frequency circuits, they may affect the speed of the circuits that they feed. Furthermore, various building blocks may experience "crosstalk" through reference lines. These difficulties arise because of the finite output impedance of reference voltage generators, especially if they incorporate op amps. As an example, let us consider the configuration shown in Fig. 12.23, assuming that the voltage at node N is heavily disturbed by the circuit fed by M_5. For fast changes in V_N, the op amp cannot maintain V_P constant, and the bias currents of M_5 and M_6 experience large transient changes. Also, the duration of the transient at node P may be quite long if the op amp suffers from a slow response. For this reason, many applications may require a high-speed op amp in the reference generator.

In systems where the power consumed by the reference circuit must be small, the use of a high-speed op amp may not be feasible. Alternatively, the critical node, e.g., node P in Fig. 12.23, can be bypassed to ground by means of a large capacitor (C_B) so as to suppress the effect of external disturbances. This approach involves two issues. First, the stability of the op amp must not degrade with the addition of the capacitor, requiring the op amp to be of a one-stage nature (Chapter 10). Second, since C_B generally slows down the transient response of the op amp, its value must be much greater than the capacitance that couples the disturbance to node P. As illustrated in Fig. 12.24, if C_B is not sufficiently large, then V_P

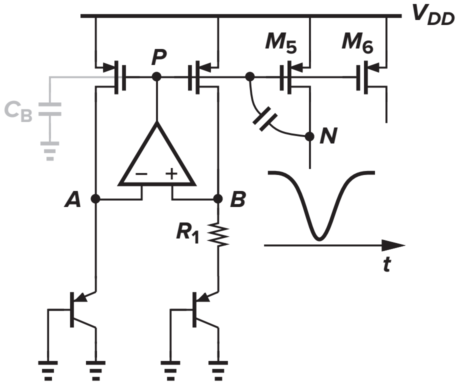


Figure 12.23 Effect of circuit transients on reference voltages and currents.

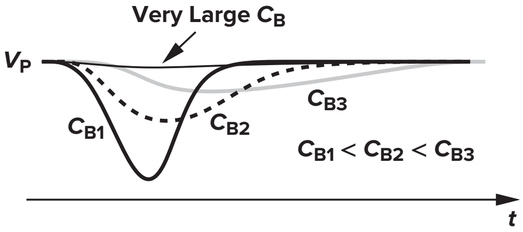


Figure 12.24 Effect of increasing bypass capacitor on the response of a reference generator.

experiences a change and takes a long time to return to its original value, possibly degrading the settling speed of the circuits biased by the reference generator. In other words, depending on the environment, it may be preferable to leave node P agile so that it can quickly recover from transients. In general, as depicted in Fig. 12.25, the response of the circuit must be analyzed by applying a disturbance at the output and observing the settling behavior.

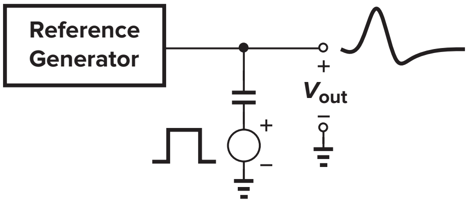


Figure 12.25 Setup for testing the transient response of a reference generator.

▶ Example 12.7

Determine the small-signal output impedance of the bandgap reference shown in Fig. 12.23 and examine its behavior with frequency.

Solution

Figure 12.26 depicts the equivalent circuit, modeling the open-loop op amp by a one-pole transfer function $A(s) = A_0/(1 + s/\omega_0)$ and an output resistance R_{out} and each bipolar transistor by a resistance $1/g_{mN}$. If M_1 and M_2 are identical, each having a transconductance of g_{mP}, then their drain currents are equal to $g_{mP}V_X$, producing a differential voltage at the input of the op amp equal to

$$V_{AB} = -g_{mP}V_X\frac{1}{g_{mN}} + g_{mP}V_X\left(\frac{1}{g_{mN}} + R_1\right) \tag{12.55}$$

$$= g_{mP}V_XR_1 \tag{12.56}$$

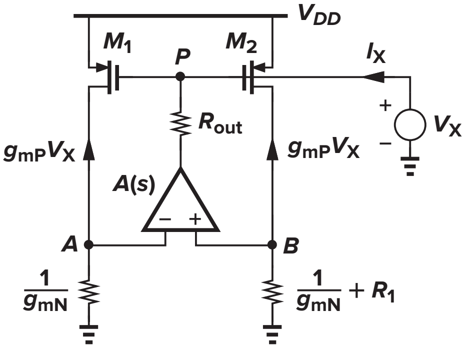


Figure 12.26 Circuit for calculation of the output impedance of a reference generator.

The current flowing through R_{out} is therefore given by

$$I_X = \frac{V_X + g_{mP}V_X R_1 A(s)}{R_{out}} \tag{12.57}$$

yielding

$$\frac{V_X}{I_X} = \frac{R_{out}}{1 + g_{mP}R_1 A(s)} \tag{12.58}$$

$$= \frac{R_{out}}{1 + g_{mP}R_1 \dfrac{A_0}{1 + s/\omega_0}} \tag{12.59}$$

$$= \frac{R_{out}}{1 + g_{mP}R_1 A_0} \frac{1 + \dfrac{s}{\omega_0}}{1 + \dfrac{s}{(1 + g_{mP}R_1 A_0)\omega_0}} \tag{12.60}$$

Thus, the output impedance exhibits a zero at ω_0 and a pole at $(1 + g_{mP}R_1A_0)\omega_0$, with the magnitude behavior plotted in Fig. 12.27. Note that $|Z_{out}|$ is small for $\omega < \omega_0$, but it rises to a high value as the frequency approaches the pole. In fact, setting $\omega = (1 + g_{mP}R_1A_0)\omega_0$ and assuming $g_{mP}R_1A_0 \gg 1$, we have

$$|Z_{out}| = \frac{R_{out}}{1 + g_{mP}R_1 A_0} \left| \frac{1 + j(1 + g_{mP}R_1 A_0)}{1 + j} \right| \tag{12.61}$$

$$= \frac{R_{out}}{\sqrt{2}} \tag{12.62}$$

which is only 30% lower than the open-loop value.

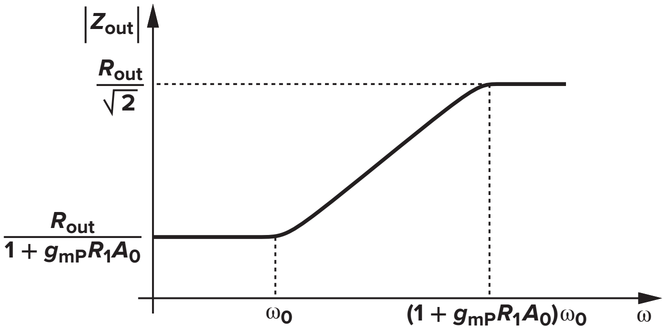


Figure 12.27 Variation of the reference generator output impedance with frequency.

The output noise of reference generators may affect the performance of low-noise circuits considerably. Figure 12.28 illustrates an example: the load current source of a common-source stage is driven by a bandgap circuit with a current multiplication factor of N. Thus, the noise current of M_1 (or M_2) is amplified by the same factor as it appears in M_3. Note that M_1–M_3 carry noise due to the op amp A_1 as well.

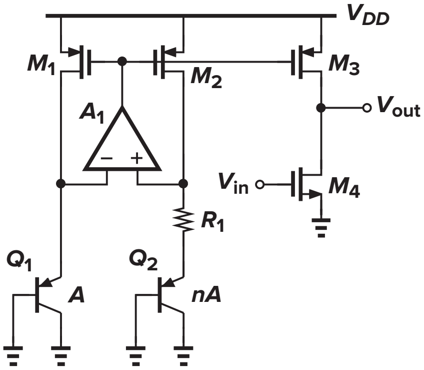


Figure 12.28 Effect of bandgap circuit noise on a CS stage.

As another example, if a high-precision A/D converter employs a bandgap voltage as the reference with which the analog input signal is compared (Fig. 12.29), then the noise in the reference is directly added to the input.

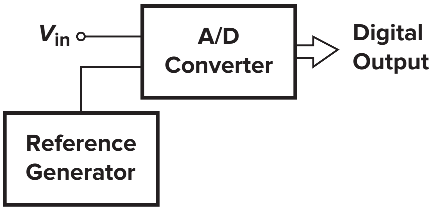


Figure 12.29 A/D converter using a reference generator.

As a simple example, let us calculate the output noise voltage of the circuit shown in Fig. 12.30, taking into account only the input-referred noise voltage of the op amp, $V_{n,op}$. Since the small-signal drain currents of M_1 and M_2 are equal to $V_{n,out}/(R_1 + g_{mN}^{-1})$, we have $V_P = -g_{mP}^{-1} V_{n,out}/(R_1 + g_{mN}^{-1})$, obtaining the differential voltage at the input of the op amp as $-g_{mP}^{-1} A_0^{-1} V_{n,out}/(R_1 + g_{mN}^{-1})$. Beginning

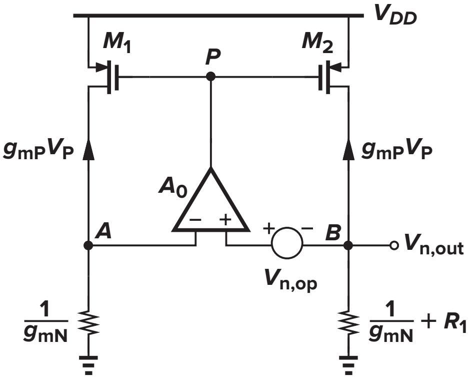


Figure 12.30 Circuit for calculation of noise in a reference generator.

from node A, we can then write

$$\frac{V_{n,out}}{R_1+g_{mN}^{-1}}\cdot\frac{1}{g_{mN}}-\frac{V_{n,out}}{g_{mP}A_0\left(R_1+g_{mN}^{-1}\right)}=V_{n,op}+V_{n,out} \tag{12.63}$$

and hence

$$V_{n,out}\left[\frac{1}{R_1+g_{mN}^{-1}}\left(\frac{1}{g_{mN}}-\frac{1}{g_{mP}A_0}\right)-1\right]=V_{n,op} \tag{12.64}$$

Since typically $g_{mP}A_0 \gg g_{mN} \gg R_1^{-1}$,

$$|V_{n,out}|\approx V_{n,op} \tag{12.65}$$

suggesting that the noise of the op amp directly appears at the output. Note that even the addition of a large capacitor from the output to ground may not suppress low-frequency $1/f$ noise components, a serious difficulty in low-noise applications. The noise contributed by other devices in the circuit is studied in Problem 12.6.

12.7 ■ Low-Voltage Bandgap References

The bandgap voltage expressed by Eq. (12.20) is around 1.25 V, eluding implementation with today's low supplies. The fundamental limitation is that we must add about $17.2V_T$ to one V_{BE} so as to achieve a net zero temperature coefficient.

Is it possible to add two *currents* with positive and negative TCs and then convert the result to an arbitrary voltage that has a zero TC (Fig. 12.31)? Recall from Fig. 12.18 that we can readily generate a PTAT current given by $V_T \ln n/R$. We also envision another current of the form V_{BE}/R serving as that with a negative TC, but how can we generate such a current with minimal complexity?

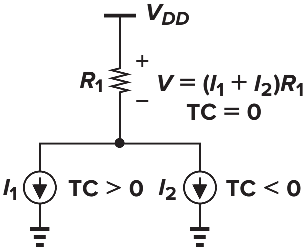


Figure 12.31 Summation of two currents with opposite TCs to obtain a result with zero TC.

Let us return to the circuit of Fig. 12.18, assume that M_3 and M_4 are identical, and note that $|I_{D4}| = V_T \ln n/R_1$ is a PTAT current. We place a resistor in parallel with Q_2 as shown in Fig. 12.32(a). We recognize that R_1 now carries an additional current equal to $|V_{BE2}|/R_2$, i.e., a current with a negative TC. Unfortunately, however, the PTAT behavior is now disturbed because $I_{C1} \neq I_{C2}$. Fortunately, a simple modification resolves this issue: as shown in Fig. 12.32(b), we tie R_2 from Y to ground and place another resistor in parallel with Q_1. Proposed by Banba et al. [8], this topology lends itself to low-voltage implementation, requiring a minimum V_{DD} of $V_{BE1}+|V_{DS3}|$.

To analyze the circuit, we observe that $V_X \approx V_Y \approx |V_{BE1}|$ and $I_{D3} = I_{D4}$. Thus,

$$I_{C1}+\frac{|V_{BE1}|}{R_3}=I_{C2}+\frac{|V_{BE1}|}{R_2} \tag{12.66}$$

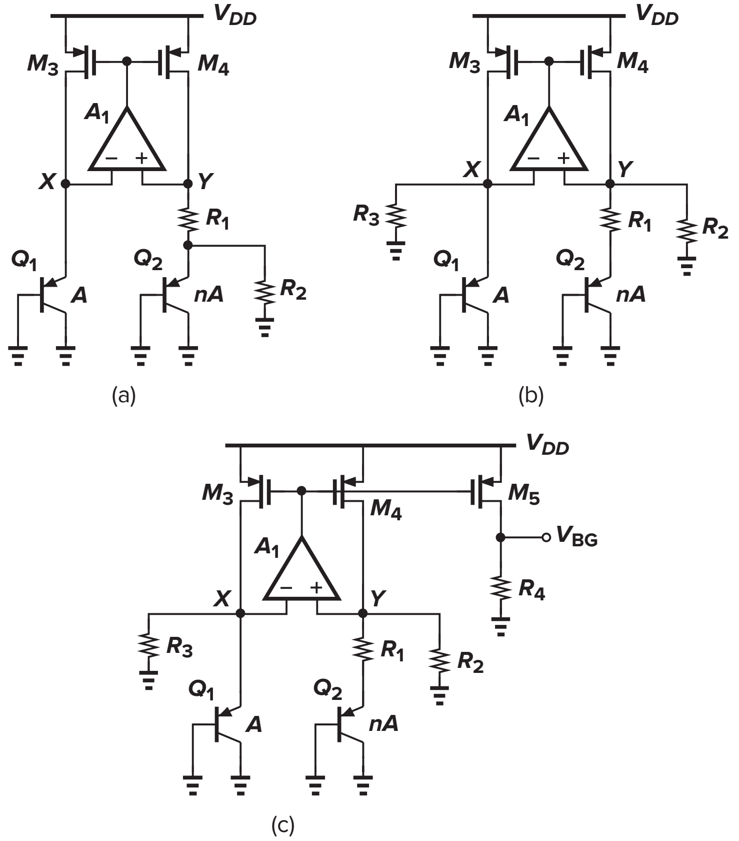


Figure 12.32 (a) Attempt to make drain current of M_4 temperature-independent, (b) circuit modification resulting in a zero-TC current, and (c) generation of arbitrarily small voltage with zero TC.

which yields $I_{C1} = I_{C2}$ if $R_2 = R_3$. We still have $|V_{BE1}| = |V_{BE2}| + I_{C2}R_1$ and hence $I_{C2} = V_T \ln n/R_1$. This current and the current flowing through R_2, $|V_{BE1}|/R_2$, constitute $|I_{D4}|$:

$$|I_{D4}| = \frac{V_T \ln n}{R_1} + \frac{|V_{BE1}|}{R_2} \tag{12.67}$$

$$= \frac{1}{R_2}\left(|V_{BE1}| + \frac{R_2}{R_1} V_T \ln n\right) \tag{12.68}$$

Selecting $(R_2/R_1)V_T \ln n$ approximately equal to $17.2V_T$ renders a zero TC for I_{D4}. This current is then copied and passed through a resistor to generate a zero-TC voltage [Fig. 12.32(c)] [8]:

$$V_{BG} = \frac{R_4}{R_2}\left(|V_{BE1}| + \frac{R_2}{R_1} V_T \ln n\right) \tag{12.69}$$

(if M_5 is identical to M_4). We choose $(R_2/R_1)\ln n \approx 17.2$, observing that V_{BG} has a zero TC and its value can be lower than the conventional limit of 1.25 V.

▶ Example 12.8

If the op amp in Fig. 12.32(c) has an input-referred offset voltage, V_{OS}, determine V_{BG}.

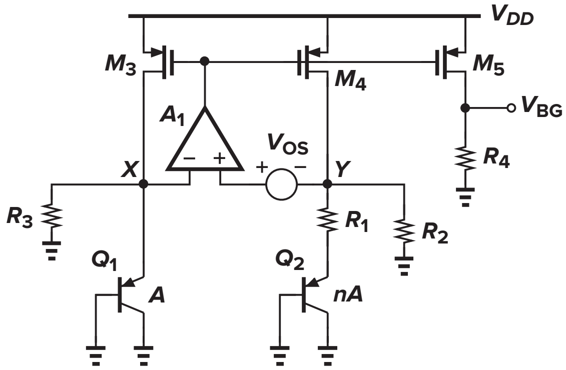


Figure 12.33

Solution

As shown in Fig. 12.33, we now have $V_X \approx V_Y + V_{OS} \approx |V_{BE1}|$ and

$$I_{C1} + \frac{|V_{BE1}|}{R_3} = I_{C2} + \frac{|V_{BE1}| - V_{OS}}{R_2} \tag{12.70}$$

which implies that $I_{C1} = I_{C2} - V_{OS}/R_2$ if $R_2 = R_3$. Since $|V_{BE1}| = |V_{BE2}| + R_1 I_{C2} + V_{OS}$, we have $I_{C2} = (V_T \ln n - V_{OS})/R_1$. This current and the current flowing through R_2, $(|V_{BE1}| - V_{OS})/R_2$, add up to $|I_{D4}|$:

$$|I_{D4}| = \frac{V_T \ln n - V_{OS}}{R_1} + \frac{|V_{BE1}| - V_{OS}}{R_2} \tag{12.71}$$

It follows that

$$V_{BG} = \frac{R_4}{R_2}\left(|V_{BE1}| + \frac{R_2}{R_1} V_T \ln n\right) - \frac{R_4}{R_1||R_2} V_{OS} \tag{12.72}$$

revealing that the op amp offset is amplified by a factor of $R_4/(R_1||R_2)$. Alternatively, we can write

$$V_{BG} = \frac{R_4}{R_2}\left[|V_{BE1}| + \frac{R_2}{R_1} V_T \ln n - \left(1 + \frac{R_2}{R_1}\right) V_{OS}\right] \tag{12.73}$$

concluding that the effect of V_{OS} can be minimized only by maximizing n.

It is instructive to estimate the lowest supply voltage with which the circuit of Fig. 12.32(c) can operate properly. With large bipolar transistors and a small bias current, e.g., 10 μA, the base-emitter voltage can be as low as 0.7 V. Similarly, wide PMOS devices allow a $|V_{DS}|$ of about 50 mV. The circuit can thus operate with a minimum V_{DD} of around 0.75 V. In this case, R_4 tends to be a large resistor, e.g., 50 kΩ, producing significant noise and requiring a bypass capacitor at the output. Also, if the PMOS drain currents are copied to generate a larger current, say, 0.5 mA, then their noise is amplified by the same factor. This noise contains thermal and flicker components due to the PMOS devices and the noise of the op amp. In Problem 12.24, we analyze the noise behavior of this circuit, but from Example 12.8, we observe that the op amp input noise is amplified by a factor of $R_4/(R_1||R_2)$.

The op amp in Fig. 12.32(c) can be realized as a five-transistor OTA. Depicted in Fig. 12.34(a) is an example. The OTA design proceeds according to the following guidelines. (1) Large transistor dimensions are chosen so as to minimize their flicker noise and offset. (2) The gate-source voltage of M_a and M_b

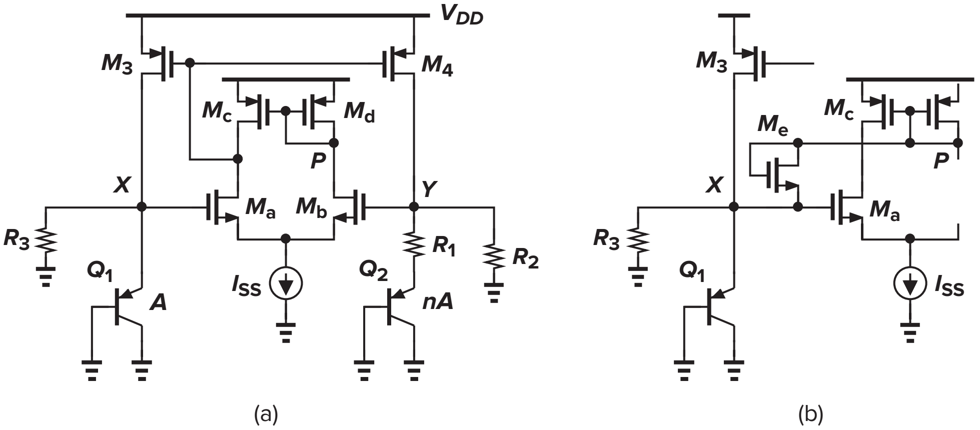


Figure 12.34 (a) Implementation of low-voltage BG circuit using a five-transistor OTA, and (b) addition of start-up device.

plus the headroom required by I_{SS} must not exceed $|V_{BE1}|$. (3) The transistors are chosen long enough to yield a reasonable loop gain, e.g., 5 to 10.

The foregoing topology must incorporate a start-up mechanism. Otherwise, the circuit begins with $V_X = V_Y = 0$, M_a and M_b remain off, and so do M_3 and M_4. Since, with $V_{DD} < 1$ V, the voltage difference between node P and node X is initially positive but finally negative (why?), we can tie a diode-connected NMOS transistor between these two nodes to ensure start-up [Fig. 12.34(b)]. Alternatively, the NMOS device can be connected between X and V_{DD}.

Another low-voltage bandgap circuit can be derived from the topology of Fig. 12.20 by simply tying a resistor from the output node to ground [9]. Shown in Fig. 12.35, the circuit now allows some of I_{D5} to flow through R_3:

$$|I_{D5}| = \frac{V_{out}}{R_3} + \frac{V_{out} - |V_{BE3}|}{R_2} \tag{12.74}$$

If the PMOS devices are identical, $|I_{D5}| = V_T \ln n / R_1$, yielding

$$V_{out} = \frac{R_3}{R_2 + R_3}\left(|V_{BE3}| + \frac{R_2}{R_1} V_T \ln n\right) \tag{12.75}$$

The standard bandgap voltage is thus scaled down by a factor of $R_3/(R_2 + R_3)$. The reader is encouraged to compute the effect of the op amp offset at the output and compare the result with (12.72).

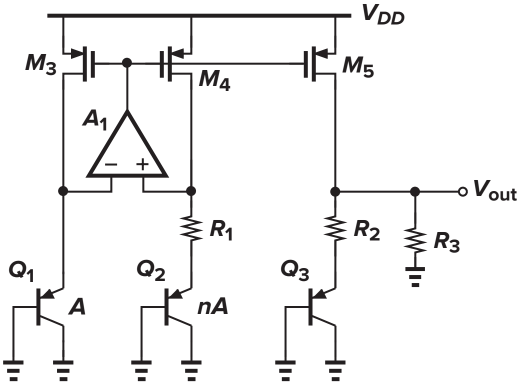


Figure 12.35 Alternative low-voltage BG circuit.

It is possible to add other bias branches to the foregoing circuits so as to provide curvature correction, but such schemes typically rely on trimming because the various mismatches within the circuit tend to shift the zero-TC temperature randomly. Other low-voltage bandgaps are described in [10].

12.8 ■ Case Study

In this section, we study a bandgap reference circuit designed for high-precision analog systems [7]. The reference generator incorporates the topology of Fig. 12.19, but with two series base-emitter voltages in each branch so as to reduce the effect of MOSFET mismatches. A simplified version of the core is depicted in Fig. 12.36, where the PMOS current mirror arrangement ensures equal collector currents for Q_1–Q_4. While requiring a high supply voltage, this design exemplifies issues that prove important in practice.

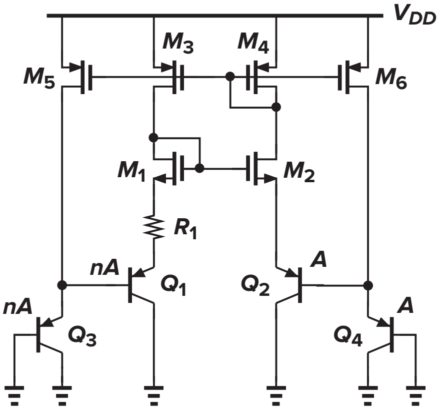


Figure 12.36 Simplified core of the bandgap circuit reported in [7].

Channel-length modulation of the MOS devices in Fig. 12.36 still results in significant supply dependence. To resolve this issue, each branch can employ both NMOS and PMOS cascode topologies. Figure 12.37(a) shows an example in which the low-voltage cascode current mirror described in Chapter 5

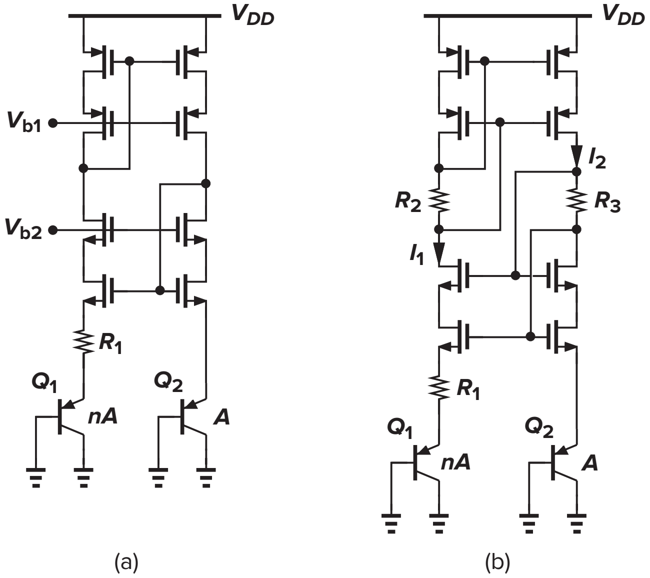


Figure 12.37 (a) Addition of cascode devices to improve supply rejection; (b) use of self-biased cascode to eliminate V_{b1} and V_{b2}.

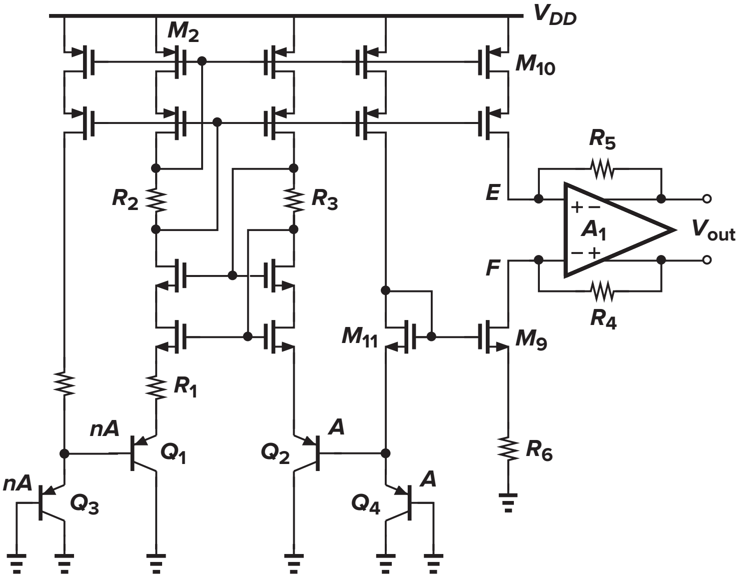


Figure 12.38 Generation of a floating reference voltage.

is utilized. To obviate the need for V_{b1} and V_{b2}, this design actually introduces a "self-biased" cascode, shown in Fig. 12.37(b), where R_2 and R_3 sustain proper voltages to allow all MOSFETs to remain in saturation. This cascode topology is analyzed in Problem 12.7.

The bandgap circuit reported in [7] is designed to generate a *floating* reference. This is accomplished by the modification shown in Fig. 12.38, where the drain currents of M_9 and M_{10} flow through R_4 and R_5, respectively. Note that M_{11} sets the gate voltage of M_9 at $V_{BE4} + V_{GS11}$, establishing a voltage equal to V_{BE4} across R_6 if M_9 and M_{11} are identical. Thus, $I_{D9} = V_{BE4}/R_6$, yielding $V_{R4} = V_{BE4}(R_4/R_6)$. Also, if M_{10} is identical to M_2, then $|I_{D10}| = 2(V_T \ln n)/R_1$, and hence $V_{R5} = 2(V_T \ln n)(R_5/R_1)$. Since the op amp ensures that $V_E \approx V_F$, we have

$$V_{out} = \frac{R_4}{R_6} V_{BE4} + 2\frac{R_5}{R_1} V_T \ln n \tag{12.76}$$

Proper choice of the resistor ratios and n therefore provides a zero temperature coefficient.

In order to further enhance the supply rejection, this design regulates the supply voltage of the core and the op amp. Illustrated in Fig. 12.39, the idea is to generate a local supply, V_{DDL}, that is defined by a reference V_{R1} and the ratio of R_{r1} and R_{r2} and hence remains relatively independent of the global supply voltage. But how is V_{R1} itself generated? To minimize the dependence of V_{R1} upon the supply,

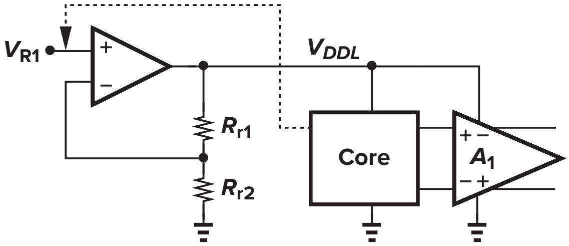


Figure 12.39 Regulation of the supply voltage of the core and op amp to improve supply rejection.

this voltage is established *inside* the core, as depicted in Fig. 12.40. In fact, R_M is chosen such that V_{R1} is a bandgap reference.

Figure 12.41 shows the overall implementation, omitting a few details for simplicity. A start-up circuit is also used. Operating from a 5-V supply, the reference generator produces a 2.00-V output while consuming 2.2 mW. The supply rejection is 94 dB at low frequencies, dropping to 58 dB at 100 kHz [7].

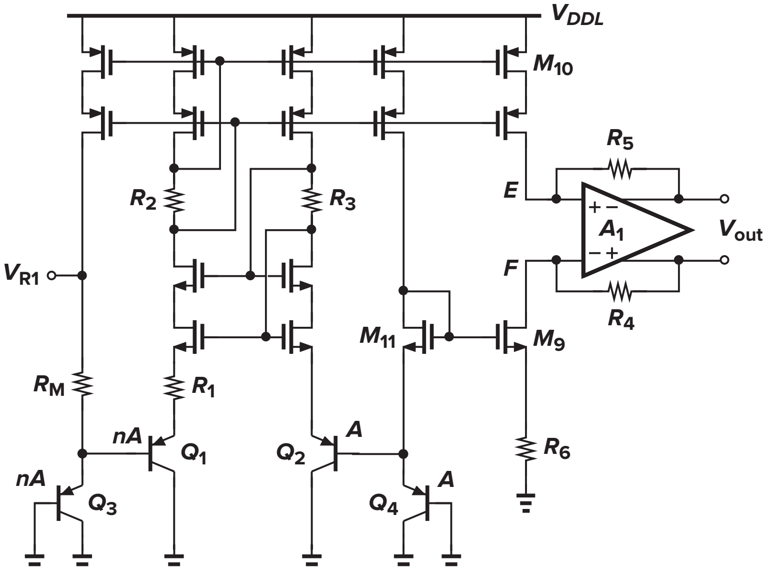


Figure 12.40 Generation of V_{R1}, used in Fig. 12.39.

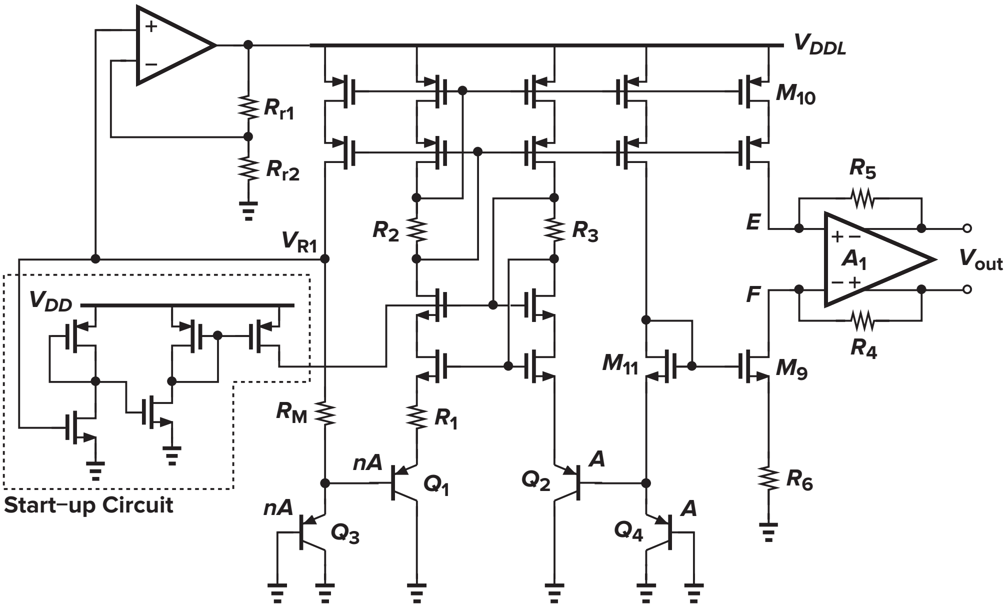


Figure 12.41 Overall circuit of the bandgap generator reported in [7].

References

[1] R. A. Blauschild et al., "A New NMOS Temperature-Stable Voltage Reference," *IEEE J. of Solid-State Circuits,* vol. 13, pp. 767–774, December 1978.

[2] Y. P. Tsividis and R. W. Ulmer, "A CMOS Voltage Reference," *IEEE J. of Solid-State Circuits*, vol. 13, pp. 774–778, December 1978.

[3] D. Hilbiber, "A New Semiconductor Voltage Standard," *ISSCC Dig. of Tech. Papers*, pp. 32–33, February 1964.

[4] K. E. Kujik, "A Precision Reference Voltage Source," *IEEE J. of Solid-State Circuits*, vol. 8, pp. 222–226, June 1973.

[5] G. C. M. Meijer, P. C. Schmall, and K. van Zalinge, "A New Curvature-Corrected Bandgap Reference," *IEEE J. of Solid-State Circuits*, vol. 17, pp. 1139–1143, December 1982.

[6] M. Gunawan et al., "A Curvature-Corrected Low-Voltage Bandgap Reference," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 667–670, June 1993.

[7] T. Brooks and A. L. Westwisk, "A Low-Power Differential CMOS Bandgap Reference," *ISSCC Dig. of Tech. Papers,* pp. 248–249, February 1994.

[8] H. Banba et al., "A CMOS Bandgap Reference Circuit with Sub-1-V Operation," *IEEE J. of Solid-State Circuits*, vol. 34, pp. 670–674, May 1999.

[9] H. Neuteboom et al., "A DSP-Based Hearing Instrument IC," *IEEE J. of Solid-State Circuits*, vol. 32, pp. 1790–1806, November 1997.

[10] C. J. B. Fayomi et al., "Sub-1-V CMOS Bandgap Reference Design Techniques: A Survey," *Analog Integrated Circuits and Signal Processing*, vol. 62, pp. 141–157, February 2010.

[11] B. Gilbert, "Monolithic Voltage and Current References: Themes and Variations," pp. 269–352 in *Analog Circuit Design*, J. H. Huijsing, R. J. van de Plassche, and W. M. C. Sansen, eds. (Boston: Kluwer Academic Publishers, 1996).

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary.

12.1. Derive an expression for I_{out} in Fig. 12.42.

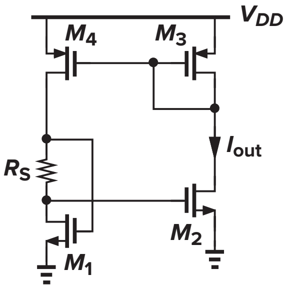


Figure 12.42

12.2. Explain how the start-up circuit shown in Fig. 12.43 operates. Derive a relationship that guarantees that $V_X < V_{TH}$ after the circuit turns on.

12.3. Consider the circuit of Fig. 12.15.

(a) If M_1 and M_2 suffer from channel-length modulation, what is the error in the output voltage?

(b) Repeat part (a) for M_3 and M_4.

(c) If M_1 and M_2 have a threshold mismatch of ΔV, i.e., $V_{TH1} = V_{TH}$ and $V_{TH2} = V_{TH} + \Delta V$, what is the error in the output voltage?

(d) Repeat part (c) for M_3 and M_4.

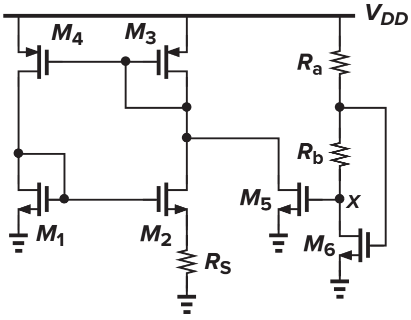


Figure 12.43

12.4. In Fig. 12.15, if the open-loop gain of the op amp A_1 is not sufficiently large, then $|V_X - V_Y|$ exceeds V_e, where V_e is the maximum tolerable error. Calculate the minimum value of A_1 in terms of V_e such that the condition $|V_X - V_Y| < V_e$ is satisfied.

12.5. In the circuit of Fig. 12.15, assume that Q_2 and Q_4 have a finite current gain β. Calculate the error in the output voltage.

12.6. Calculate the output noise voltage of the circuit shown in Fig. 12.30 due to the thermal and flicker noise of M_1 and M_2.

12.7. Consider the self-biased cascode shown in Fig. 12.44. Determine the minimum and maximum values of RI_{REF} such that both M_1 and M_2 remain in saturation.

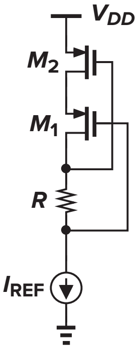


Figure 12.44

12.8. The circuit of Fig. 12.3(a) sometimes turns on even with no explicit start-up mechanism. Identify the capacitive path(s) that couple the transition on V_{DD} to the internal nodes and hence provide the start-up current.

12.9. Sketch the temperature coefficient of V_{BE} [Eq. (12.13)] versus temperature. Some iteration may be necessary.

12.10. Determine the derivative of Eq. (12.13) with respect to temperature and sketch the result versus T. This quantity reveals the curvature of the voltage.

12.11. Suppose that in Fig. 12.9, the amplifier has an output resistance R_{out}. Calculate the error in V_{out}.

12.12. The circuit of Fig. 12.9 is designed with $R_3 = 1\ \text{k}\Omega$ and a current of 50 μA through it. Calculate $R_1 = R_2$ and n for a zero TC.

12.13. In the circuit of Fig. 12.15, Q_1 and Q_2 are biased at 100 μA and Q_3 and Q_4 at 50 μA. If $R_1 = 1\ \text{k}\Omega$, calculate R_2 and $(W/L)_{1-4}$ such that the circuit operates with $V_{DD} = 3$ V. Which op amp topology can be used here?

12.14. Since the bandgap of silicon exhibits a small temperature coefficient, Eq. (12.48) suggests that $\partial V_{REF}/\partial T \propto (4+m)k/q$, a relatively large value, whereas we derived V_{REF} such that it has a zero TC. Explain the flaw in this argument.

12.15. A differential pair with resistive loads is designed such that its voltage gain, $g_m R_D$, has a zero TC at room temperature. If only the temperature dependence of the mobility is considered, determine the required temperature behavior of the tail current. Design a circuit that roughly approximates this behavior.

12.16. In Problem 12.15, assume that the tail current is constant, but the load resistors exhibit a finite TC. What resistor temperature coefficient cancels the variation of the mobility at room temperature?

12.17. In the circuit of Fig. 12.32(b), how should R_1–R_3 be chosen so that the negative-feedback loop is stronger than the positive-feedback loop?

12.18. Does the five-transistor OTA in Fig. 12.34(a) impose additional supply voltage constraints?

12.19. Figure 12.45 illustrates a "single-junction" bandgap design [11]. Here, switches S_1 and S_2 are driven by complementary clocks.

(a) What is V_{out} when S_1 is on and S_2 is off?

(b) What is the change in V_{out} when S_1 turns off and S_2 turns on?

(c) How are I_1, I_2, C_1, and C_2 chosen to produce a zero-TC output when S_1 is off?

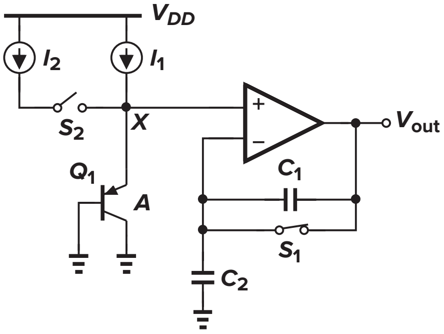


Figure 12.45

12.20. Suppose that in Fig. 12.45, I_2/I_1 deviates from its nominal value by a small error ϵ. Calculate V_{out} when S_1 is off.

12.21. The circuit of Fig. 12.20 is designed with $(W/L)_{1-4} = 50/0.5$, $I_{D1} = I_{D2} = 50\ \mu$A, $R_1 = 1$ kΩ, and $R_2 = 2$ kΩ. Assume that $\lambda = \gamma = 0$ and Q_3 is identical to Q_1.

(a) Determine n and $(W/L)_5$ such that V_{out} has a zero TC at room temperature.

(b) Neglecting the noise contribution of Q_1–Q_3, calculate the output thermal noise.

12.22. Consider the circuit of Fig. 12.21. Assume $K = 4$, $f_{CK} = 50$ MHz, and a power budget of 1 mW. Determine the aspect ratio of M_1–M_4 and the value of C_S such that $g_{m1} = 1/(500\ \Omega)$.

12.23. Suppose $(W/L)_3 = K(W/L)_4$ in Fig. 12.32(c). How should R_2 and R_3 be chosen?

12.24. Determine the output noise voltage of the circuit in Fig. 12.32(c).

12.25. Analyze the circuit of Fig. 12.3(a) if R_S is placed in series with the source of M_1.

CHAPTER

13

Introduction to Switched-Capacitor Circuits

Our study of amplifiers in previous chapters has dealt only with cases in which the input signal is continuously available and applied to the circuit and the output signal is continuously observed. Called "continuous-time" circuits, such amplifiers find wide application in audio, video, and high-speed analog systems. In many situations, however, we may sense the input only at periodic instants of time, ignoring its value at other times. The circuit then processes each "sample," producing a valid output at the end of each period. Such circuits are called "discrete-time" or "sampled-data" systems.

In this chapter, we study a common class of discrete-time systems called "switched-capacitor (SC) circuits." Our objective is to provide the foundation for more advanced topics such as filters, comparators, ADCs, and DACs. Most of our study deals with switched-capacitor amplifiers, but the concepts can be applied to other discrete-time circuits as well. Beginning with a general view of SC circuits, we describe sampling switches and their speed and precision issues. Next, we analyze switched-capacitor amplifiers, considering unity-gain, noninverting, and multiply-by-two topologies. Finally, we examine a switched-capacitor integrator.

13.1 ■ General Considerations

In order to understand the motivation for sampled-data circuits, let us first consider the simple continuous-time amplifier shown in Fig. 13.1(a), where V_{out}/V_{in} is ideally equal to $-R_2/R_1$. Used extensively with bipolar op amps, this circuit presents a difficult issue if implemented in CMOS technology. Recall that, to achieve a high voltage gain, the open-loop output resistance of CMOS op amps is maximized, typically approaching hundreds of kilohms. We therefore suspect that R_2 heavily drops the open-loop gain, degrading the precision of the circuit. In fact, with the aid of the simple equivalent circuit shown in

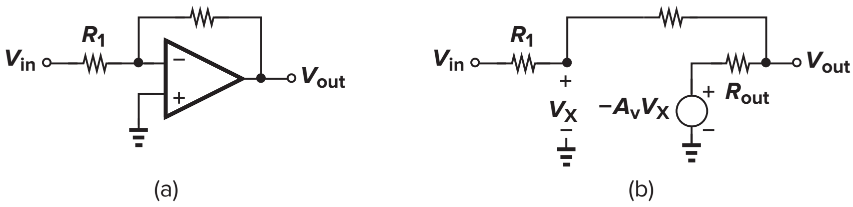


Figure 13.1 (a) Continuous-time feedback amplifier; (b) equivalent circuit of (a).

Fig. 13.1(b), we can write

$$-A_v\left(\frac{V_{out}-V_{in}}{R_1+R_2}R_1+V_{in}\right)-R_{out}\frac{V_{out}-V_{in}}{R_1+R_2}=V_{out} \tag{13.1}$$

and hence

$$\frac{V_{out}}{V_{in}}=-\frac{R_2}{R_1}\cdot\frac{A_v-\dfrac{R_{out}}{R_2}}{1+\dfrac{R_{out}}{R_1}+A_v+\dfrac{R_2}{R_1}} \tag{13.2}$$

Equation (13.2) implies that, compared to the case where $R_{out}=0$, the closed-loop gain suffers from inaccuracies in both the numerator and the denominator. Also, the input resistance of the amplifier, approximately equal to R_1, loads the preceding stage while introducing thermal noise.

▶ Example 13.1

Using the feedback techniques described in Chapter 8, calculate the closed-loop gain of the circuit of Fig. 13.1(a) and compare the result with Eq. (13.2).

Solution

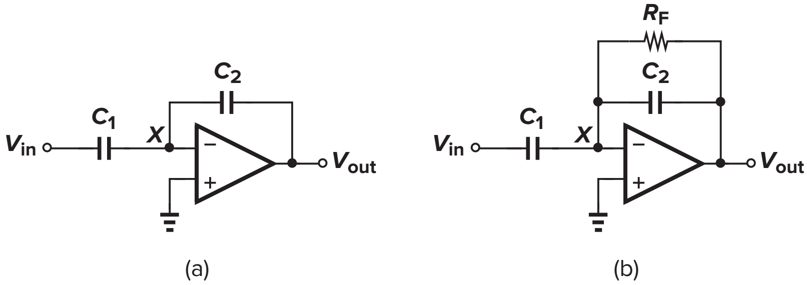
With the aid of the approach described in Example 8.16, the reader can prove that

$$\frac{V_{out}}{V_{in}}=\frac{-R^2A_v}{R_2^2+R_1R_{out}+R_2R_{out}+(1+A_v)R_1R_2} \tag{13.3}$$

$$=-\frac{R_2}{R_1}\cdot\frac{A_v}{\dfrac{R_2}{R_1}+\dfrac{R_{out}}{R_2}+\dfrac{R_{out}}{R_1}+1+A_v} \tag{13.4}$$

The two results are approximately equal if $R_{out}/R_2 \ll A_v$, a condition required to ensure that the transmission through R_2 is negligible.

In the circuit of Fig. 13.1(a), the closed-loop gain is set by the ratio of R_2 and R_1. In order to avoid reducing the open-loop gain of the op amp, we postulate that the resistors can be replaced by capacitors [Fig. 13.2(a)]. Ideally, the gain of the circuit is equal to the impedance of C_2 divided by the impedance of C_1 and multiplied by -1, i.e., equal to $-C_1/C_2$.



(a) (b)

Figure 13.2 (a) Continuous-time feedback amplifier using capacitors; (b) use of resistor to define bias point.

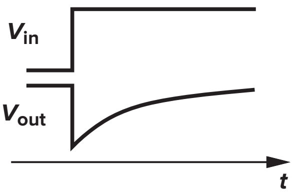


Figure 13.3 Step response of the amplifier of Fig. 13.2(b).

But, how is the bias voltage at node X set?[1] We may add a large feedback resistor as in Fig. 13.2(b), providing dc feedback while negligibly affecting the ac behavior of the amplifier in the frequency band of interest. Such an arrangement is indeed practical if the circuit senses *only* high-frequency signals. But suppose, for example, the circuit is to amplify a voltage step. Illustrated in Fig. 13.3, the response contains a step change due to the initial amplification by the circuit consisting of C_1, C_2, and the op amp, followed by a "tail" resulting from the loss of charge on C_2 through R_F. From another point of view, the circuit may not be suited to amplify *wideband* signals because it exhibits a high-pass transfer function. In fact, the transfer function is given by

$$\frac{V_{out}}{V_{in}}(s) \approx -\frac{R_F \dfrac{1}{C_2 s}}{R_F + \dfrac{1}{C_2 s}} \div \frac{1}{C_1 s} \tag{13.5}$$

$$= -\frac{R_F C_1 s}{R_F C_2 s + 1} \tag{13.6}$$

indicating that $V_{out}/V_{in} \approx -C_1/C_2$ only if $\omega \gg (R_F C_2)^{-1}$.

The above difficulty can be remedied by increasing $R_F C_2$, but in many applications the required values of the two components become prohibitively large. We must therefore seek other methods of establishing the bias while utilizing capacitive feedback networks.

It is possible to replace R_F in Fig. 13.2(b) with a *switch*. Illustrated in Fig. 13.4, the idea is to turn S_2 on so as to place the op amp in unity-gain feedback and force V_X to V_B, an appropriately chosen input common-mode level for the op amp. After the switch turns off, node X retains this voltage, allowing proper operation. Of course, when S_2 is on, the circuit does not amplify V_{in}.

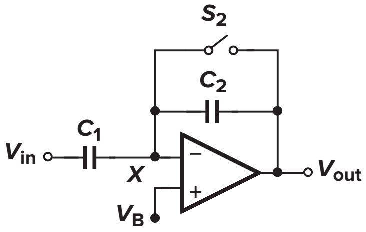


Figure 13.4 Use of feedback switch to define dc input level.

Let us now consider the switched-capacitor circuit depicted in Fig. 13.5, where three switches control the operation: S_1 and S_3 connect the left plate of C_1 to V_{in} and ground, respectively, and S_2 provides unity-gain feedback. We first assume that the open-loop gain of the op amp is very large and study the circuit in two phases. First, S_1 and S_2 are on and S_3 is off, yielding the equivalent circuit of Fig. 13.6(a). For a high-gain op amp, $V_B = V_{out} \approx 0$, and hence the voltage across C_1 is approximately equal to V_{in}.

[1] The bias voltage is given by the initial condition at this node and hence is ambiguous.

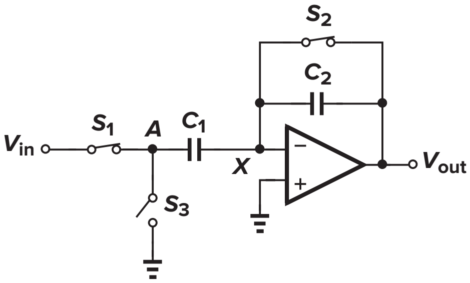


Figure 13.5 Switched-capacitor amplifier.

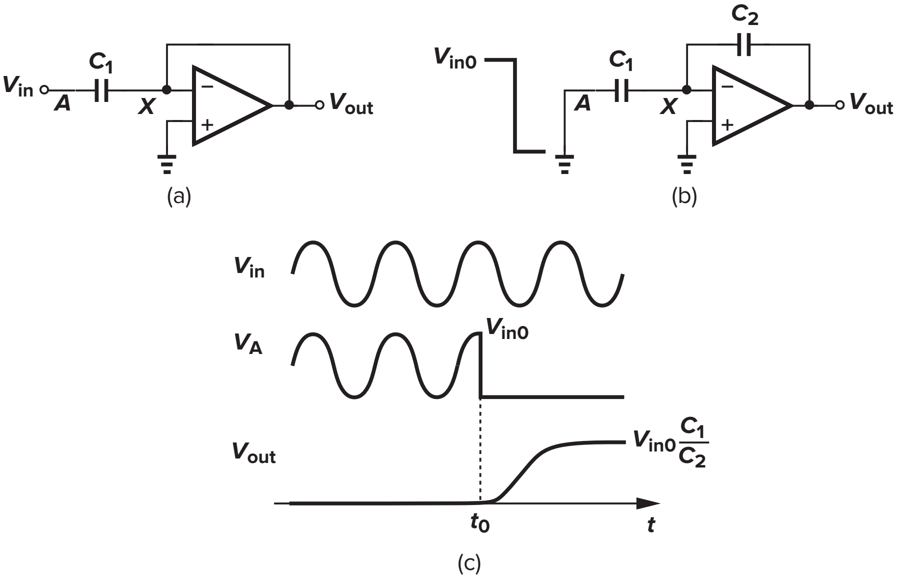


Figure 13.6 Circuit of Fig. 13.5 in (a) sampling mode, (b) amplification mode (c) input and output waveforms in the two modes.

We say that C_1 samples the input. Next, at $t = t_0$, S_1 and S_2 turn off and S_3 turns on, pulling node A to ground. Since the gain is equal to $-C_1/C_2$ and since V_A changes from V_{in0} to 0, the output voltage must change from zero to $V_{in0}C_1/C_2$.

The output voltage change can also be calculated by examining the transfer of charge. Note that the charge stored on C_1 just before t_0 is equal to $V_{in0}C_1$. After $t = t_0$, the negative feedback through C_2 drives the op amp input differential voltage, and hence the voltage across C_1, to zero (Fig. 13.7). The charge stored on C_1 at $t = t_0$ must then be transferred to C_2, producing an output voltage equal to $V_{in0}C_1/C_2$. Thus, the circuit amplifies V_{in0} by a factor of C_1/C_2.

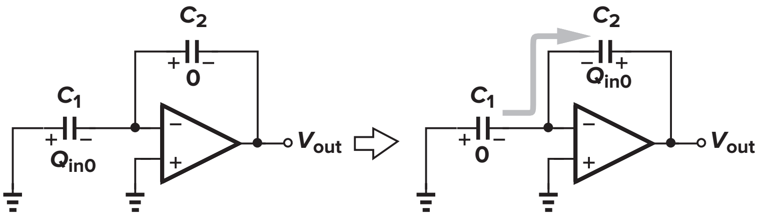


Figure 13.7 Transfer of charge from C_1 to C_2.

Several attributes of the circuit of Fig. 13.5 distinguish it from continuous-time implementations. First, the circuit devotes some time to "sampling" the input, setting the output to zero and providing no amplification during this period. Second, after sampling, for $t > t_0$, the circuit ignores the input voltage V_{in}, amplifying the sampled voltage. Third, the circuit configuration changes considerably from one phase to another, as seen in Fig. 13.6(a) and (b), raising concern about its stability. Note that S_2 must turn on periodically to compensate for the leakage currents that slowly discharge X. These currents arise from S_2 itself and the gate leakage of the op amp.

What is the advantage of the amplifier of Fig. 13.5 over that in Fig. 13.1? In addition to sampling capability, we note from the waveforms depicted in Fig. 13.6 that after V_{out} settles to $V_{in} \cdot C_1/C_2$, the current through C_2 approaches zero. That is, the feedback capacitor does not reduce the open-loop gain of the amplifier if the output voltage is given enough time to settle. In Fig. 13.1, on the other hand, R_2 loads the amplifier continuously.

The switched-capacitor amplifier of Fig. 13.5 lends itself to implementation in CMOS technology much more easily than in other technologies. This is because discrete-time operations require switches to perform sampling as well as a high input impedance to sense the stored quantities with no corruption. For example, if the op amp of Fig. 13.5 incorporates bipolar transistors at its input, the base current drawn from the inverting input in the amplification phase [Fig. 13.6(b)] creates an error in the output voltage. The existence of simple switches and a high input impedance have made CMOS technology the dominant choice for sampled-data applications.

The foregoing discussion leads to the conceptual view illustrated in Fig. 13.8 for switched-capacitor amplifiers. In the simplest case, the operation takes place in two phases: sampling and amplification. Thus, in addition to the analog input, V_{in}, the circuit requires a clock to define each phase.

Our study of SC amplifiers proceeds according to these two phases. First, we analyze various sampling techniques. Second, we consider SC amplifier topologies.

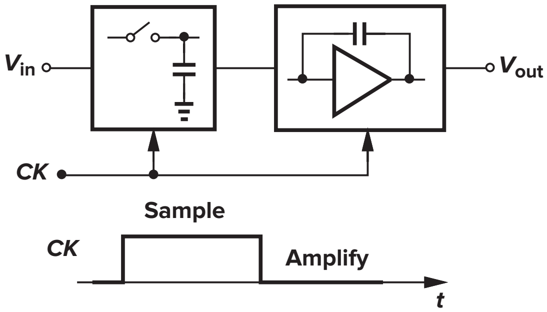


Figure 13.8 General view of switched-capacitor amplifier.

13.2 ■ Sampling Switches

13.2.1 MOSFETS as Switches

A simple sampling circuit consists of a switch and a capacitor [Fig. 13.9(a)]. A MOS transistor can serve as a switch [Fig. 13.9(b)] because it can be on while carrying a zero current.

To understand how the circuit of Fig. 13.9(b) samples the input, first consider the simple cases depicted in Fig. 13.10, where the gate command, CK, goes high at $t = t_0$. In Fig. 13.10(a), we assume that $V_{in} = 0$ and the capacitor has an initial voltage equal to V_{DD}. Thus, at $t = t_0$, M_1 senses a gate-source voltage equal to V_{DD} while its drain voltage is also equal to V_{DD}. The transistor therefore operates in saturation, drawing a current of $I_{D1} = (\mu_n C_{ox}/2)(W/L)(V_{DD} - V_{TH})^2$ from the capacitor. As V_{out} falls, at some

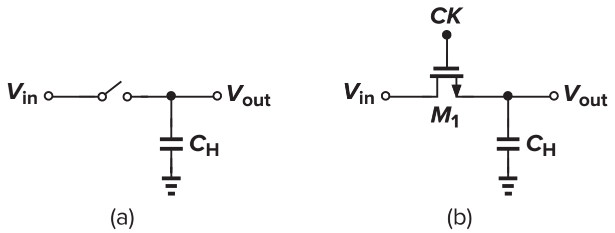


Figure 13.9 (a) Simple sampling circuit; (b) implementation of the switch by a MOS device.

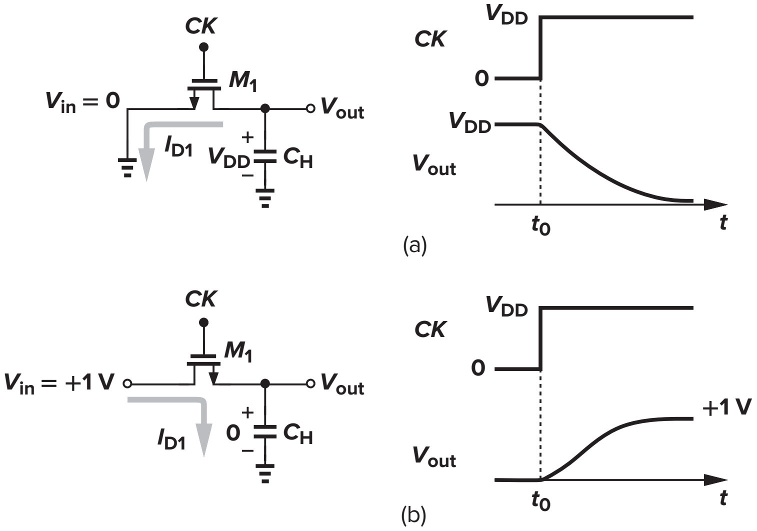


Figure 13.10 Response of a sampling circuit to different input levels and initial conditions.

point $V_{out} = V_{DD} - V_{TH}$, driving M_1 into the triode region. The device nevertheless continues to discharge C_H until V_{out} approaches zero. We note that for $V_{out} \ll 2(V_{DD} - V_{TH})$, the transistor can be viewed as a resistor equal to $R_{on} = [\mu_n C_{ox}(W/L)(V_{DD} - V_{TH})]^{-1}$.

Now consider the case in Fig. 13.10(b), where $V_{in} = +1$ V, $V_{out}(t = t_0) = 0$ V, and $V_{DD} = 3$ V. Here, the terminal of M_1 connected to C_H acts as the source, and the transistor turns on with $V_{GS} = +3$ V, but $V_{DS} = +1$ V. Thus, M_1 operates in the triode region, charging C_H until V_{out} approaches $+1$ V. For $V_{out} \approx +1$ V, M_1 exhibits an on-resistance of $R_{on} = [\mu_n C_{ox}(W/L)(V_{DD} - V_{in} - V_{TH})]^{-1}$.

The above observations reveal two important points. First, a MOS switch can conduct current in either direction simply by exchanging the role of its source and drain terminals. Second, as shown in Fig. 13.11, when the switch is on, V_{out} follows V_{in}, and when the switch is off, V_{out} remains constant. Thus, the circuit "tracks" the signal when CK is high and "freezes" the instantaneous value of V_{in} across C_H when CK goes low.

▶ Example 13.2

In the circuit of Fig. 13.10(a), calculate V_{out} as a function of time. Assume that $\lambda = 0$.

Solution

Before V_{out} drops below $V_{DD} - V_{TH}$, M_1 is saturated and we have

$$V_{out}(t) = V_{DD} - \frac{I_{D1}t}{C_H} \tag{13.7}$$

$$= V_{DD} - \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{DD} - V_{TH})^2 \frac{t}{C_H} \tag{13.8}$$

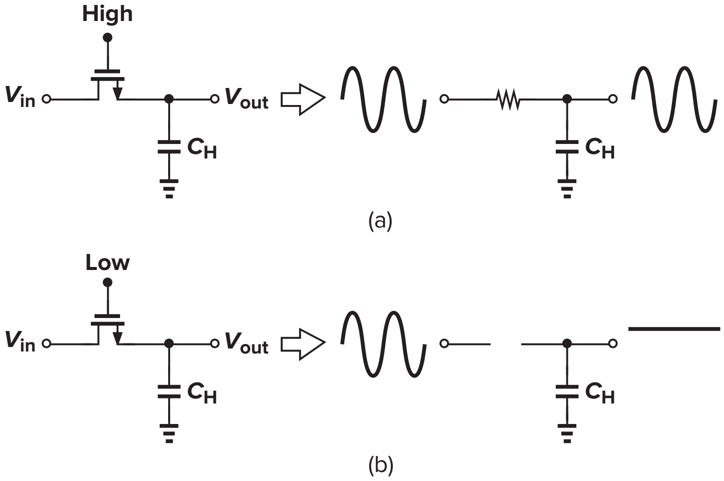


Figure 13.11 Track and hold capabilities of a sampling circuit.

After

$$t_1 = \frac{2V_{TH}C_H}{\mu_n C_{ox}\dfrac{W}{L}(V_{DD} - V_{TH})^2} \tag{13.9}$$

M_1 enters the triode region, yielding a time-dependent current. We therefore write

$$C_H\frac{dV_{out}}{dt} = -I_{D1} \tag{13.10}$$

$$= -\frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2(V_{DD} - V_{TH})V_{out} - V_{out}^2\right] \quad t > t_1 \tag{13.11}$$

Rearranging (13.11), we have

$$\frac{dV_{out}}{[2(V_{DD} - V_{TH}) - V_{out}]V_{out}} = -\frac{1}{2}\mu_n\frac{C_{ox}}{C_H}\frac{W}{L}dt \tag{13.12}$$

which, upon separation into partial fractions, is written as

$$\left[\frac{1}{V_{out}} + \frac{1}{2(V_{DD} - V_{TH}) - V_{out}}\right]\frac{dV_{out}}{V_{DD} - V_{TH}} = -\mu_n\frac{C_{ox}}{C_H}\frac{W}{L}dt \tag{13.13}$$

Thus,

$$\ln V_{out} - \ln[2(V_{DD} - V_{TH}) - V_{out}] = -(V_{DD} - V_{TH})\mu_n\frac{C_{ox}}{C_H}\frac{W}{L}(t - t_1) \tag{13.14}$$

that is

$$\ln\frac{V_{out}}{2(V_{DD} - V_{TH}) - V_{out}} = -(V_{DD} - V_{TH})\mu_n\frac{C_{ox}}{C_H}\frac{W}{L}(t - t_1) \tag{13.15}$$

Taking the exponential of both sides and solving for V_{out}, we obtain

$$V_{out} = \frac{2(V_{DD} - V_{TH})\exp\left[-(V_{DD} - V_{TH})\mu_n\dfrac{C_{ox}}{C_H}\cdot\dfrac{W}{L}(t - t_1)\right]}{1 + \exp\left[-(V_{DD} - V_{TH})\mu_n\dfrac{C_{ox}}{C_H}\cdot\dfrac{W}{L}(t - t_1)\right]} \tag{13.16}$$

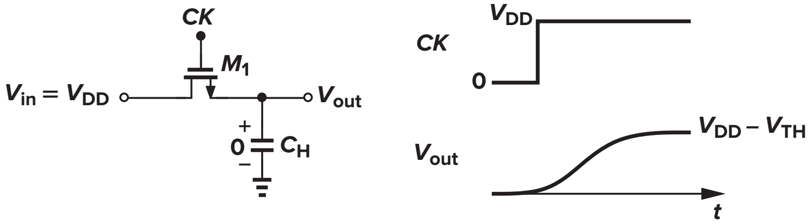


Figure 13.12 Maximum output level in an NMOS sampler.

In the circuit of Fig. 13.10(b), we assumed that $V_{in} = +1$ V (Fig. 13.12). Now suppose $V_{in} = V_{DD}$. How does V_{out} vary with time? Since the gate and drain of M_1 are at the same potential, the transistor is saturated, and we have

$$C_H \frac{dV_{out}}{dt} = I_{D1} \tag{13.17}$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{DD} - V_{out} - V_{TH})^2 \tag{13.18}$$

where channel-length modulation is neglected. It follows that

$$\frac{dV_{out}}{(V_{DD} - V_{out} - V_{TH})^2} = \frac{1}{2}\mu_n \frac{C_{ox}}{C_H}\frac{W}{L}dt \tag{13.19}$$

and hence

$$\left.\frac{1}{V_{DD} - V_{out} - V_{TH}}\right|_0^{Vout} = \left.\frac{1}{2}\mu_n \frac{C_{ox}}{C_H}\frac{W}{L}t\right|_0^t \tag{13.20}$$

where body effect is neglected and $V_{out}(t = 0)$ is assumed zero. Thus,

$$V_{out} = V_{DD} - V_{TH} - \frac{1}{\dfrac{1}{2}\mu_n \dfrac{C_{ox}}{C_H}\dfrac{W}{L}t + \dfrac{1}{V_{DD} - V_{TH}}} \tag{13.21}$$

Equation (13.21) implies that as $t \to \infty$, $V_{out} \to V_{DD} - V_{TH}$. This is because as V_{out} approaches $V_{DD} - V_{TH}$, the overdrive voltage of M_1 vanishes, reducing the current available for charging C_H to negligible values. Of course, even for $V_{out} = V_{DD} - V_{TH}$, the transistor conducts some subthreshold current and, given enough time, eventually brings V_{out} to V_{DD}. Nonetheless, as mentioned in Chapter 3, for typical operation speeds, it is reasonable to assume that V_{out} does not exceed $V_{DD} - V_{TH}$.

The foregoing analysis demonstrates a serious limitation of MOS switches: if the input signal level is close to V_{DD}, then the output provided by an NMOS switch cannot track the input. From another point of view, the on-resistance of the switch increases considerably as the input and output voltages approach $V_{DD} - V_{TH}$. We may then ask—What is the maximum input level that the switch can pass to the output faithfully? In Fig. 13.12, for $V_{out} \approx V_{in}$, the transistor must operate in the deep triode region, and hence the upper bound of V_{in} equals $V_{DD} - V_{TH}$. As explained later, in practice, V_{in} must be quite lower than this value.

▶ **Example 13.3**

In the circuit of Fig. 13.13, calculate the minimum and maximum on-resistance of M_1. Assume that $\mu_n C_{ox} = 50\ \mu$A/V^2, $W/L = 10/1$, $V_{TH} = 0.7$ V, $V_{DD} = 3$ V, and $\gamma = 0$.

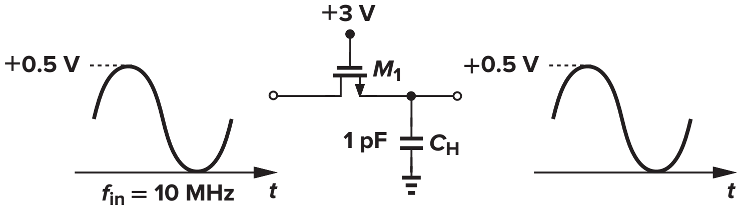


Figure 13.13

Solution

We note that in the steady state, M_1 remains in the triode region because the gate voltage is higher than both V_{in} and V_{out} by a value greater than V_{TH}. If $f_{in} = 10$ MHz, we predict that V_{out} tracks V_{in} with a negligible phase shift due to the on-resistance of M_1 and C_H. Assuming that $V_{out} \approx V_{in}$, we need not distinguish between the source and drain terminals, obtaining

$$R_{on1} = \frac{1}{\mu_n C_{ox} \dfrac{W}{L}(V_{DD} - V_{in} - V_{TH})} \tag{13.22}$$

Thus, $R_{on1,max} \approx 1.11$ kΩ and $R_{on1,min} \approx 870$ Ω. By contrast, if the maximum input level is raised to 1.5 V, then $R_{on1,max} = 2.5$ kΩ.

MOS devices operating in the deep triode region are sometimes called "zero-offset" switches to emphasize that they exhibit no dc shift between the input and output voltages of the simple sampling circuit of Fig. 13.9(b).[2] This is evident from the examples of Fig. 13.10, where the output eventually becomes equal to the input. Nonexistent in bipolar technology, the zero-offset property proves crucial in precise sampling of analog signals.

We have thus far considered only NMOS switches. The reader can verify that the foregoing principles apply to PMOS switches as well. In particular, as shown in Fig. 13.14, a PMOS transistor fails to operate as a switch if its gate is grounded and its drain terminal senses an input voltage of $|V_{THP}|$ or less. In other words, the on-resistance of the device rises rapidly as the input and output levels drop to $|V_{THP}|$ above ground.

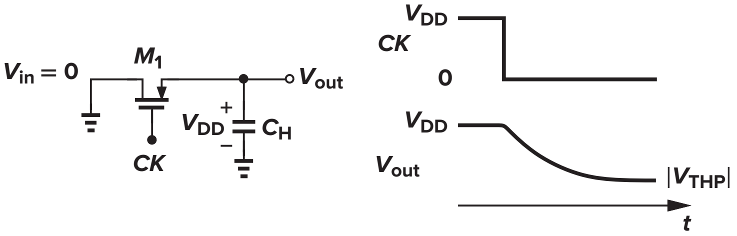


Figure 13.14 Sampling circuit using PMOS switch.

13.2.2 Speed Considerations

What determines the speed of the sampling circuits of Fig. 13.9? We must first define the speed here. Illustrated in Fig. 13.15, a simple measure of speed is the time required for the output voltage to go

[2] We assume the circuit following the sampler draws no input dc current.

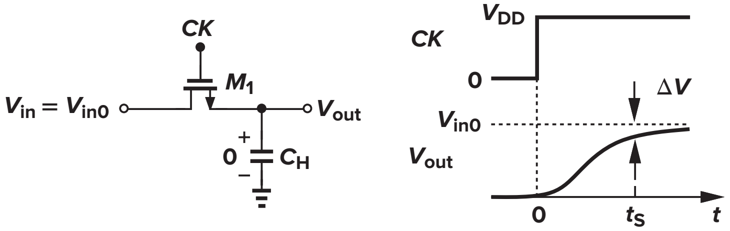


Figure 13.15 Definition of speed in a sampling circuit.

from zero to the maximum input level after the switch turns on. Since V_{out} would take infinite time to become equal to V_{in0}, we consider the output settled when it is within a certain "error band," ΔV, around the final value. For example, we say that the output settles to 0.1% accuracy after t_S seconds, meaning that in Fig. 13.15, $\Delta V/V_{in0} = 0.1\%$. Thus, the speed specification must be accompanied by an accuracy specification as well. Note that after $t = t_S$, we can consider the source and drain voltages to be approximately equal.

From the circuit of Fig. 13.15, we surmise that the sampling speed is given by two factors: the on-resistance of the switch and the value of the sampling capacitor. Thus, to achieve a higher speed, a large aspect ratio and a small capacitor must be used. However, as illustrated in Fig. 13.13, the on-resistance also depends on the input level, yielding a greater time constant for more positive inputs (in the case of NMOS switches). From Eq. (13.22), we plot the on-resistance of the switch as a function of the input level [Fig. 13.16(a)], noting the sharp rise as V_{in} approaches $V_{DD} - V_{TH}$. For example, if we restrict the variation of R_{on} to a range of 4 to 1, then the maximum input level is given by

$$\frac{1}{\mu_n C_{ox}\dfrac{W}{L}(V_{DD} - V_{in,max} - V_{TH})} = \frac{4}{\mu_n C_{ox}\dfrac{W}{L}(V_{DD} - V_{TH})} \tag{13.23}$$

That is

$$V_{in,max} = \frac{3}{4}(V_{DD} - V_{TH}) \tag{13.24}$$

This value falls around $V_{DD}/2$, translating to severe swing limitations. Note that the device threshold voltage directly limits the voltage swings.[3]

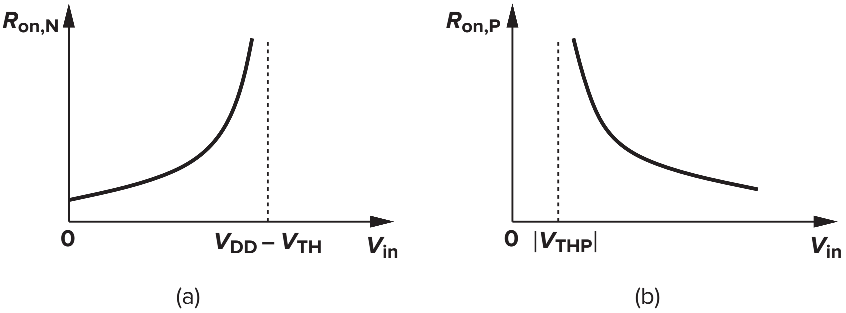


Figure 13.16 On-resistance of (a) NMOS and (b) PMOS devices as a function of input voltage.

[3] By contrast, the output swing of cascode stages is limited by overdrive voltages rather than by the threshold voltage.

In order to accommodate greater voltage swings in a sampling circuit, we first observe that a PMOS switch exhibits an on-resistance that *decreases* as the input voltage becomes more positive [Fig. 13.16(b)]. It is then plausible to employ "complementary" switches so as to allow rail-to-tail swings. Shown in Fig. 13.17(a), such a combination requires complementary clocks, producing an equivalent resistance:

$$\begin{aligned} R_{on,eq} &= R_{on,N} || R_{on,P} \\ &= \frac{1}{\mu_n C_{ox}(W/L)_N(V_{DD} - V_{in} - V_{THN})} \Big|\Big| \frac{1}{\mu_p C_{ox}(W/L)_P(V_{in} - |V_{THP}|)} \end{aligned}$$

It follows that

$$R_{on,eq} = \frac{1}{\mu_n C_{ox}(W/L)_N(V_{DD} - V_{THN}) - [\mu_n C_{ox}(W/L)_N - \mu_p C_{ox}(W/L)_P]V_{in} - \mu_p C_{ox}(W/L)_P|V_{THP}|}$$

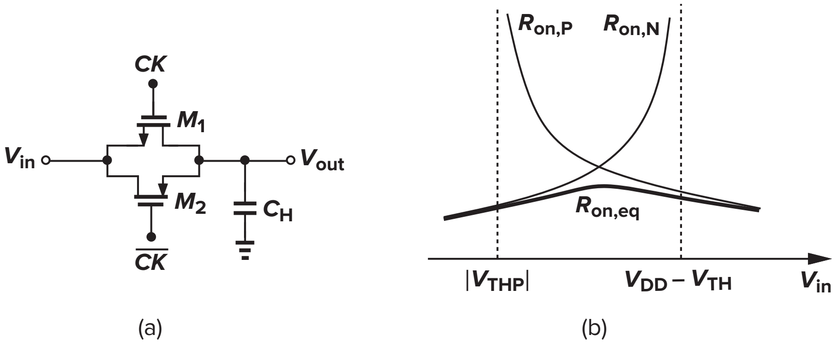


Figure 13.17 (a) Complementary switch; (b) on-resistance of the complementary switch.

Interestingly, if $\mu_n C_{ox}(W/L)_N = \mu_p C_{ox}(W/L)_P$, then $R_{on,eq}$ is independent of the input level.[4] Figure 13.17(b) plots the behavior of $R_{on,eq}$ in the general case, revealing much less variation than that corresponding to each switch alone. We quantify the effect of switch nonlinearity in Chapter 14.

For high-speed input signals, it is critical that the NMOS and PMOS switches in Fig. 13.17(a) turn off simultaneously so as to avoid ambiguity in the sampled value. If, for example, the NMOS device turns off Δt seconds earlier than the PMOS device, then the output voltage tends to track the input for the remaining Δt seconds, but with a large, input-dependent time constant (Fig. 13.18). This effect gives rise to distortion in the sampled value. For moderate precision, the simple circuit shown in Fig. 13.19 provides complementary clocks by duplicating the delay of inverter I_1 through the pass gate G_2.

13.2.3 Precision Considerations

Our foregoing study of MOS switches indicates that a larger W/L or a smaller sampling capacitor results in a higher speed. In this section, we show that these methods of increasing the speed degrade the precision with which the signal is sampled.

Three mechanisms in MOS transistor operation introduce error at the instant the switch turns off. We study each effect individually.

[4] In reality, V_{THN} and V_{THP} vary with V_{in} through body effect, but we ignore this variation here.

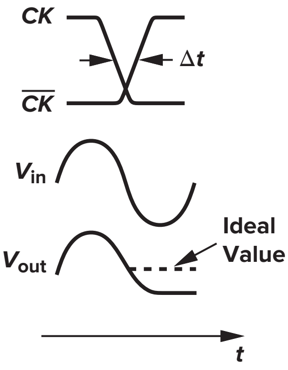


Figure 13.18 Distortion generated if complementary switches do not turn off simultaneously.

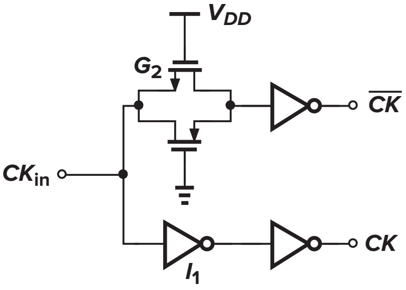


Figure 13.19 Simple circuit generating complementary clocks.

Channel Charge Injection Consider the sampling circuit of Fig. 13.20, and recall that for a MOSFET to be on, a channel must exist at the oxide-silicon interface. Assuming that $V_{in} \approx V_{out}$, we use our derivations in Chapter 2 to express the total charge in the inversion layer as

$$Q_{ch} = WLC_{ox}(V_{DD} - V_{in} - V_{TH}) \tag{13.25}$$

where L denotes the effective channel length. When the switch turns off, Q_{ch} exits through the source and drain terminals, a phenomenon called "channel charge injection."

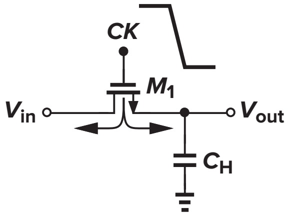


Figure 13.20 Charge injection when a switch turns off.

The charge injected to the left side of Fig. 13.20 is absorbed by the input source, creating no error. On the other hand, the charge injected to the right side is deposited on C_H, introducing an error in the voltage stored on the capacitor. For example, if half of Q_{ch} is injected onto C_H, the resulting error equals

$$\Delta V = \frac{WLC_{ox}(V_{DD} - V_{in} - V_{TH})}{2C_H} \tag{13.26}$$

Illustrated in Fig. 13.21, the error for an NMOS switch appears as a negative "pedestal" at the output. Note that the error is directly proportional to WLC_{ox} and inversely proportional to C_H.

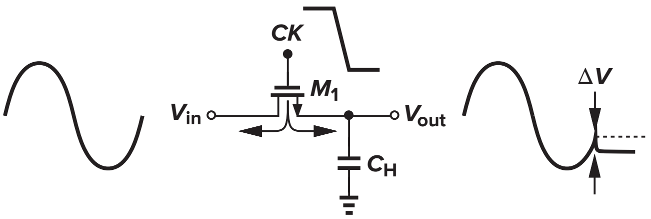


Figure 13.21 Effect of charge injection.

An important question that arises now is—Why did we assume in arriving at (13.26) that exactly *half* of the channel charge is injected onto C_H? In reality, the fraction of charge that exits through the source and drain terminals is a relatively complex function of various parameters, such as the impedance seen at each terminal to ground and the transition time of the clock [1, 2]. Investigations of this effect have not yielded any rule of thumb that can predict the charge splitting in terms of such parameters. Furthermore, in many cases, these parameters, e.g., the clock transition time, are poorly controlled. Also, most circuit simulation programs model charge injection quite inaccurately. As a worst-case estimate, we can assume that the entire channel charge is injected onto the sampling capacitor.

How does charge injection affect the precision? Assuming that all of the charge is deposited on the capacitor, we express the sampled output voltage as

$$V_{out} \approx V_{in} - \frac{WLC_{ox}(V_{DD} - V_{in} - V_{TH})}{C_H} \tag{13.27}$$

where the phase shift between the input and the output is neglected. Thus,

$$V_{out} = V_{in}\left(1 + \frac{WLC_{ox}}{C_H}\right) - \frac{WLC_{ox}}{C_H}(V_{DD} - V_{TH}) \tag{13.28}$$

suggesting that the output deviates from the ideal value through two effects: a nonunity gain equal to $1 + WLC_{ox}/C_H$,[5] and a constant offset voltage $-WLC_{ox}(V_{DD} - V_{TH})/C_H$ (Fig. 13.22). In other words, since we have assumed that channel charge is a *linear* function of the input voltage, the circuit exhibits only gain error and dc offset.

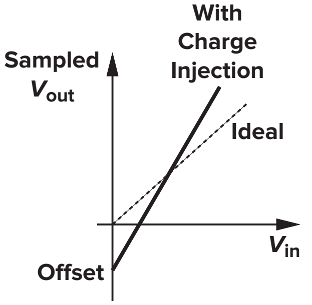


Figure 13.22 Input/output characteristic of sampling circuit in the presence of charge injection.

In the foregoing discussion, we tacitly assumed that V_{TH} is constant. However, for NMOS switches (in an n-well technology), body effect must be taken into account.[6] Since $V_{TH} = V_{TH0} + \gamma(\sqrt{2\phi_B + V_{SB}} - \sqrt{2\phi_B})$,

[5] The voltage gain is *greater* than unity because the pedestal becomes smaller as the input level rises.

[6] Even for PMOS switches, the n-well is connected to the most positive supply voltage because the source and drain terminals of the switch may interchange during sampling.

and $V_{BS} \approx -V_{in}$, we have

$$V_{out} = V_{in} - \frac{WLC_{ox}}{C_H}\left(V_{DD} - V_{in} - V_{TH0} - \gamma\sqrt{2\phi_B + V_{in}} + \gamma\sqrt{2\phi_B}\right), \tag{13.29}$$

$$= V_{in}\left(1 + \frac{WLC_{ox}}{C_H}\right) + \gamma\frac{WLC_{ox}}{C_H}\sqrt{2\phi_B + V_{in}} - \frac{WLC_{ox}}{C_H}\left(V_{DD} - V_{TH0} + \gamma\sqrt{2\phi_B}\right) \tag{13.30}$$

It follows that the nonlinear dependence of V_{TH} upon V_{in} introduces nonlinearity in the input/output characteristic.

In summary, charge injection contributes three types of errors in MOS sampling circuits: gain error, dc offsets, and nonlinearity. In many applications, the first two can be tolerated or corrected whereas the last cannot.

It is instructive to consider the speed-precision trade-off resulting from charge injection. Representing the speed by a simple time constant τ and the precision by the error ΔV due to charge injection, we define a figure of merit as $F = (\tau\Delta V)^{-1}$. Writing

$$\tau = R_{on}C_H \tag{13.31}$$

$$= \frac{1}{\mu_n C_{ox}(W/L)(V_{DD} - V_{in} - V_{TH})}C_H \tag{13.32}$$

and

$$\Delta V = \frac{WLC_{ox}}{C_H}(V_{DD} - V_{in} - V_{TH}) \tag{13.33}$$

we have

$$F = \frac{\mu_n}{L^2} \tag{13.34}$$

Thus, to the first order, the trade-off is independent of the switch width and the sampling capacitor.

Clock Feedthrough In addition to channel charge injection, a MOS switch couples the clock transitions to the sampling capacitor through its gate-drain or gate-source overlap capacitance. Depicted in Fig. 13.23, the effect introduces an error in the sampled output voltage. Assuming the overlap capacitance is constant, we express the error as

$$\Delta V = V_{CK}\frac{WC_{ov}}{WC_{ov} + C_H} \tag{13.35}$$

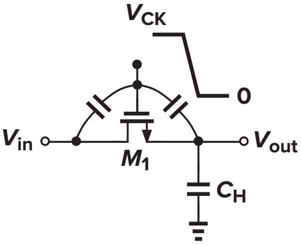


Figure 13.23 Clock feedthrough in a sampling circuit.

where C_{ov} is the overlap capacitance per unit width. The error ΔV is independent of the input level, manifesting itself as a constant offset in the input/output characteristic. As with charge injection, clock feedthrough leads to a trade-off between speed and precision as well.

***kT/C* Noise** Recall from Example 7.3 that a resistor charging a capacitor gives rise to a total rms noise voltage of $\sqrt{kT/C}$. As shown in Fig. 13.24, a similar effect occurs in sampling circuits. The on-resistance of the switch introduces thermal noise at the output and, when the switch turns off, this noise is stored on the capacitor along with the instantaneous value of the input voltage. It can be proved that the rms voltage of the sampled noise in this case is still approximately equal to $\sqrt{kT/C}$ [3, 4].

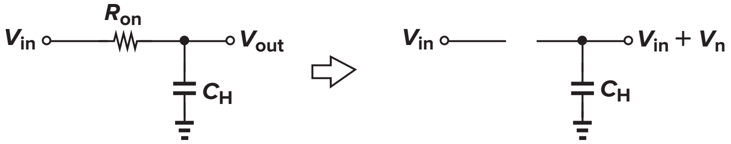


Figure 13.24 Thermal noise in a sampling circuit.

The problem of kT/C noise limits the performance in many high-precision applications. In order to achieve low noise, the sampling capacitor must be sufficiently large, thus loading other circuits and degrading the speed.

13.2.4 Charge Injection Cancellation

The dependence of charge injection upon the input level and the trade-off expressed by (13.34) make it necessary to seek methods of canceling the effect of charge injection so as to achieve a higher F. We consider a few such techniques here.

To arrive at the first technique, we postulate that the charge injected by the main transistor can be *removed* by means of a second transistor. As shown in Fig. 13.25, a "dummy" switch, M_2, driven by $\overline{CK}$ is added to the circuit such that after M_1 turns off and M_2 turns on, the channel charge deposited by the former on C_H is absorbed by the latter to create a channel. Note that both the source and drain of M_2 are connected to the output node.

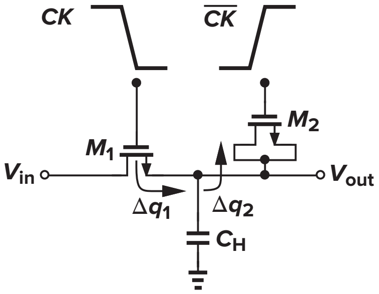


Figure 13.25 Addition of dummy device to reduce charge injection and clock feedthrough.

How do we ensure that the charge injected by M_1, Δq_1, is equal to that absorbed by M_2, Δq_2? Suppose half of the channel charge of M_1 is injected onto C_H, i.e.,

$$\Delta q_1 = \frac{W_1 L_1 C_{ox}}{2}(V_{CK} - V_{in} - V_{TH1}) \tag{13.36}$$

Since $\Delta q_2 = W_2 L_2 C_{ox}(V_{CK} - V_{in} - V_{TH2})$, if we choose $W_2 = 0.5W_1$ and $L_2 = L_1$, then $\Delta q_2 = \Delta q_1$. Unfortunately, the assumption of equal splitting of charge between source and drain is generally invalid, making this approach less attractive.

Interestingly, with the choice $W_2 = 0.5W_1$ and $L_2 = L_1$, the effect of clock feedthrough is suppressed. As depicted in Fig. 13.26, the total charge in V_{out} is zero because

$$-V_{CK}\frac{W_1C_{ov}}{W_1C_{ov}+C_H+2W_2C_{ov}}+V_{CK}\frac{2W_2C_{ov}}{W_1C_{ov}+C_H+2W_2C_{ov}}=0 \tag{13.37}$$

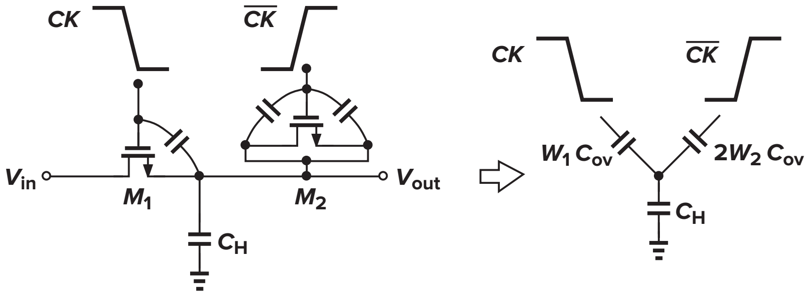


Figure 13.26 Clock feedthrough suppression by dummy switch.

Another approach to lowering the effect of charge injection incorporates both PMOS and NMOS devices such that the opposite charge packets injected by the two cancel each other (Fig. 13.27). For Δq_1 to cancel Δq_2, we must have $W_1L_1C_{ox}(V_{CK}-V_{in}-V_{THN}) = W_2L_2C_{ox}(V_{in}-|V_{THP}|)$. Thus, the cancellation occurs for only one input level. Even for clock feedthrough, the circuit does not provide complete cancellation because the gate-drain overlap capacitance of NFETs is not equal to that of PFETs.

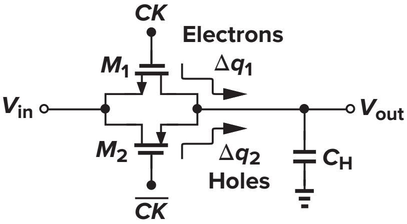


Figure 13.27 Use of complementary switches to reduce charge injection.

Our knowledge of the advantages of differential circuits suggests that the problem of charge injection may be relieved through differential operation. As shown in Fig. 13.28, we surmise that charge injection appears as a common-mode disturbance. But, writing $\Delta q_1 = WLC_{ox}(V_{CK}-V_{in1}-V_{TH1})$ and $\Delta q_2 = WLC_{ox}(V_{CK}-V_{in2}-V_{TH2})$, we recognize that $\Delta q_1 = \Delta q_2$ only if $V_{in1} = V_{in2}$. In other words, the overall error is not suppressed for differential signals. Nevertheless, this technique both removes the

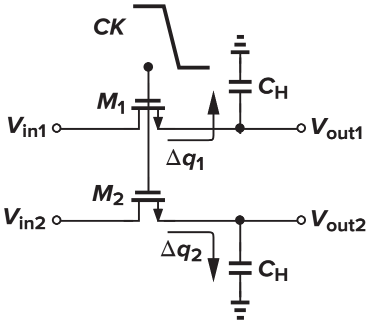


Figure 13.28 Differential sampling circuit.

constant offset and lowers the nonlinear component. This can be understood by writing

$$\Delta q_1 - \Delta q_2 = WLC_{ox}[(V_{in2} - V_{in1}) + (V_{TH2} - V_{TH1})] \tag{13.38}$$

$$= WLC_{ox}\left[V_{in2} - V_{in1} + \gamma\left(\sqrt{2\phi_F + V_{in2}} - \sqrt{2\phi_F + V_{in1}}\right)\right] \tag{13.39}$$

Since for $V_{in1} = V_{in2}$, $\Delta q_1 - \Delta q_2 = 0$, the characteristic exhibits no offset. Also, the nonlinearity of body effect now appears in both square-root terms of (13.39), leading to only odd-order distortion (Chapter 14).

The problem of charge injection continues to limit the speed-precision envelope in sampled-data systems. Many cancellation techniques have been introduced, but each leads to other trade-offs. One such technique, called "bottom-plate sampling," is widely used in switched-capacitor circuits and is described later in this chapter.

13.3 ■ Switched-Capacitor Amplifiers

As mentioned in Sec. 13.1 and exemplified by the circuit of Fig. 13.5, CMOS feedback amplifiers are more easily implemented with a capacitive feedback network than with a resistive one. Having examined sampling techniques, we are now ready to study a number of switched-capacitor amplifiers. Our objective is to understand the underlying principles as well as the speed-precision trade-offs encountered in the design of each circuit.

Before studying SC amplifiers, it is helpful to look briefly at the physical implementation of capacitors in CMOS technology. A simple capacitor structure is shown in Fig. 13.29(a), where the "top plate" and the "bottom plate" are realized by metal layers. An important concern in using this structure is the parasitic capacitance between each plate and the substrate. In particular, the bottom plate suffers from capacitance, C_p, to the underlying substrate—a value typically 5 to 10% of the main capacitance. For this reason, we usually model the capacitor as in Fig. 13.29(b). Monolithic capacitors are described in more detail in Chapters 18 and 19.

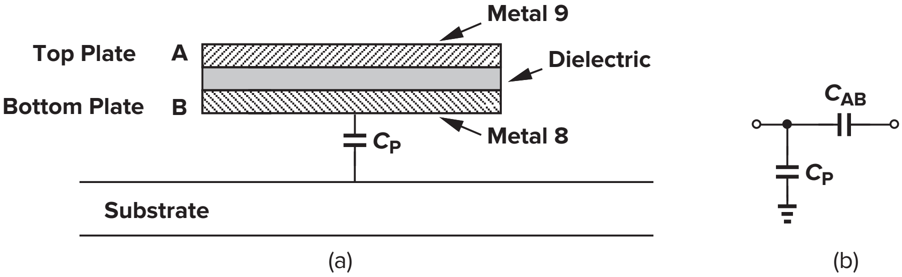


Figure 13.29 (a) Monolithic capacitor structure; (b) circuit model of (a) including parasitic capacitance to the substrate.

13.3.1 Unity-Gain Sampler/Buffer

While a unity-gain amplifier can be realized with no resistors or capacitors in the feedback network [Fig. 13.30(a)], for discrete-time applications, it still requires a sampling circuit. We may therefore conceive the circuit shown in Fig. 13.30(b) as a sampler/buffer. However, the input-dependent charge injected by S_1 onto C_H limits the accuracy here.

Now consider the topology depicted in Fig. 13.31(a), where three switches control the sampling and amplification modes. In the sampling mode, S_1 and S_2 are on and S_3 is off, yielding the topology shown in Fig. 13.31(b). Thus, $V_{out} = V_X \approx 0$, and the voltage across C_H tracks V_{in}. At $t = t_0$, when $V_{in} = V_0$,

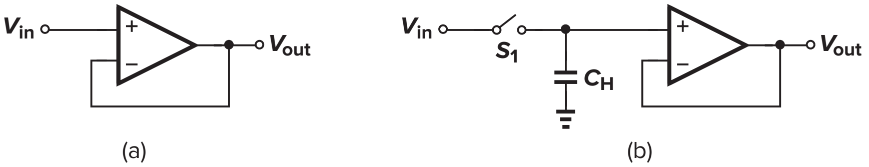


Figure 13.30 (a) Unity-gain buffer; (b) sampling circuit followed by unity-gain buffer.

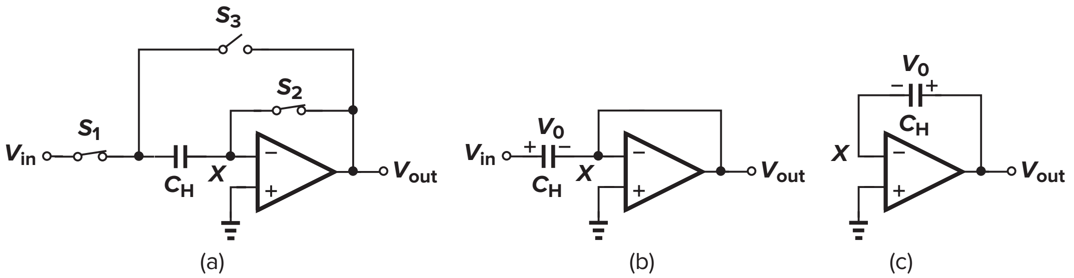


Figure 13.31 (a) Unity-gain sampler; (b) circuit of (a) in sampling mode; (c) circuit of (a) in amplification mode.

S_1 and S_2 turn off and S_3 turns on, flipping the capacitor around the op amp and entering the circuit into the amplification mode [Fig. 13.31(c)]. Since the op amp's high gain requires that node X still be a virtual ground and since the charge on the capacitor must be conserved, V_{out} rises to a value approximately equal to V_0. This voltage is therefore "frozen," and it can be processed by subsequent stages.

With proper timing, the circuit of Fig. 13.31(a) can substantially alleviate the problem of channel charge injection. As Fig. 13.32 illustrates in "slow motion," during the transition from the sampling mode to the amplification mode, S_2 turns off slightly *before* S_1 does. We carefully examine the effect of the charge injected by S_2 and S_1. When S_2 turns off, it injects a charge packet Δq_2 onto C_H, producing an error equal to $\Delta q_2/C_H$. However, this charge is independent of the input level because node X is a virtual ground. For example, if S_2 is realized by an NMOS device whose gate voltage equals V_{CK}, then $\Delta q_2 = WLC_{ox}(V_{CK} - V_{TH} - V_X)$.

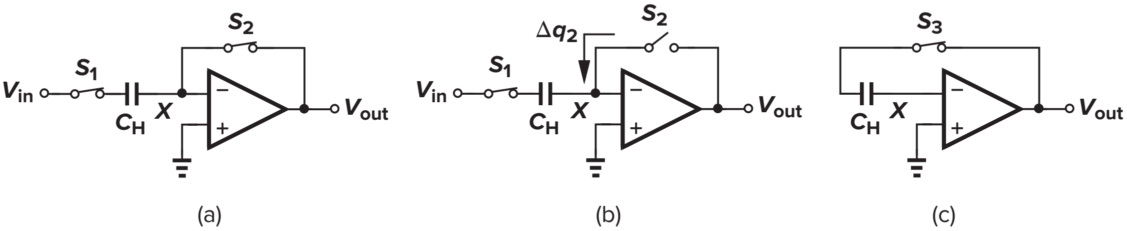


Figure 13.32 Operation of the unity-gain sampler in slow motion.

The constant magnitude of Δq_2 means that the channel charge of S_2 introduces only an offset (rather than gain error or nonlinearity) in the input/output characteristic. As described below, this offset can easily be removed by differential operation. But, how about the charge injected by S_1 onto C_H? Let us set V_{in} to zero and suppose that S_1 injects a charge packet Δq_1 onto node P (after S_2 has turned off) [Fig. 13.33(a)]. If the capacitance connected from X to ground (including the input capacitance of the op amp) is zero, V_P and V_X jump to infinity. To simplify the analysis, we assume a capacitance equal to C_X from X to ground [Fig. 13.33(b)], and we will see shortly that its value does not affect the results. In

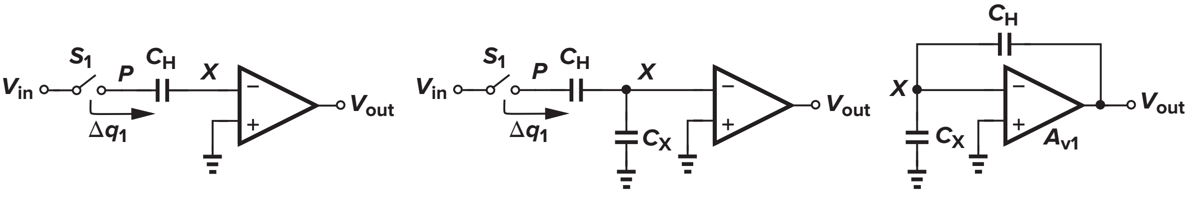


Figure 13.33 Effect of charge injected by S_1 with (a) zero and (b) finite op amp input capacitance; (c) transition of circuit to amplification mode.

Fig. 13.33(b), each of the series capacitors C_H and C_X carries a charge equal to Δq_1. Now, as shown in Fig. 13.33(c), we place C_H around the op amp, seeking to obtain the resulting output voltage.

To calculate the output voltage, we must make an important observation: the total charge at node X cannot change after S_2 turns off because no path exists for electrons to flow into or out of this node. Thus, if before S_1 turns off, the total charge on the right plate of C_H and the top plate of C_X is zero, it must still add up to zero after S_1 injects charge because no *resistive* path is connected to X. The same holds true after C_H is placed around the op amp.

Now consider the circuit of Fig. 13.33(c), assuming that the total charge at node X is zero. We can write $C_X V_X - (V_{out} - V_X)C_H = 0$, and $V_X = -V_{out}/A_{v1}$. Thus, $-(C_X + C_H)V_{out}/A_{v1} - V_{out}C_H = 0$, i.e., $V_{out} = 0$. Note that this result is independent of Δq_1, capacitor values, or the gain of the op amp, thereby revealing that the charge injection by S_1 introduces no error *if S_2 turns off first.*

In summary, in Fig. 13.31(a), after S_2 turns off, node X "floats," maintaining a constant total charge regardless of the transitions at other nodes of the circuit. As a result, after the feedback configuration is formed, the output voltage is not influenced by the charge injection due to S_1. From another point of view, node X is a virtual ground at the moment S_2 turns off, freezing the instantaneous input level across C_H and yielding a charge equal to $V_0 C_H$ on the left plate of C_H. After settling with feedback, node X is again a virtual ground, forcing C_H to still carry $V_0 C_H$ and hence the output voltage to be approximately equal to V_0.

The effect of the charge injected by S_1 can be studied from yet another perspective. Suppose that in Fig. 13.33(c), the output voltage is finite and positive. Then, since $V_X = V_{out}/(-A_{v1})$, V_X must be finite and negative, requiring negative charge on the top plate of C_X. For the total charge at X to be zero, the charge on the left plate of C_H must be positive and that on its right plate negative, giving $V_{out} \leq 0$. Thus, the only valid solution is $V_{out} = 0$.

The third switch in Fig. 13.31(a), S_3, also merits attention. In order to turn on, S_3 must establish an inversion layer at its oxide interface. Does the required channel charge come from C_H or from the op amp? We note from the foregoing analysis that after the feedback circuit has settled, the charge on C_H equals $V_0 C_H$, unaffected by S_3. The channel charge of this switch is therefore entirely supplied by the op amp, introducing no error.

Our study of Fig. 13.31(a) thus far suggests that, with proper timing, the charge injected by S_1 and S_3 is unimportant and the channel charge of S_2 results in a constant offset voltage. Figure 13.34 depicts a simple realization of the clock edges to ensure that S_1 turns off after S_2 does.

The input-independent nature of the charge injected by the reset switch allows complete cancellation by differential operation. Illustrated in Fig. 13.35, such an approach employs a differential op amp along with two sampling capacitors so that the charge injected by S_2 and S_2' appears as a *common-mode* disturbance at nodes X and Y. This is in contrast to the behavior of the differential circuit shown in Fig. 13.28, where the input-dependent charge injection still leads to nonlinearity. In reality, S_2 and S_2' exhibit a finite charge injection mismatch, an issue resolved by adding another switch, S_{eq}, that turns off slightly after S_2 and S_2' (and before S_1 and S_1'), thereby equalizing the charge at nodes X and Y.

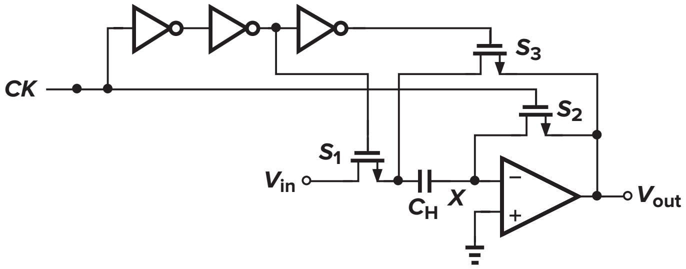


Figure 13.34 Generation of proper clock edges for unity-gain sampler.

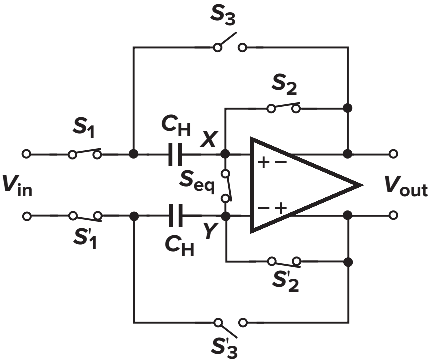


Figure 13.35 Differential realization of unity-gain sampler.

Precision Considerations The circuit of Fig. 13.31(a) operates as a unity-gain buffer in the amplification mode, producing an output voltage approximately equal to the voltage stored across the capacitor. How close to unity is the gain here? As a general case, we assume that the op amp exhibits a finite input capacitance C_{in} and calculate the output voltage when the circuit goes from the sampling mode to the amplification mode (Fig. 13.36). Owing to the finite gain of the op amp, $V_X \neq 0$ in the amplification mode, giving a charge equal to $C_{in}V_X$ on C_{in}. The conservation of charge at X requires that $C_{in}V_X$ come from C_H, raising the charge on C_H to $C_H V_0 + C_{in}V_X$.[7] It follows that the voltage across C_H equals $(C_H V_0 + C_{in}V_X)/C_H$. We therefore write $V_{out} - (C_H V_0 + C_{in}V_X)/C_H = V_X$ and $V_X = -V_{out}/A_{v1}$. Thus,

$$V_{out} = \frac{V_0}{1 + \frac{1}{A_{v1}}\left(\frac{C_{in}}{C_H} + 1\right)} \tag{13.40}$$

$$\approx V_0\left[1 - \frac{1}{A_{v1}}\left(\frac{C_{in}}{C_H} + 1\right)\right] \tag{13.41}$$

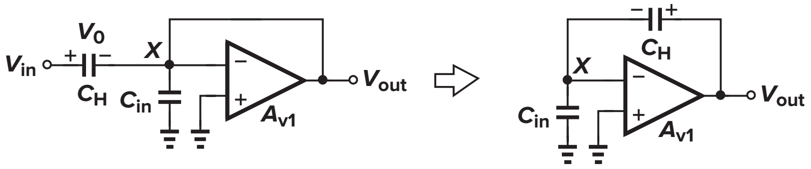


Figure 13.36 Equivalent circuit for accuracy calculations.

[7] The charge on C_H *increases* because positive charge transfer from the left plate of C_H to the top plate of C_{in} leads to a more positive voltage across C_H.

As expected, if $C_{in}/C_H \ll 1$, then $V_{out} \approx V_0/(1 + A_{v1}^{-1})$. In general, however, the circuit suffers from a gain error of approximately $-(C_{in}/C_H + 1)/A_{v1}$, suggesting that the input capacitance must be minimized even if speed is not critical. Recall from Chapter 9 that to increase A_{v1}, we may choose a large width for the input transistors of the op amp, but at the cost of higher input capacitance. An optimum device size must therefore yield minimum gain error rather than maximum A_{v1}.

▶ **Example 13.4**

In the circuit of Fig. 13.36, $C_{in} = 0.5$ pF and $C_H = 2$ pF. What is the minimum op amp gain that guarantees a gain error of 0.1%?

Solution

Since $C_{in}/C_H = 0.25$, we have $A_{v1,min} = 1000 \times 1.25 = 1250$. ◀

Speed Considerations Let us first examine the circuit in the sampling mode [Fig. 13.37(a)]. What is the time constant in this phase? The total resistance in series with C_H is given by R_{on1} and the resistance between X and ground, R_X. Using the simple op amp model shown in Fig. 13.37(b), where R_0 denotes the open-loop output impedance of the op amp, we have

$$(I_X - G_m V_X)R_0 + I_X R_{on2} = V_X \tag{13.42}$$

that is

$$R_X = \frac{R_0 + R_{on2}}{1 + G_m R_0} \tag{13.43}$$

Since typically $R_{on2} \ll R_0$ and $G_m R_0 \gg 1$, we have $R_X \approx 1/G_m$. For example, in a telescopic op amp employing differential to single-ended conversion, G_m equals the transconductance of each input transistor.

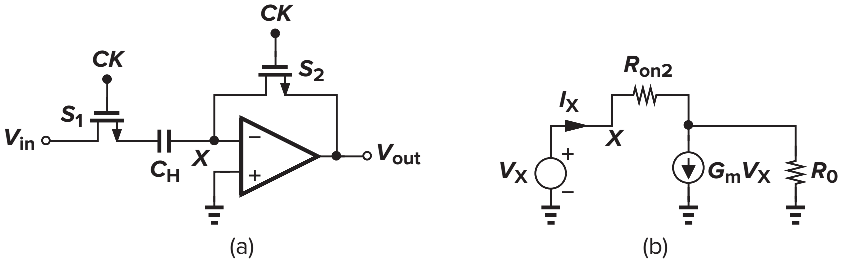


Figure 13.37 (a) Unity-gain sampler in sampling mode; (b) equivalent circuit of (a).

The time constant in the sampling mode is thus equal to

$$\tau_{sam} = \left(R_{on1} + \frac{1}{G_m}\right) C_H \tag{13.44}$$

The magnitude of τ_{sam} must be sufficiently small to allow settling in the test case of Fig. 13.15 to the required precision.

Now let us consider the circuit as it enters the amplification mode. Shown in Fig. 13.38 along with both the op amp input capacitance and the load capacitance, the circuit must begin with $V_{out} \approx 0$ and eventually produce $V_{out} \approx V_0$. If C_{in} is relatively small, we can assume that the voltages across C_L and

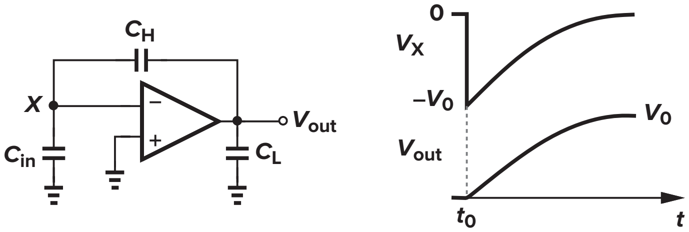


Figure 13.38 Time response of unity-gain sampler in amplification mode.

C_H do not change instantaneously, concluding that if $V_{out} \approx 0$ and $V_{CH} \approx V_0$, then $V_X = -V_0$ at the beginning of the amplification mode. In other words, the input difference sensed by the op amp initially jumps to a large value, possibly causing the op amp to slew. But, let us first assume that the op amp can be modeled by a linear model and determine the output response.

To simplify the analysis, we represent the charge on C_H by an explicit series voltage source, V_S, that goes from zero to V_0 at $t = t_0$ while C_H carries no charge itself (Fig. 13.39). The objective is to obtain the transfer function $V_{out}(s)/V_S(s)$ and hence the step response. We have

$$V_{out}\left(\frac{1}{R_0} + C_L s\right) + G_m V_X = (V_S + V_X - V_{out})C_H s \tag{13.45}$$

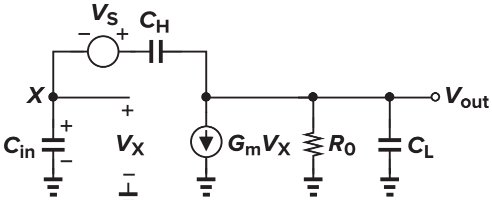


Figure 13.39 Equivalent circuit of unity-gain circuit in amplification mode.

Also, since the current through C_{in} equals $V_X C_{in} s$,

$$V_X \frac{C_{in}s}{C_H s} + V_X + V_S = V_{out} \tag{13.46}$$

Calculating V_X from (13.46) and substituting in (13.45), we arrive at the transfer function:

$$\frac{V_{out}}{V_S}(s) = R_0 \frac{(G_m + C_{in}s)C_H}{R_0(C_L C_{in} + C_{in}C_H + C_H C_L)s + G_m R_0 C_H + C_H + C_{in}} \tag{13.47}$$

Note that for $s = 0$, (13.47) reduces to a form similar to (13.40). Since typically $G_m R_0 C_H \gg C_H + C_{in}$, we can simplify (13.47) as

$$\frac{V_{out}}{V_S}(s) = \frac{(G_m + C_{in}s)C_H}{(C_L C_{in} + C_{in}C_H + C_H C_L)s + G_m C_H} \tag{13.48}$$

Thus, the response is characterized by a time constant equal to

$$\tau_{amp} = \frac{C_L C_{in} + C_{in}C_H + C_H C_L}{G_m C_H} \tag{13.49}$$

$$= \frac{1}{G_m}\left[C_{in} + \left(1 + \frac{C_{in}}{C_H}\right)C_L\right] \tag{13.50}$$

which is independent of the op amp output resistance. This is because a higher R_0 leads to a greater loop gain, eventually yielding a constant closed-loop speed. Another interesting interpretation of this result is described later (Fig. 13.52).

▶ **Example 13.5**

Consider the special cases $C_L = 0$ and $C_{in} = 0$ and explain the results intuitively.

Solution

If $C_L = 0$, then $\tau = C_{in}/G_m$. This occurs because the equivalent resistance seen by C_{in} is simply equal to $1/G_m$ if $C_L = 0$ [Fig. 13.40(a)].

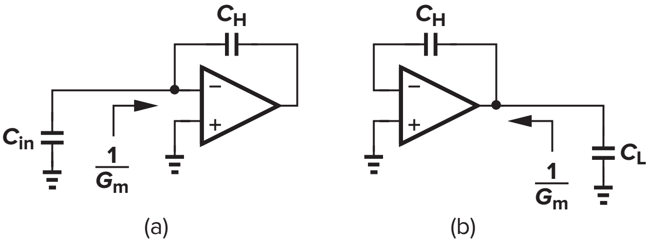


Figure 13.40

If $C_{in} = 0$, we have $\tau = C_L/G_m$ because C_L now sees a driving resistance equal to $1/G_m$ [Fig. 13.40(b)]. ◀

We now study the slewing behavior of the circuit, considering a telescopic op amp as an example. Upon entering the amplification mode, the circuit may experience a large step at the inverting input (Fig. 13.38). As shown in Fig. 13.41, the tail current of the op amp's input differential pair is then steered to one side, and its mirror current charges the capacitance seen at the output. Since M_2 is off during slewing, C_{in} is negligible and the slew rate is approximately equal to I_{SS}/C_L. The slewing continues until V_X is sufficiently close to the gate voltage of M_1, after which point the settling progresses with the time constant given in (13.50).

Our foregoing studies reveal that the input capacitance of the op amp degrades both the speed and the precision of the unity-gain sampler/buffer. For this reason, the bottom plate of C_H in Fig. 13.31 is usually driven by the input signal or the output of the op amp, and the top plate is connected to node X

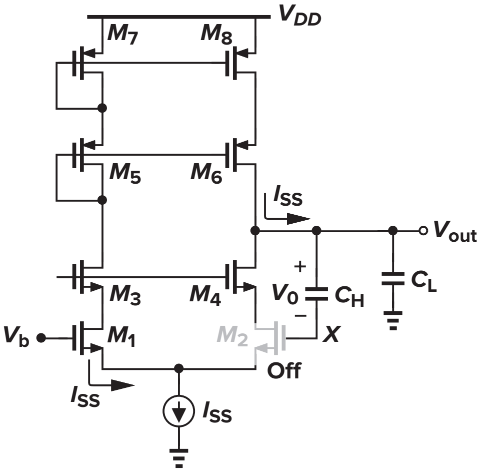


Figure 13.41 Unity-gain sampler during slewing.

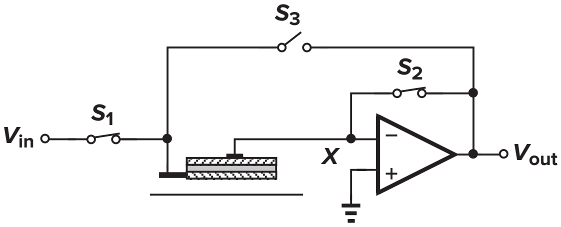


Figure 13.42 Connection of capacitor to the unity-gain sampler.

(Fig. 13.42), minimizing the parasitic capacitance seen from node X to ground. This technique is called "bottom-plate sampling." Driving the bottom plate by the input or the output also avoids the injection of substrate noise to node X (Chapter 19).

It is instructive to compare the performance of the sampling circuits shown in Figs. 13.30(b) and 13.31(a). In Fig. 13.30(b), the sampling time constant is smaller because it depends on only the on-resistance of the switch. More important, in Fig. 13.30(b), the amplification after the switch turns off is almost instantaneous, whereas in Fig. 13.31, it requires a finite settling time. However, the critical advantage of the unity-gain sampler is the input-independent charge injection.

13.3.2 Noninverting Amplifier

In this section, we revisit the amplifier of Fig. 13.5, studying its speed and precision properties. Repeated in Fig. 13.43(a), the amplifier operates as follows. In the sampling mode, S_1 and S_2 are on and S_3 is off, creating a virtual ground at X and allowing the voltage across C_1 to track the input voltage [Fig. 13.43(b)]. At the end of the sampling mode, S_2 turns off first, injecting a constant charge, Δq_2, onto node X. Subsequently, S_1 turns off and S_3 turns on [Fig. 13.43(c)]. Since V_P goes from V_{in0} to 0, the output voltage changes from 0 to approximately $V_{in0}(C_1/C_2)$, providing a voltage gain equal to C_1/C_2. We call the circuit a "noninverting amplifier" because the final output has the same polarity as V_{in0} and the gain can be greater than unity.

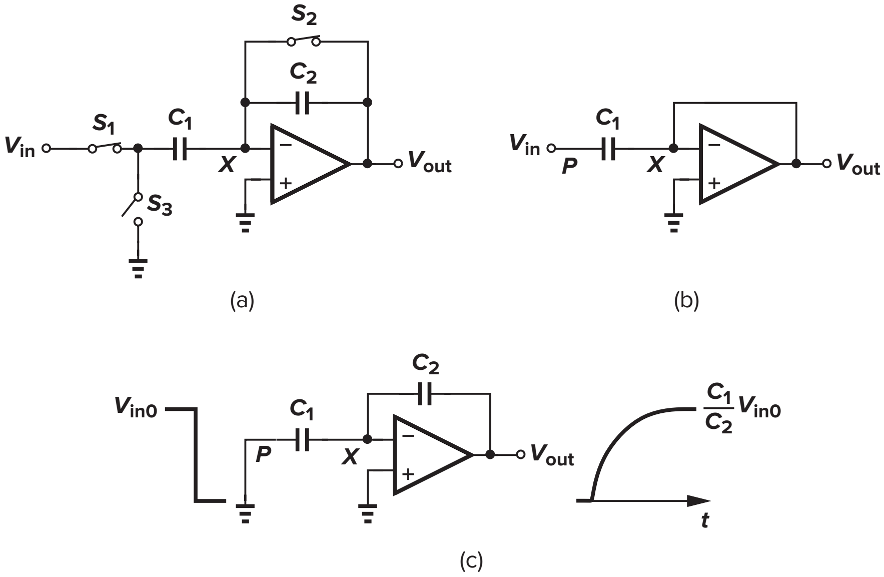


Figure 13.43 (a) Noninverting amplifier; (b) circuit of (a) in sampling mode; (c) transition of circuit to amplification mode.

As with the unity-gain circuit of Fig. 13.31(a), the noninverting amplifier avoids input-dependent charge injection by proper timing, namely, turning S_2 off before S_1 (Fig. 13.44). After S_2 is off, the total charge at node X remains constant, making the circuit insensitive to charge injection of S_1 or charge “absorption” of S_3. Let us first study the effect of S_1 carefully. As illustrated in Fig. 13.45, the charge injected by S_1, Δq_1, changes the voltage at node P by approximately $\Delta V_P = \Delta q_1/C_1$, and hence the output voltage by $-\Delta q_1 C_1/C_2$. However, after S_3 turns on, V_P drops to zero. Thus, the *overall* change in V_P is equal to $0 - V_{in0} = -V_{in0}$, producing an overall change in the output equal to $-V_{in0}(-C_1/C_2) = V_{in0}C_1/C_2$.

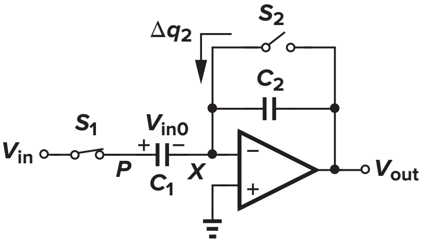


Figure 13.44 Transition of noninverting amplifier to amplification mode.

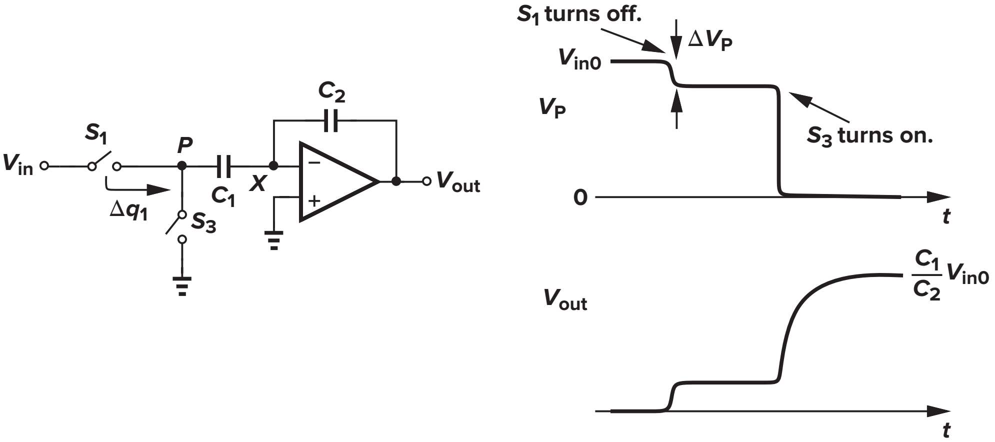


Figure 13.45 Effect of charge injected by S_1.

The key point here is that V_P goes from one fixed voltage, V_0, to another, 0, with an intermediate perturbation due to S_1. Since the output voltage of interest is measured after node P is connected to ground, the charge injected by S_1 does not affect the final output. From another perspective, as shown in Fig. 13.46, the charge on the right plate of C_1 at the instant S_2 turns off is approximately equal

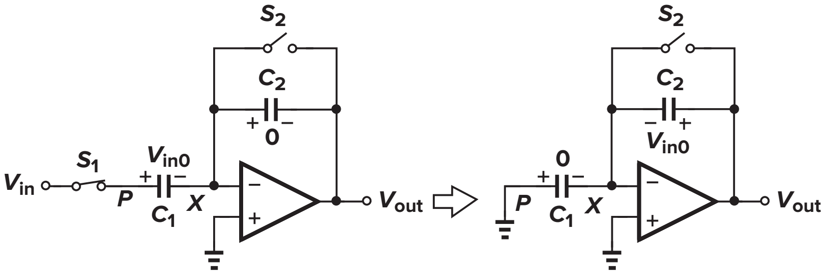


Figure 13.46 Charge redistribution in noninverting amplifier.

to $-V_{in0}C_1$. Also, the total charge at node X must remain constant after S_2 turns off. Thus, when node P is connected to ground and the circuit settles, the voltage across C_1, and hence its charge, are nearly zero, and the charge $-V_{in0}C_1$ must reside on the left plate of C_2. In other words, the output voltage is approximately equal to $V_{in0}C_1/C_2$ regardless of the intermediate excursions at node P.

The foregoing discussion indicates that two other phenomena have no effect on the final output. First, from the time S_2 turns off until the time S_1 turns off, the input voltage may change significantly (Fig. 13.47) without introducing any error. In other words, the sampling instant is defined by the turn-off of S_2. Second, when S_3 turns on, it requires some channel charge, but since the final value of V_P is zero, this charge is unimportant. Neither of these effects introduces error because the total charge at node X is conserved and V_P is eventually set by a fixed (zero) potential. To emphasize that V_P is initially and finally determined by fixed voltages, we say that node P is "driven" or node P switches from a low-impedance node to another low-impedance node. Here the term low-impedance distinguishes node P, at which charge is not conserved, from "floating" nodes such as X, where charge is conserved.

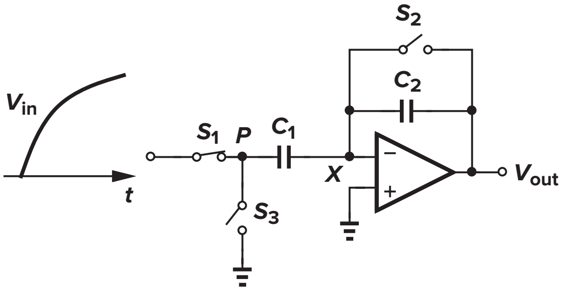


Figure 13.47 Effect of input change after S_2 turns off.

In summary, proper timing in Fig. 13.43(a) ensures that node X is perturbed only by the charge injection of S_2, making the final value of V_{out} free from errors due to S_1 and S_3. The constant offset due to S_2 can be suppressed by differential operation (Fig. 13.48).

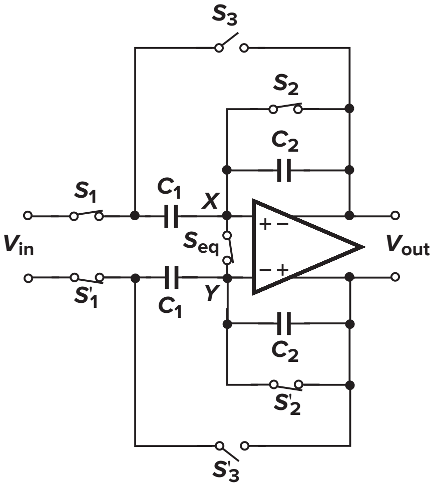


Figure 13.48 Differential realization of noninverting amplifier.

▶ **Example 13.6**

In the differential circuit of Fig. 13.48, suppose the equalizing switch is not used and S_2 and S_2' exhibit a threshold voltage mismatch of 10 mV. If $C_1 = 1$ pF, $C_2 = 0.5$ pF, $V_{TH} = 0.6$ V, and for all switches $WLC_{ox} = 50$ fF, calculate the dc offset measured at the output assuming that all of the channel charge of S_2 and S_2' is injected onto X and Y, respectively.

Solution

Simplifying the circuit as in Fig. 13.49, we have $V_{out} \approx \Delta q/C_2$, where $\Delta q = WLC_{ox}\Delta V_{TH}$. Note that C_1 does not appear in the result because X is a virtual ground, i.e., the voltage across C_1 changes only negligibly. Thus, the injected charge resides primarily on the left plate of C_2, giving an output error voltage equal to $\Delta V_{out} = WLC_{ox}\Delta V_{TH}/C_2 = 1$ mV.

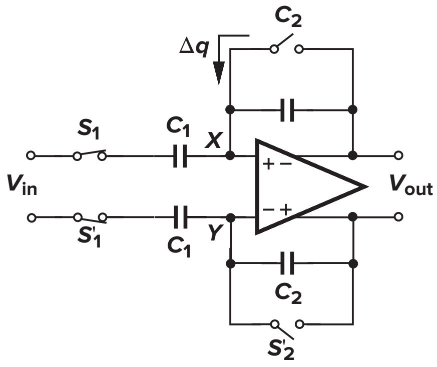


Figure 13.49

Precision Considerations As mentioned above, the circuit of Fig. 13.43(a) provides a nominal voltage gain of C_1/C_2. We now calculate the actual gain if the op amp exhibits a finite open-loop gain equal to A_{v1}. Depicted in Fig. 13.50 along with the input capacitance of the op amp, the circuit amplifies the input voltage change such that

$$(V_{out} - V_X)C_2 s = V_X C_{in} s + (V_X - V_{in})C_1 s \tag{13.51}$$

Since $V_{out} = -A_{v1}V_X$, we have

$$\left|\frac{V_{out}}{V_{in}}\right| = \frac{C_1}{C_2 + \dfrac{C_2 + C_1 + C_{in}}{A_{v1}}} \tag{13.52}$$

For large A_{v1},

$$\left|\frac{V_{out}}{V_{in}}\right| \approx \frac{C_1}{C_2}\left(1 - \frac{C_2 + C_1 + C_{in}}{C_2}\cdot\frac{1}{A_{v1}}\right) \tag{13.53}$$

implying that the amplifier suffers from a gain error of $(C_2 + C_1 + C_{in})/(C_2 A_{v1})$. Note that the gain error increases with the nominal gain C_1/C_2.

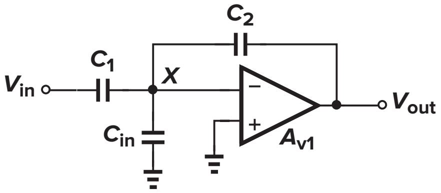


Figure 13.50 Equivalent circuit of noninverting amplifier during amplification.

Comparing (13.41) with (13.53), we note that with $C_H = C_2$ and for a nominal gain of unity, the noninverting amplifier exhibits greater gain error than does the unity-gain sampler. This is because the feedback factor equals $C_2/(C_1 + C_{in} + C_2)$ in the former and $C_H/(C_H + C_{in})$ in the latter. For example, if C_{in} is negligible, the unity-gain sampler's gain error is half that of the noninverting amplifier.

Speed Considerations The smaller feedback factor in Fig. 13.50 suggests that the time response of the amplifier may be slower than that of the unity-gain sampler. This is indeed true. Consider the equivalent circuit shown in Fig. 13.51(a). Since the only difference between this circuit and that in Fig. 13.39 is the capacitor C_1, which is connected from node X to an ideal voltage source, we expect that (13.50) gives the time constant of this amplifier as well if C_{in} is replaced by $C_{in} + C_1$. But for a more rigorous analysis, we substitute V_{in}, C_1, and C_{in} in Fig. 13.51(a) with a Thevenin equivalent as in Fig. 13.51(b), where $\alpha = C_1/(C_1 + C_{in})$, and $C_{eq} = C_1 + C_{in}$ and note that

$$V_X = (\alpha V_{in} - V_{out})\frac{C_{eq}}{C_{eq} + C_2} + V_{out} \tag{13.54}$$

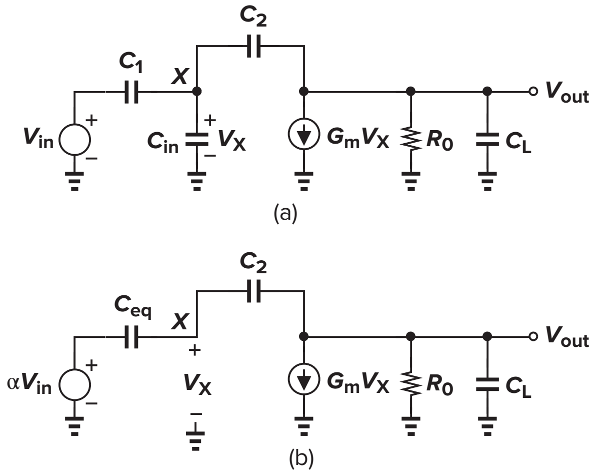


Figure 13.51 (a) Equivalent circuit of noninverting amplifier in amplification mode; (b) circuit of (a) with V_{in}, C_1, and C_{in} replaced by a Thevenin equivalent.

Thus,

$$\left[(\alpha V_{in} - V_{out})\frac{C_{eq}}{C_{eq} + C_2} + V_{out}\right] G_m + V_{out}\left(\frac{1}{R_0} + C_L s\right) = (\alpha V_{in} - V_{out})\frac{C_{eq}C_2}{C_{eq} + C_2}s \tag{13.55}$$

and hence

$$\frac{V_{out}}{V_{in}}(s) = \frac{-C_{eq}\dfrac{C_1}{C_1 + C_{in}}(G_m - C_2 s)R_0}{C_2 G_m R_0 + C_{eq} + C_2 + R_0[C_L(C_{eq} + C_2) + C_{eq}C_2]s} \tag{13.56}$$

Note that for $s = 0$, (13.56) reduces to (13.52). For a large $G_m R_0$, we can simplify (13.56) to

$$\frac{V_{out}}{V_{in}}(s) \approx \frac{-C_{eq}\dfrac{C_1}{C_1 + C_{in}}(G_m - C_2 s)R_0}{R_0(C_L C_{eq} + C_L C_2 + C_{eq}C_2)s + G_m R_0 C_2} \tag{13.57}$$

obtaining a time constant of

$$\tau_{amp} = \frac{C_L C_{eq} + C_L C_2 + C_{eq} C_2}{G_m C_2} \tag{13.58}$$

which is the same as the time constant of Fig. 13.38 if C_{in} is replaced by $C_{in} + C_1$. Note the direct dependence of τ_{amp} upon the nominal gain, C_1/C_2.

This expression can be rewritten as

$$\tau = \frac{C_1 + C_2 + C_{in}}{C_2} \cdot \frac{C_L + \dfrac{C_2(C_{in} + C_1)}{C_2 + C_{in} + C_1}}{G_m} \tag{13.59}$$

yielding interesting insights: the time constant is given by an equivalent capacitance, $C_L + C_2(C_{in} + C_1)/(C_2 + C_{in} + C_1)$, and an equivalent resistance, $(C_1 + C_2 + C_{in})/(G_m C_2)$ (Fig. 13.52). We can roughly say that the op amp sees the series combination of C_2 and $C_1 + C_{in}$ in parallel with C_L, and its G_m is reduced by the feedback factor, $C_2/(C_1 + C_2 + C_{in})$.

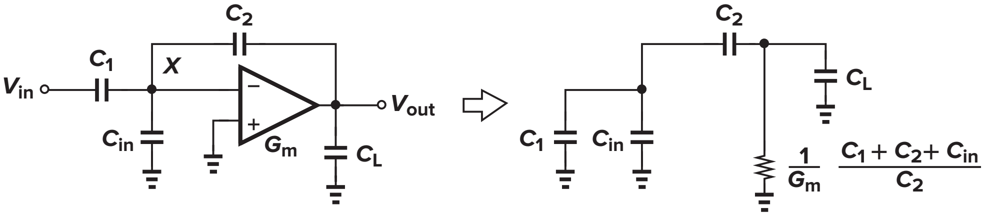


Figure 13.52 Equivalent circuit showing settling time constant.

It is instructive to examine the amplifier's time constant for the special case $C_L = 0$. Equation (13.58) yields $\tau_{amp} = (C_1 + C_{in})/G_m$, a value *independent* of the feedback capacitor. This is because, while a larger C_2 introduces heavier loading at the output, it also provides a greater feedback factor.

The reader may wonder why Eq. (13.56) yields a negative gain for the circuit that we have called a "noninverting" amplifier. This equation simply means that if the left plate of C_1 is stepped *down*, then the output goes *up*. This does not contradict the operation of the original circuit (Fig. 13.43), where the *change* in V_P is equal to $-V_{in}$.

13.3.3 Precision Multiply-by-Two Circuit

The circuit of Fig. 13.43(a) can operate with a relatively high closed-loop gain, but it suffers from speed and precision degradation due to the low feedback factor. In this section, we study a topology that provides a nominal gain of two while achieving a higher speed and lower gain error [5]. Shown in Fig. 13.53(a), the amplifier incorporates two equal capacitors, $C_1 = C_2 = C$. In the sampling mode, the circuit is configured as in Fig. 13.53(b), establishing a virtual ground at X and allowing the voltage across C_1 and C_2 to track V_{in}. In the transition to the amplification mode, S_3 turns off first, C_1 is placed around the op amp, and the left plate of C_2 is switched to ground [Fig. 13.53(c)]. Since at the moment S_3 turns off, the total charge on C_1 and C_2 equals $2V_{in0}C$ (if the charge injected by S_3 is neglected), and since the voltage across C_2 approaches zero in the amplification mode, the final voltage across C_1 and hence the output voltage are approximately equal to $2V_{in0}$. This can also be seen from the slow-motion illustration of Fig. 13.54.

The reader can show that the charge injected by S_1 and S_2 and absorbed by S_4 and S_5 is unimportant, and that injected by S_3 introduces a constant offset. The offset can be suppressed by differential operation.

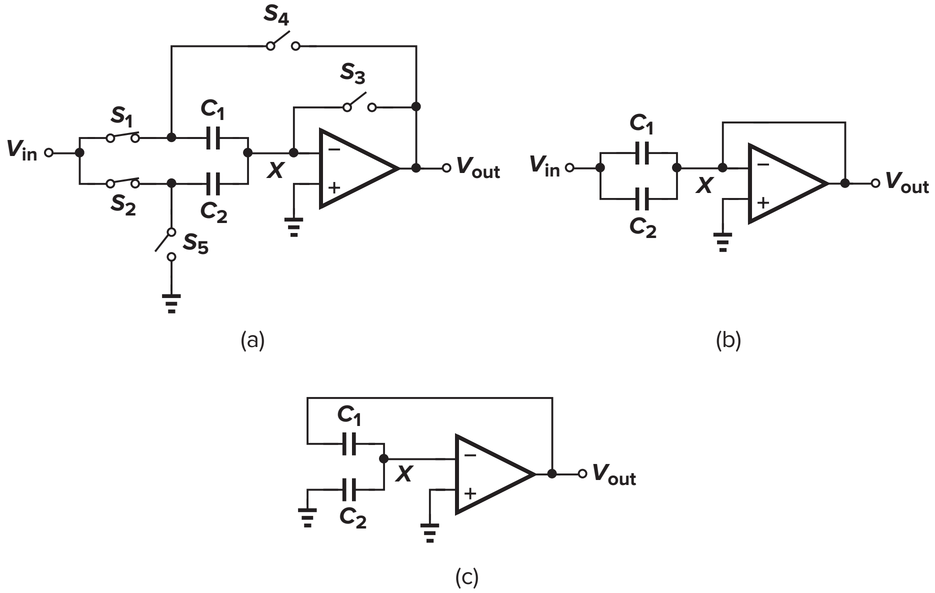


Figure 13.53 (a) Multiply-by-two circuit; (b) circuit of (a) in sampling mode; (c) circuit of (a) in amplification mode.

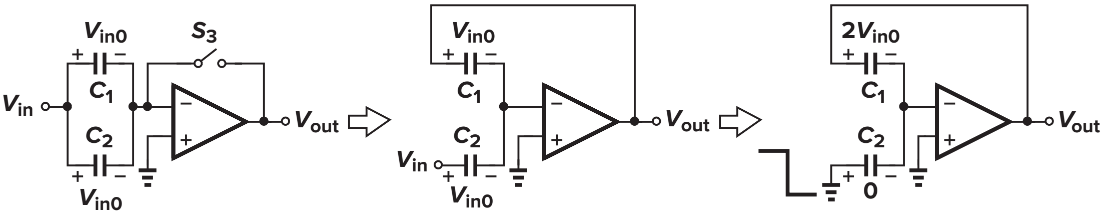


Figure 13.54 Transition of multiply-by-two circuit to amplification mode in slow motion.

The speed and precision of the multiply-by-two circuit are expressed by (13.58) and (13.53), respectively, but the advantage of the circuit is the higher feedback factor for a given closed-loop gain. Note, however, that the input capacitance of the multiply-by-two circuit in the sampling mode is higher.

13.4 ■ Switched-Capacitor Integrator

Integrators are used in many analog systems. Examples include filters and oversampled analog-to-digital converters. Figure 13.55 depicts a continuous-time integrator, whose output can be expressed as

$$V_{out} = -\frac{1}{RC_F}\int V_{in}dt \tag{13.60}$$

if the op amp gain is very large. For sampled-data systems, we must devise a discrete-time counterpart of this circuit.

Before studying SC integrators, let us first point out an interesting property. Consider a resistor connected between two nodes [Fig. 13.56(a)], carrying a current equal to $(V_A - V_B)/R$. The role of the resistor is to take a certain amount of charge from node A every second and move it to node B. Can we

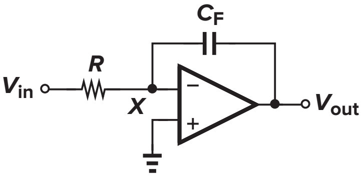


Figure 13.55 Continuous-time integrator.

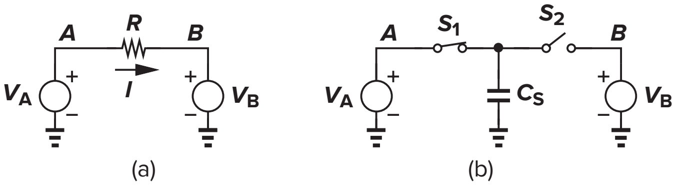


Figure 13.56 (a) Continuous-time and (b) discrete-time resistors.

perform the same function with a capacitor? Suppose that in the circuit of Fig. 13.56(b), capacitor C_S is alternately connected to nodes A and B at a clock rate f_{CK}. The *average* current flowing from A to B is then equal to the charge moved in one clock period:

$$\overline{I_{AB}} = \frac{C_S(V_A - V_B)}{f_{CK}^{-1}} \tag{13.61}$$

$$= C_S f_{CK}(V_A - V_B) \tag{13.62}$$

We can therefore view the circuit as a "resistor" equal to $(C_S f_{CK})^{-1}$. Recognized by James Clark Maxwell, this property formed the foundation for many modern switched-capacitor circuits.

Let us now replace resistor R in Fig. 13.55 by its discrete-time equivalent, arriving at the integrator of Fig. 13.57(a). We note that in every clock cycle, C_1 absorbs a charge equal to C_1V_{in} when S_1 is on and deposits the charge on C_2 when S_2 is on (node X is a virtual ground). For example, if V_{in} is constant, the output changes by $V_{in}C_1/C_2$ every clock cycle [Fig. 13.57(b)]. Approximating the staircase waveform by a ramp, we note that the circuit behaves as an integrator.

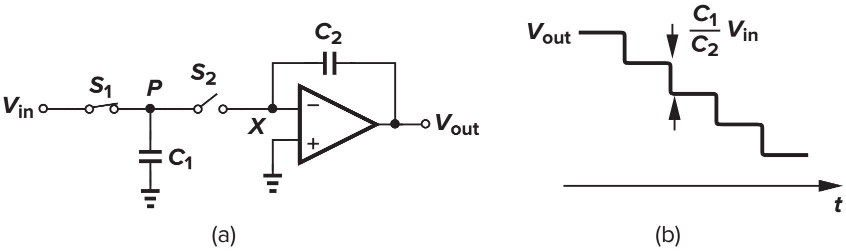


Figure 13.57 (a) Discrete-time integrator; (b) response of circuit to a constant input voltage.

The final value of V_{out} in Fig. 13.57(a) after every clock cycle can be written as

$$V_{out}(kT_{CK}) = V_{out}[(k-1)T_{CK}] - V_{in}[(k-1)T_{CK}] \cdot \frac{C_1}{C_2} \tag{13.63}$$

where the gain of the op amp is assumed large. Note that the small-signal settling time constant as charge is transferred from C_1 to C_2 is given by (13.50).

The integrator of Fig. 13.57(a) suffers from two important drawbacks. First, the input-dependent charge injection of S_1 introduces nonlinearity in the charge stored on C_1 and hence the output voltage. Second, the nonlinear capacitance at node P resulting from the source/drain junctions of S_1 and S_2 leads to a nonlinear charge-to-voltage conversion when C_1 is switched to X. This can be understood with the aid of Fig. 13.58, where the charge stored on the total junction capacitance, C_j, is *not* equal to $V_{in0}C_j$, but rather equal to

$$q_{cj} = \int_0^{Vin0} C_j dV \tag{13.64}$$

Since C_j is a function of voltage, q_{cj} exhibits a nonlinear dependence on V_{in0}, thereby creating a nonlinear component at the output after the charge is transferred to the integration capacitor.

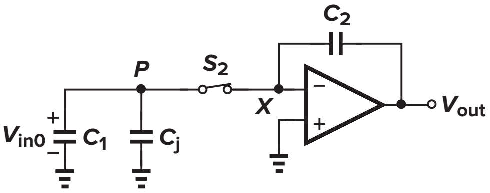


Figure 13.58 Effect of junction capacitance nonlinearity in SC integrator.

An integrator topology that resolves both of the foregoing issues is shown in Fig. 13.59(a). We study the circuit's operation in the sampling and integration modes. As shown in Fig. 13.59(b), in the sampling mode, S_1 and S_3 are on and S_2 and S_4 are off, allowing the voltage across C_1 to track V_{in} while the op amp and C_2 hold the previous value. In the transition to the integration mode, S_3 turns off first, injecting a constant charge onto C_1; S_1 turns off next; and subsequently S_2 and S_4 turn on [Fig. 13.59(c)]. The charge stored on C_1 is therefore transferred to C_2 through the virtual ground node.

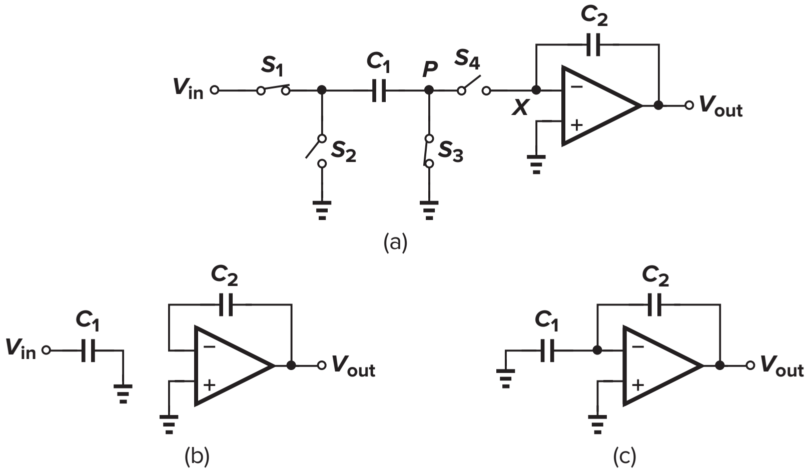


Figure 13.59 (a) Parasitic-insensitive integrator; (b) circuit of (a) in sampling mode; (c) circuit of (a) in integration mode.

Since S_3 turns off first, it introduces only a constant offset, which can be suppressed by differential operation. Moreover, because the left plate of C_1 is "driven" (Sec. 13.3.2), the charge injection or absorption of S_1 and S_2 contributes no error. Also, since node X is a virtual ground, the charge injected or absorbed by S_4 is constant and independent of V_{in}.

How about the nonlinear junction capacitance of S_3 and S_4? We observe that the voltage across this capacitance goes from near zero in the sampling mode to virtual ground in the integration mode. Since the voltage across the nonlinear capacitance changes by a very small amount, the resulting nonlinearity is negligible.

13.5 ■ Switched-Capacitor Common-Mode Feedback

Our study of common-mode feedback in Chapter 9 suggested that sensing the output CM level by means of resistors lowers the differential voltage gain of the circuit. We also observed that sensing techniques using MOSFETs that operate as source followers or variable resistors suffer from a limited linear range. Switched-capacitor CMFB networks provide an alternative that avoids both of these difficulties (but the circuit must be refreshed periodically).

In switched-capacitor common-mode feedback, the outputs are sensed by capacitors rather than resistors. Figure 13.60 depicts a simple example, where equal capacitors C_1 and C_2 reproduce at node X the average of the changes in each output voltage. Thus, if V_{out1} and V_{out2} experience, say, a positive CM change, then V_X and hence I_{D5} increase, pulling V_{out1} and V_{out2} down. The output CM level is then equal to V_{GS2} plus the voltage across C_1 and C_2.

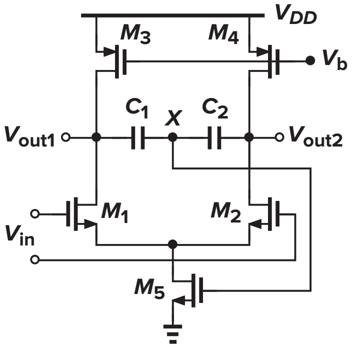


Figure 13.60 Simple SC common-mode feedback.

How is the voltage across C_1 and C_2 defined? This is typically carried out when the amplifier is in the sampling (or reset) mode and can be accomplished as shown in Fig. 13.61. Here, during CM level definition, the amplifier differential input is zero and switch S_1 is on. Transistors M_6 and M_7 operate as a linear sense circuit because their gate voltages are nominally equal. Thus, the circuit settles such that the ouput CM level is equal to $V_{GS6,7} + V_{GS5}$. At the end of this mode, S_1 turns off, leaving a voltage equal

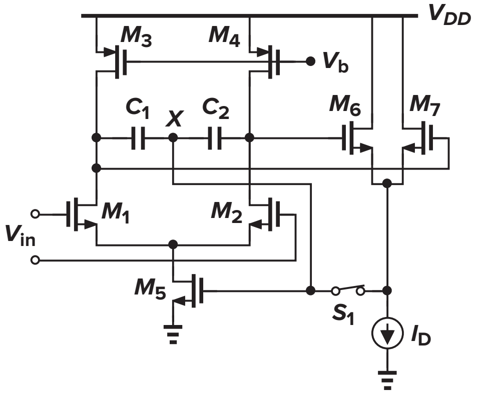


Figure 13.61 Definition of the voltage across C_1 and C_2.

to $V_{GS6,7}$ across C_1 and C_2. In the amplification mode, M_6 and M_7 may experience a large nonlinearity, but they do not affect the performance of the main circuit because S_1 is off.

In applications where the output CM level must be defined more accurately than in the previous example, the topology shown in Fig. 13.62 may be used. Here, in the reset mode, one plate of C_1 and C_2 is switched to V_{CM} while the other is connected to the gate of M_6. Each capacitor therefore sustains a voltage equal to $V_{CM} - V_{GS6}$. In the amplification mode, S_2 and S_3 are on and the other switches are off, yielding an output CM level equal to $V_{CM} - V_{GS6} + V_{GS5}$. This value is equal to V_{CM} if I_{D3} and I_{D4} are copied properly from I_{REF} so that $V_{GS6} = V_{GS5}$.

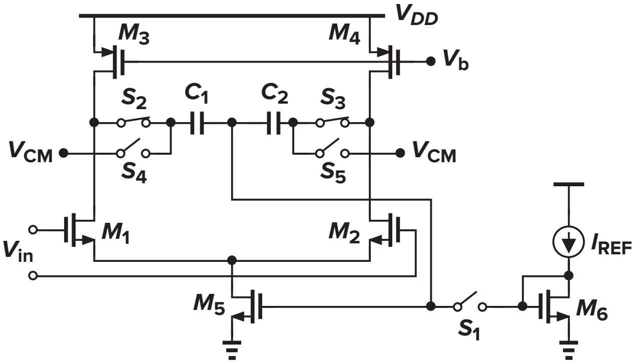


Figure 13.62 Alternative topology for definition of output CM level.

With large output swings, the speed of the CMFB loop may in fact influence the settling of the differential output [6]. For this reason, part of the tail current of the differential pairs in Figs. 13.61 and 13.62 can be provided by a *constant* current source so that M_5 makes only small adjustments to the circuit.

References

[1] G. Wegmann, E. A. Vittoz, and F. Rahali, "Charge Injection in Analog MOS Switches," *IEEE J. of Solid-State Circuits*, vol. SC-22, pp. 1091–1097, December 1987.

[2] B. J. Sheu and C. Hu, "Switch-Induced Error Voltage on a Switched Capacitor," *IEEE J. of Solid-State Circuits*, vol. SC-19, pp. 519–525, April 1984.

[3] R. Gregorian and G. C. Temes, *Analog MOS Integrated Circuits for Signal Processing* (New York: John Wiley and Sons, 1986).

[4] J. H. Fischer, "Noise Sources and Calculation Techniques for Switched Capacitor Filters," *IEEE J. of Solid-State Circuits*, vol. 17, pp. 742–752, August 1982.

[5] B. S. Song, M. F. Tompsett, and K. R. Lakshmikumar, "A 12-Bit 1-Msample/s Capacitor-Averaging Pipelined A/D Converter," *IEEE J. of Solid-State Circuits*, vol. SC-23, pp. 1324–1333, December 1988.

[6] B. Razavi, *Principles of Data Conversion System Design* (New York: IEEE Press, 1995).

[7] P. C. Yu and H.-S. Lee, "A High-Swing 2-V CMOS Op Amp with Replica-Amp Gain Enhancement," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 1265–1272, December 1993.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that transistors are in saturation.

13.1. The circuit of Fig. 13.2(b) is designed with $C_1 = 2$ pF and $C_2 = 0.5$ pF.

13.2. Assuming that $R_F = \infty$, but the op amp has an output resistance R_{out}, derive the transfer function $V_{out}(s)/V_{in}(s)$.

13.3. If the op amp is ideal, determine the minimum value of R_F that guarantees a gain error of 1% for an input frequency of 1 MHz.

13.4. Suppose that in Fig. 13.6(a), the op amp is characterized by a transconductance G_m and an output resistance R_{out}.

13.5. Determine the transfer function V_{out}/V_{in} in this mode.

13.6. Plot the waveform at node B if V_{in} is a 100-MHz sinusoid with a peak amplitude of 1 V, $C_1 = 1$ pF, $G_m = 1/(100\ \Omega)$, and $R_{out} = 20\ \text{k}\Omega$.

13.7. In Fig. 13.6(b), node A is in fact connected to ground through a switch (Fig. 13.5). If the switch introduces a series resistance R_{on} and the op amp is ideal, calculate the time constant of the circuit in this mode. What is the total energy dissipated in the switch as the circuit enters the amplification mode and V_{out} settles to its final value?

13.8. The circuit of Fig. 13.10(a) is designed with $(W/L)_1 = 20/0.5$ and $C_H = 1$ pF.

13.9. Using Eqs. (13.9) and (13.16), calculate the time required for V_{out} to drop to +1 mV.

13.10. Approximating M_1 by a linear resistor equal to $[\mu_n C_{ox}(W/L)_1(V_{DD} - V_{TH})]^{-1}$, calculate the time required for V_{out} to drop to +1 mV and compare the result with that obtained in part (a).

13.11. The circuit of Fig. 13.12 cannot be characterized by a single time constant because the resistance charging C_H (equal to $1/g_{m1}$ if $\gamma = 0$) varies with the output level. Assume that $(W/L)_1 = 20/0.5$ and $C_H = 1$ pF.

13.12. Using Eq. (13.21), calculate the time required for V_{out} to reach 2.1 V.

13.13. Sketch the transconductance of M_1 versus time.

13.14. In the circuit of Fig. 12.8(b), $(W/L)_1 = 20/0.5$ and $C_H = 1$ pF. Assume that $\lambda = \gamma = 0$ and $V_{in} = V_0 \sin \omega_{in} t + V_m$, where $\omega_{in} = 2\pi \times (100\ \text{MHz})$.

13.15. Calculate R_{on1} and the phase shift from the input to the output if $V_0 = V_m = 10$ mV.

13.16. Repeat part (a) if $V_0 = 10$ mV but $V_m = 1$ V. The variation of the phase shift translates to distortion.

13.17. Describe an efficient SPICE simulation that yields the plot of $R_{on,eq}$ for the circuit of Fig. 13.17.

13.18. The sampling network of Fig. 13.17 is designed with $(W/L)_1 = 20/0.5$, $(W/L)_2 = 60/0.5$, and $C_H = 1$ pF. If $V_{in} = 0$ and the initial value of V_{out} is +3 V, estimate the time required for V_{out} to drop to +1 mV.

13.19. In the circuit of Fig. 13.20, $(W/L)_1 = 20/0.5$ and $C_H = 1$ pF. Calculate the maximum error at the output due to charge injection. Compare this error with that resulting from clock feedthrough.

13.20. The circuit of Fig. 13.63 samples the input on C_1 when CK is high and connects C_1 and C_2 when CK is low. Assume that $(W/L)_1 = (W/L)_2$ and $C_1 = C_2$.

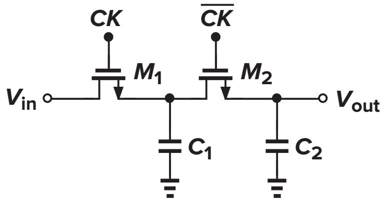


Figure 13.63

13.21. If the initial voltages across C_1 and C_2 are zero and $V_{in} = 2$ V, plot V_{out} versus time for many clock cycles. Neglect charge injection and clock feedthrough.

13.22. What is the maximum error in V_{out} due to charge injection and clock feedthrough of M_1 and M_2? Assume that the channel charge of M_2 splits equally between C_1 and C_2.

13.23. Determine the sampled kT/C noise at the output after M_2 turns off.

13.24. For $V_{in} = V_0 \sin \omega_0 t + V_0$, where $V_0 = 0.5$ V and $\omega_0 = 2\pi \times (10\ \text{MHz})$, plot the output waveforms of the circuits shown in Fig. 13.30(b) and 13.31(a). Assume a clock frequency of 50 MHz.

13.25. In Fig. 13.47, S_1 turns off Δt seconds after S_2, and S_3 turns on Δt seconds after S_1 turns off. Plot the output waveform, taking into account the charge injection and clock feedthough of S_1–S_3. Assume that all of the switches are NMOS devices.

13.26. The circuit of Fig. 13.50 is designed with $C_1 = 2$ pF, $C_{in} = 0.2$ pF, and $A_v = 1000$. What is the maximum nominal gain, C_1/C_2, that the circuit can provide with a gain error of 1%?

13.27. In Problem 13.26, what is the maximum nominal gain if $G_m = 1/(100\ \Omega)$ and the circuit must achieve a time constant of 2 ns in the amplification mode? Assume that $C_{in} = 0.2$ pF, and calculate C_1 and C_2.

13.28. The integrator of Fig. 13.57 is designed with $C_1 = C_2 = 1$ pF and a clock frequency of 100 MHz. Neglecting charge injection and clock feedthrough, sketch the output if the input is a 10-MHz sinusoid with a peak amplitude of 0.5 V. Approximating C_1, S_1, and S_2 by a resistor, estimate the output amplitude.

13.29. Consider the switched-capacitor amplifier depicted in Fig. 13.64, where the common-mode feedback is not shown. Assume that $(W/L)_{1-4} = 50/0.5$, $I_{SS} = 1$ mA, $C_1 = C_2 = 2$ pF, $C_3 = C_4 = 0.5$ pF, and the output CM level is 1.5 V. Neglect the transistor capacitances.

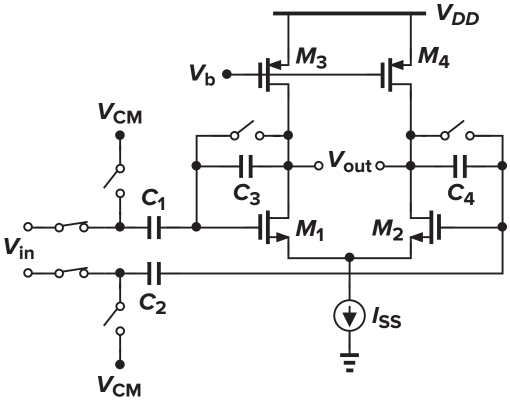


Figure 13.64

13.30. What is the maximum allowable output voltage swing in the amplification mode?

13.31. Determine the gain error of the amplifier.

13.32. What is the small-signal time constant in the amplification mode?

13.33. Repeat Problem 13.32 if the gate-source capacitance of M_1 and M_2 is not neglected.

13.34. A differential circuit incorporating a well-designed common-mode feedback network exhibits the open-loop input-output characteristic shown in Fig. 13.65(a). In some circuits, however, the characteristic appears as in Fig. 13.65(b). Explain how this effect occurs.

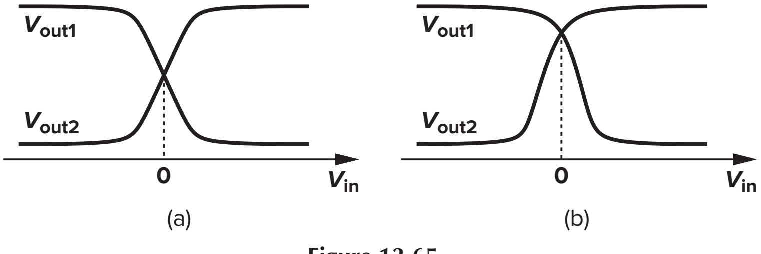


Figure 13.65

13.35. In the common-mode feedback network of Fig. 13.61, assume that $W/L = 50/0.5$ for all transistors, $I_{D5} = 1$ mA, and $I_{D6,7} = 50\ \mu$A. Determine the allowable range of the input common-mode level.

13.36. Repeat Problem 13.35 if $(W/L)_{6,7} = 10/0.5$.

13.37. Suppose that in the common-mode feedback network of Fig. 13.61, S_1 injects a charge of Δq onto the gate of M_5. How much do the gate voltage of M_5 and the output common-mode level change due to this error?

13.38. In the circuit of Fig. 13.66, each op amp is represented by a Norton equivalent and characterized by G_m and R_{out}. The output currents of two op amps are summed at node Y [7]. (The circuit is shown in the amplification mode.) Note that the main amplifier and the auxiliary amplifier are identical and that the error amplifier senses the voltage variation at node X and injects a proportional current into node Y. The output impedance of the error amplifier is much greater than R_{out}. Assume that $G_m R_{out} \gg 1$.

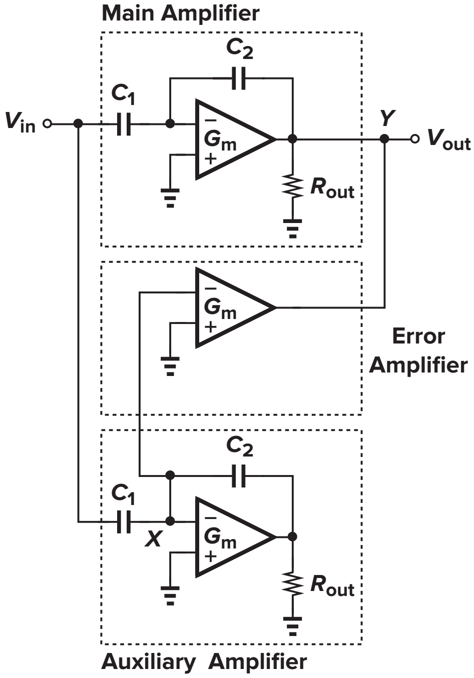


Figure 13.66

13.39. Calculate the gain error of the circuit.

13.40. Repeat part (a) if the auxiliary and error amplifiers are eliminated and compare the results.

CHAPTER

14

Nonlinearity and Mismatch

In Chapters 6 and 7, we dealt with two types of nonidealities, namely, frequency response and noise, that limit the performance of analog circuits. In this chapter, we study two other imperfections that prove critical in high-precision analog design and trade off with many other performance parameters. These effects are nonlinearity and mismatch.

We first define metrics for quantifying the effects of nonlinearity. Next, we study nonlinearity in differential circuits and feedback systems and examine several linearization techniques. We then deal with the problem of mismatch and dc offsets in differential circuits. Finally, we consider a number of offset cancellation methods and describe the effect of offset cancellation on random noise.

14.1 ■ Nonlinearity

14.1.1 General Considerations

As we have observed in the large-signal analysis of single-stage and differential amplifiers, circuits usually exhibit a nonlinear input/output characteristic. Depicted in Fig. 14.1, such a characteristic deviates from a straight line as the input swing increases. Two examples are shown in Fig. 14.2. In a common-source stage or a differential pair, the output variation becomes heavily nonlinear as the input level increases. In other words, for a small input swing, the output is a reasonable replica of the input, but for large swings the output exhibits "saturated" levels.

The nonlinear behavior of a circuit can also be viewed as *variation* of the slope, and hence the small-signal gain, with the input level. Illustrated in Fig. 14.3, this observation means that a given incremental change at the input results in different incremental changes at the output depending on the input dc level.

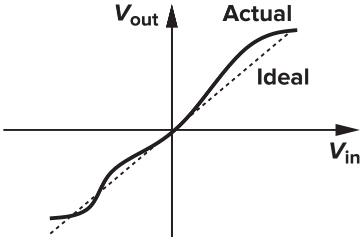


Figure 14.1 Input/output characteristic of a nonlinear system.

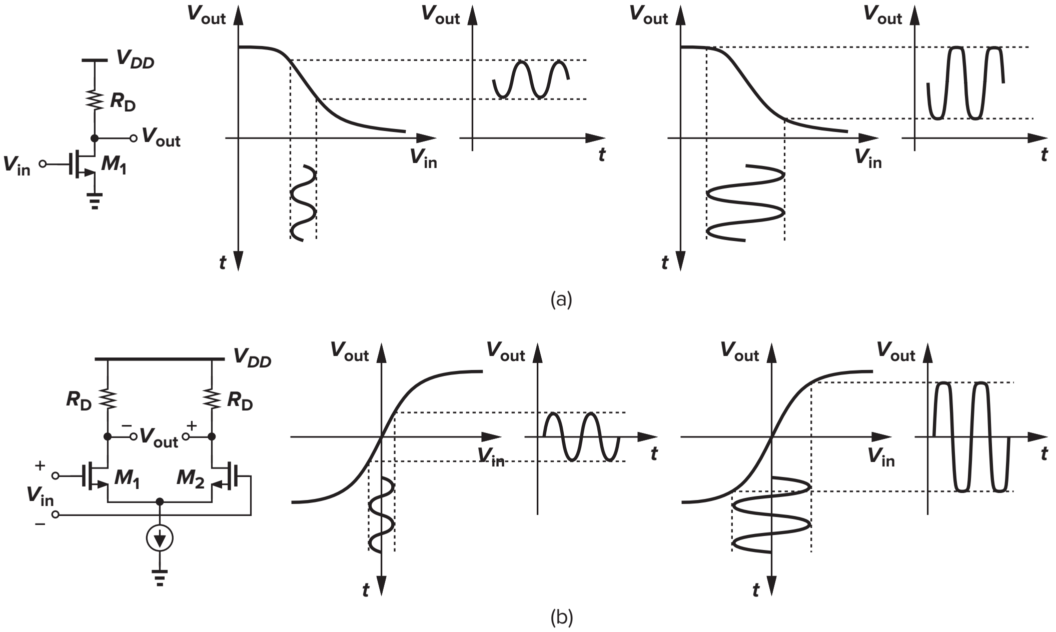


Figure 14.2 Distortion in (a) a common-source stage and (b) a differential pair.

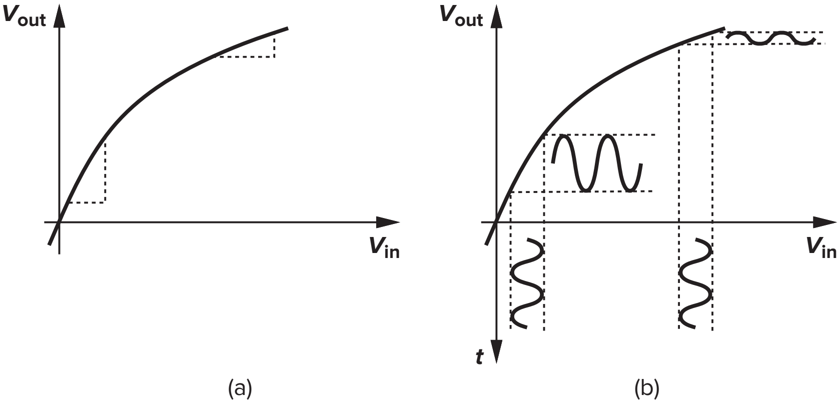


Figure 14.3 Variation of small-signal gain in a nonlinear amplifier.

In many analog circuits, precision requirements mandate relatively small nonlinearities, making it possible to approximate the input/output characteristic by a polynomial in the range of interest:

$$y(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) + \cdots \tag{14.1}$$

For small x, $y(t) \approx \alpha_1 x$, indicating that α_1 is the small-signal gain in the vicinity of $x \approx 0$.

How is the nonlinearity quantified? A simple method is to identify α_1, α_2, etc., in (14.1). Another metric that proves useful in practice is to specify the maximum deviation of the characteristic from an ideal one (i.e., a straight line). As shown in Fig. 14.4, for the voltage range of interest, $[0\ \ V_{in,max}]$, we pass a straight line through the end points of the actual characteristic, obtain the maximum deviation, ΔV, and normalize the result to the maximum output swing, $V_{out,max}$. For example, we say that an amplifier exhibits 1% nonlinearity ($\Delta V/V_{out,max} = 0.01$) for an input range of 1 V.

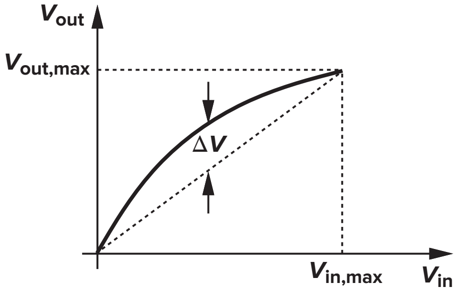


Figure 14.4 Definition of nonlinearity.

▶ Example 14.1

The input/output characteristic of a differential amplifier is approximated as $y(t) = \alpha_1 x(t) + \alpha_3 x^3(t)$. Calculate the maximum nonlinearity if the input range is from $x = -x_{max}$ to $x = +x_{max}$.

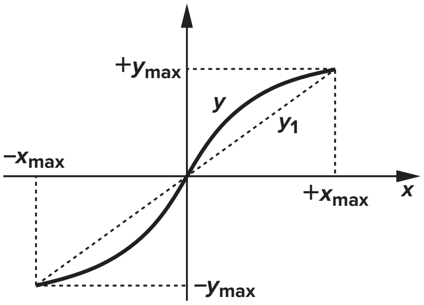


Figure 14.5

Solution

As depicted in Fig. 14.5, we can express the straight line passing through the end points as

$$y_1 = \frac{\alpha_1 x_{max} + \alpha_3 x_{max}^3}{x_{max}} x \tag{14.2}$$

$$= \left(\alpha_1 + \alpha_3 x_{max}^2\right) x \tag{14.3}$$

The difference between y and y_1 is therefore equal to

$$\Delta y = y - y_1 \tag{14.4}$$

$$= \alpha_1 x + \alpha_3 x^3 - \left(\alpha_1 + \alpha_3 x_{max}^2\right) x \tag{14.5}$$

Setting the derivative of Δy with respect to x to zero, we have $x = x_{max}/\sqrt{3}$, and the maximum deviation is equal to $2\alpha_3 x_{max}^3/(3\sqrt{3})$. Normalized to the maximum output, the nonlinearity is obtained as

$$\frac{\Delta y}{y_{max}} = \frac{2\alpha_3 x_{max}^3}{3\sqrt{3} \times 2\left(\alpha_1 x_{max} + \alpha_3 x_{max}^3\right)} \tag{14.6}$$

Note that the factor of 2 in the denominator is included because the maximum peak-to-peak output swing is equal to $2(\alpha_1 x_{max} + \alpha_3 x_{max}^3)$. For small nonlinearities, we can neglect $\alpha_3 x_{max}^3$ with respect to $\alpha_1 x_{max}$, arriving at

$$\frac{\Delta y}{y_{max}} \approx \frac{\alpha_3}{3\sqrt{3}\alpha_1} x_{max}^2 \tag{14.7}$$

Note that the relative nonlinearity is proportional to the square of the maximum input swing in this example. ◀

The nonlinearity of a circuit can also be characterized by applying a sinusoid at the input and measuring the harmonic content of the output. Specifically, if in (14.1), $x(t) = A\cos\omega t$, then

$$y(t) = \alpha_1 A\cos\omega t + \alpha_2 A^2\cos^2\omega t + \alpha_3\cos^3\omega t + \cdots \tag{14.8}$$

$$= \alpha_1 A\cos\omega t + \frac{\alpha_2 A^2}{2}[1+\cos(2\omega t)] + \frac{\alpha_3 A^3}{4}[3\cos\omega t + \cos(3\omega t)] + \cdots. \tag{14.9}$$

We observe that higher-order terms yield higher harmonics. In particular, even-order terms and odd-order terms result in even and odd harmonics, respectively. Note that the magnitude of the nth harmonic grows roughly in proportion to the nth power of the input amplitude. Called "harmonic distortion," this effect is usually quantified by summing the power of all of the harmonics (except that of the fundamental) and normalizing the result to the power of the fundamental. Such a metric is called the "total harmonic distortion" (THD). For a third-order nonlinearity,

$$\text{THD} = \frac{(\alpha_2 A^2/2)^2 + (\alpha_3 A^3/4)^2}{(\alpha_1 A + 3\alpha_3 A^3/4)^2} \tag{14.10}$$

Harmonic distortion is undesirable in most signal processing applications, including audio and video systems. High-quality audio products such as compact disc (CD) players require a THD of about 0.01% (−80 dB), and video products, about 0.1% (−60 dB).

14.1.2 Nonlinearity of Differential Circuits

Differential circuits exhibit an "odd-symmetric" input/output characteristic, i.e., $f(-x) = -f(x)$. For the polynomial of (14.1) to be an odd function, all of the even-order terms, α_{2j}, must be zero:

$$y(t) = \alpha_1 x(t) + \alpha_3 x^3(t) + \alpha_5 x^5(t) + \cdots \tag{14.11}$$

indicating that a differential circuit driven by a differential signal produces no even harmonics. This is another very important property of differential operation.

In order to appreciate the reduction of nonlinearity obtained by differential operation, let us consider the two amplifiers shown in Fig. 14.6, each of which is designed to provide a small-signal voltage gain of

$$|A_v| \approx g_m R_D \tag{14.12}$$

$$= \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})R_D \tag{14.13}$$

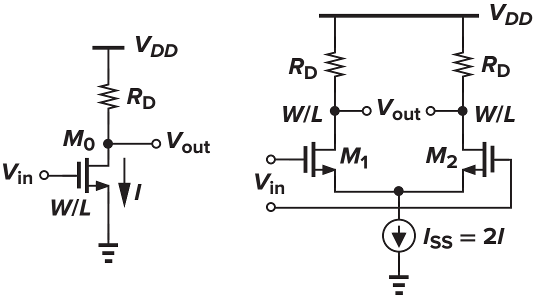


Figure 14.6 Single-ended and differential amplifiers providing the same voltage gain.

Suppose a signal $V_m \cos \omega t$ is applied to each circuit. Examining only the drain currents for simplicity, we can write for the common-source stage:

$$\begin{aligned}
I_{D0} &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH} + V_m\ cos\omega t)^2 \\
&= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 + \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})V_m \cos \omega t \\
&\quad + \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_m^2 \cos^2 \omega t \\
&= I + \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})V_m \cos \omega t + \frac{1}{4}\mu_n C_{ox}\frac{W}{L}V_m^2[1 + \cos(2\omega t)] \qquad (14.14)
\end{aligned}$$

Thus, the amplitude of the second harmonic, A_{HD2}, normalized to that of the fundamental, A_F, is

$$\frac{A_{HD2}}{A_F} = \frac{V_m}{4(V_{GS} - V_{TH})} \qquad (14.15)$$

On the other hand, for M_1 and M_2 in Fig. 14.6, we have from Chapter 4

$$\begin{aligned}
I_{D1} - I_{D2} &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{\frac{4I_{SS}}{\mu_n C_{ox}\frac{W}{L}} - V_{in}^2} \qquad (14.16) \\
&= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}V_{in}\sqrt{4(V_{GS} - V_{TH})^2 - V_{in}^2} \qquad (14.17)
\end{aligned}$$

If $|V_{in}| \ll V_{GS} - V_{TH}$, then

$$\begin{aligned}
I_{D1} - I_{D2} &= \mu_n C_{ox}\frac{W}{L}V_{in}(V_{GS} - V_{TH})\sqrt{1 - \frac{V_{in}^2}{4(V_{GS} - V_{TH})^2}} \qquad (14.18) \\
&\approx \mu_n C_{ox}\frac{W}{L}V_{in}(V_{GS} - V_{TH})\left[1 - \frac{V_{in}^2}{8(V_{GS} - V_{TH})^2}\right] \qquad (14.19) \\
&= \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})\left[V_m \cos \omega t - \frac{V_m^3 \cos^3 \omega t}{8(V_{GS} - V_{TH})^2}\right] \qquad (14.20)
\end{aligned}$$

Since $\cos^3 \omega t = [3 \cos \omega t + \cos(3\omega t)]/4$, we obtain

$$I_{D1} - I_{D2} = g_m\left[V_m - \frac{3V_m^3}{32(V_{GS} - V_{TH})^2}\right]\cos \omega t - g_m\frac{V_m^3 \cos(3\omega t)}{32(V_{GS} - V_{TH})^2} \qquad (14.21)$$

If $V_m \gg 3V_m^3/[8(V_{GS} - V_{TH})^2]$, then

$$\frac{A_{HD3}}{A_F} \approx \frac{V_m^2}{32(V_{GS} - V_{TH})^2} \qquad (14.22)$$

Comparison of (14.15) and (14.22) indicates that the differential circuit exhibits much less distortion than its single-ended counterpart while providing the same voltage gain and output swing. For example, if $V_m = 0.2(V_{GS} - V_{TH})$, (14.15) and (14.22) yield a distortion of 5% and 0.125%, respectively.

While achieving a lower distortion, the differential pair consumes twice as much power as the CS stage because $I_{SS} = 2I$. The key point, however, is that even if the bias current of M_0 is raised to $2I$, (14.15) predicts that the distortion decreases by only a factor of $\sqrt{2}$ (with W/L maintained constant).

14.1.3 Effect of Negative Feedback on Nonlinearity

In Chapter 8, we observed that negative feedback makes the closed-loop gain relatively independent of the op amp's open-loop gain. Since nonlinearity can be viewed as variation of the small-signal gain with the input level, we expect that negative feedback suppresses this variation as well, yielding higher linearity for the closed-loop system.

Analysis of nonlinearity in a feedback system is quite complex. Here, we consider a simple, "mildly nonlinear" system to gain more insight. The reason is that, if properly designed, a feedback amplifier exhibits only small distortion components, lending itself to this type of analysis.

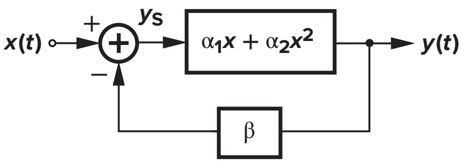


Figure 14.7 Feedback system incorporating a nonlinear feedforward amplifier.

Let us assume that the core amplifier in the system of Fig. 14.7 has an input-output characteristic $y \approx \alpha_1 x + \alpha_2 x^2$. We apply a sinusoidal input $x(t) = V_m \cos \omega t$, postulating that the output contains a fundamental component and a second harmonic and hence can be approximated as $y \approx a \cos \omega t + b \cos 2\omega t$.[1] Our objective is to determine a and b. The output of the subtractor can be written as

$$y_S = x(t) - \beta y(t) \tag{14.23}$$

$$= V_m \cos \omega t - \beta(a \cos \omega t + b \cos 2\omega t) \tag{14.24}$$

$$= (V_m - \beta a) \cos \omega t - \beta b \cos 2\omega t \tag{14.25}$$

This signal experiences the nonlinearity of the feedforward amplifier, thereby producing an output given by

$$y(t) = \alpha_1[(V_m - \beta) \cos \omega t - \beta b \cos 2\omega t] + \alpha_2[(V_m - \beta a) \cos \omega t - \beta b \cos 2\omega t]^2 \tag{14.26}$$

$$= [\alpha_1(V_m - \beta a) - \alpha_2(V_m - \beta a)\beta b] \cos \omega t + \left[-\alpha_1 \beta b + \frac{\alpha_2(V_m - \beta a)^2}{2}\right] \cos 2\omega t + \cdots \tag{14.27}$$

The coefficients of $\cos \omega t$ and $\cos 2\omega t$ in (14.27) must be equal to a and b, respectively:

$$a = (\alpha_1 - \alpha_2 \beta b)(V_m - \beta a) \tag{14.28}$$

$$b = -\alpha_1 \beta b + \frac{\alpha_2(V_m - \beta a)^2}{2} \tag{14.29}$$

[1] Note that higher harmonics and phase shifts through the system are neglected.

The assumption of small nonlinearity implies that both α_2 and b are small quantities, yielding $a \approx \alpha_1(V_m - \beta a)$ and hence

$$a = \frac{\alpha_1}{1 + \beta\alpha_1} V_m \tag{14.30}$$

which is to be expected because $\beta\alpha_1$ is the loop gain. To calculate b, we write

$$V_m - \beta a \approx \frac{a}{\alpha_1} \tag{14.31}$$

thus expressing (14.29) as

$$b = -\alpha_1 \beta b + \frac{1}{2}\alpha_2 \left(\frac{a}{\alpha_1}\right)^2 \tag{14.32}$$

That is

$$b(1 + \alpha_1\beta) = \frac{\alpha_2}{2}\left(\frac{a}{\alpha_1}\right)^2 \tag{14.33}$$

$$= \frac{\alpha_2}{2\alpha_1^2}\frac{\alpha_1^2}{(1 + \beta\alpha_1)^2} V_m^2 \tag{14.34}$$

It follows that

$$b = \frac{\alpha_2 V_m^2}{2}\frac{1}{(1 + \beta\alpha_1)^3} \tag{14.35}$$

For a meaningful comparison, we normalize the amplitude of the second harmonic to that of the fundamental:

$$\frac{b}{a} = \frac{\alpha_2 V_m}{2}\frac{1}{\alpha_1}\frac{1}{(1 + \beta\alpha_1)^2} \tag{14.36}$$

Without feedback, on the other hand, such a ratio would be equal to $(\alpha_2 V_m^2/2)/\alpha_1 V_m = \alpha_2 V_m/(2\alpha_1)$. Thus, the relative magnitude of the second harmonic has dropped by a factor of $(1 + \beta\alpha_1)^2$. Negative feedback therefore reduces the relative second harmonic by a factor of $(1 + \beta\alpha_1)^2$ and the gain by $1 + \beta\alpha_1$.

As described in Chapter 8, a feedback circuit employing a feedforward amplifier with a finite gain suffers from gain error. For a feedforward gain of A_0 and a feedback factor of β, the relative gain error is approximately equal to $1/(\beta A_0)$. If the feedforward amplifier exhibits nonlinearity, it is possible to derive a simple relationship between the gain error and the maximum nonlinearity of the overall feedback circuit. As illustrated in Fig. 14.8, we draw two straight lines, one representing the ideal characteristic (with a slope of $1/\beta$) and another passing through the end points of the actual characteristic. We note that with this construction, the nonlinearity, Δy_2, is always smaller than the gain error, Δy_1. This is, of course, true only if the small-signal gain drops monotonically as x goes from 0 to x_{max}, a typical behavior in most analog circuits. Thus, a sufficient condition to ensure that $\Delta y_2 < \epsilon$ is to guarantee that $\Delta y_1 < \epsilon$ by choosing a high open-loop gain for the amplifier.

The above condition is often applied in analog design because it is much easier to predict the open-loop gain than its nonlinearity. Of course, this simplification is obtained at the cost of a pessimistic choice of the amplifier's gain, an issue that becomes more serious as short-channel devices limit the voltage gain that can be achieved.

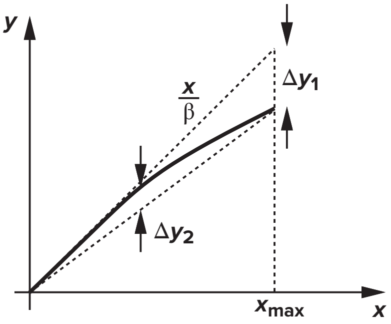


Figure 14.8 Gain error and nonlinearity in a feedback system.

14.1.4 Capacitor Nonlinearity

In switched-capacitor circuits, the voltage dependence of capacitors may introduce substantial distortion. While for a linear capacitor we have $Q = CV$, for a voltage-dependent capacitor we must write $dQ = C\,dV$. Thus, the total charge on a capacitor sustaining a voltage V_1 is

$$Q(V_1) = \int_0^{V_1} C\,dV \tag{14.37}$$

This means that the charge depends on the "history" of the voltage rather than its instantaneous value. In other words, we cannot write $Q(V_1) = CV_1$ even if C is evaluated for a voltage V_1 across the capacitor. To study the effect of capacitor nonlinearity, we express each capacitor as $C = C_0(1 + \alpha_1 V + \alpha_2 V^2 + \cdots)$.

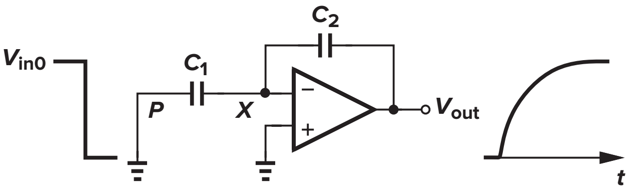


Figure 14.9 Effect of capacitor nonlinearity.

Let us consider the noninverting amplifier of Fig. 13.43(a), repeated in Fig. 14.9, as an example. At the beginning of the amplification mode, C_1 has a voltage equal to V_{in0} and C_2 a voltage of zero. Assuming that $C_1 \approx MC_0(1 + \alpha_1 V)$, where M is the nominal closed-loop gain ($C_1 = MC_2$), we obtain the charge across C_1 as

$$Q_1 = \int_0^{V_{in0}} C_1\,dV \tag{14.38}$$

$$= \int_0^{V_{in0}} MC_0(1 + \alpha_1 V)\,dV \tag{14.39}$$

$$= MC_0 V_{in0} + MC_0 \frac{\alpha_1}{2} V^2 \tag{14.40}$$

Similarly, if $C_2 \approx C_0(1 + \alpha_1 V)$, then the charge on this capacitor at the end of the amplification mode is

$$Q_2 = \int_0^{V_{out}} C_2\,dV \tag{14.41}$$

$$= C_0 V_{out} + C_0 \frac{\alpha_1}{2} V_{out}^2 \tag{14.42}$$

Equating Q_1 and Q_2 and solving for V_{out}, we have

$$V_{out} = \frac{1}{\alpha_1}\Big(-1 + \sqrt{1 + M\alpha_1^2 V_{in0}^2 + 2M\alpha_1 V_{in0}}\Big) \tag{14.43}$$

The last two terms under the square root are usually much less than unity and, since for $\epsilon \ll 1$, $\sqrt{1+\epsilon} \approx 1 + \epsilon/2 - \epsilon^2/8$, we can write

$$V_{out} \approx MV_{in0} + (1-M)\,\frac{M\alpha_1}{2}V_{in0}^2 \tag{14.44}$$

The second term in the above equation represents the nonlinearity resulting from the voltage dependence of the capacitor.

14.1.5 Nonlinearity in Sampling Circuits

Recall from Chapter 13 that the on-resistance of MOS switches in a sampling circuit varies with the input and output levels. For example, the NMOS switch in Fig. 14.10(a) exhibits a rising resistance as V_{in} and V_{out} increase. Similarly, the complementary topology in Fig. 14.10(b) displays an equivalent resistance that varies considerably as V_{in} and V_{out} go from 0 to V_{DD}. In contrast to the monotonic behavior derived in Chapter 13, R_{on} reaches a peak here due to the dependence of the mobility upon the vertical field in the channel. We wish to examine the harmonic distortion observed in the output due to this effect.

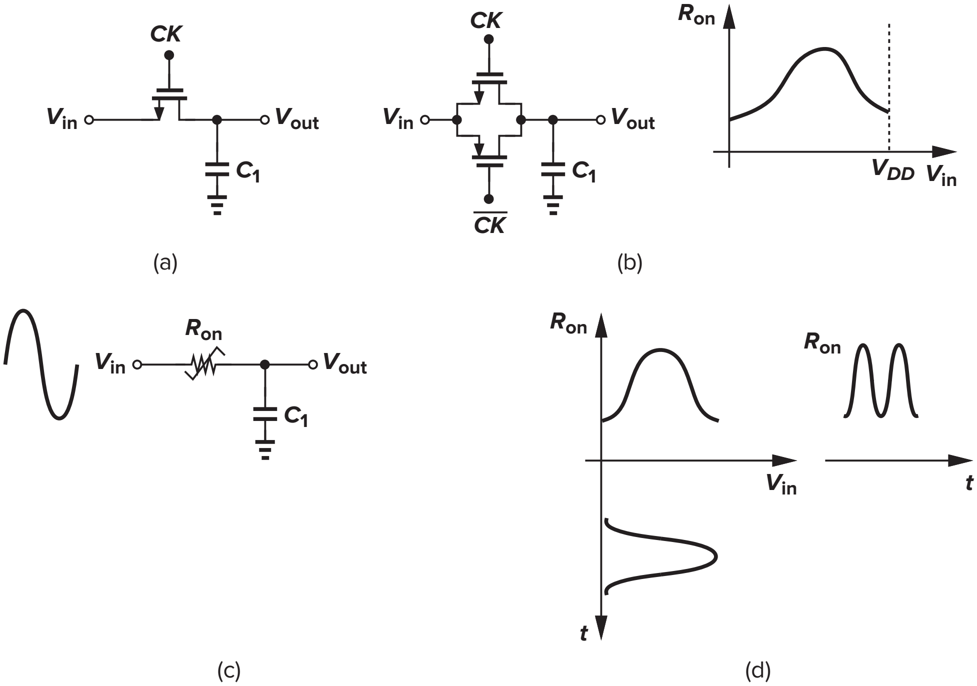


Figure 14.10 (a) Sampling circuit using NMOS switch, (b) sampling circuit using complementary devices, (c) representation of switch on-resistance by a nonlinear resistor, and (d) time-domain behavior.

As shown in Fig. 14.10(c), we apply a large sinusoid to the input, $V_{in} = V_0 \cos\omega_0 t + V_0$, where $V_0 = V_{DD}/2$, and seek the harmonics at the output. How do we analyze this circuit? The nonlinear dependence of R_{on} upon V_{in} or V_{out} presents a formidable challenge. Let us first assume that the resistance is linear and write the output as

$$V_{out}(t) = \frac{V_0}{\sqrt{R_{on}^2 C_1^2 \omega_0^2 + 1}} \cos[\omega_0 t - \tan^{-1}(R_{on} C_1 \omega_0)] + V_0 \tag{14.45}$$

In practice, the bandwidth must be large enough to negligibly attenuate the signal, i.e., $R_{on} C_1 \omega_0 \ll 1$, leading to

$$V_{out}(t) \approx V_0 \cos(\omega_0 t - R_{on} C_1 \omega_0) + V_0 \tag{14.46}$$

We now assume that this expression also holds for the nonlinear circuit if R_{on} is represented properly. It is interesting to note that R_{on}, and hence the *phase shift* from the input to the output, vary as V_{in} and V_{out} go up and down, thus creating distortion.

A key observation that simplifies our analysis is that, with a periodic input, R_{on} also varies *periodically* and can therefore be approximated by a Fourier series:

$$R_{on}(t) = R_0 + R_1 \cos \omega_0 t + R_2 \cos(2\omega_0 t) + \cdots \tag{14.47}$$

If we assume a roughly symmetric behavior for R_{on} in Fig. 14.10(b), we observe the time-domain behavior depicted in Fig. 14.10(d), where R_{on} varies at a rate equal to twice the input frequency. In this special case, $R_1 \approx 0$, but we continue with the general case. Replacing for R_{on} in Eq. (14.46), we have

$$V_{out}(t) \approx V_0 \cos[\omega_0 t - R_0 C_1 \omega_0 - R_1 C_1 \omega_0 \cos \omega_0 t - R_2 C_1 \omega_0 \cos(2\omega_0 t) - \cdots] + V_0 \tag{14.48}$$

If the cosine terms in the argument have amplitudes much less than 1 rad,

$$\begin{aligned} V_{out}(t) \approx\ & V_0 \cos(\omega_0 t - R_0 C_1 \omega_0) + \\ & [R_1 C_1 \omega_0 \cos \omega_0 t + R_2 C_1 \omega_0 \cos(2\omega_0 t) + \cdots] V_0 \sin(\omega_0 t - R_0 C_1 \omega_0) + V_0 \end{aligned}$$

We observe that the products $\cos \omega_0 t \sin(\omega_0 t - R_0 C_1 \omega_0)$, $\cos(2\omega_0 t) \sin(\omega_0 t - R_0 C_1 \omega_0)$, etc., give rise to harmonics. For example, the first two products respectively translate to a second harmonic and a third harmonic having peak amplitudes of $V_0 R_1 C_1 \omega_0 / 2$ and $V_0 R_2 C_1 \omega_0 / 2$, respectively. If we retain only these two harmonics, then

$$\text{THD} = \frac{R_1^2 + R_2^2}{4} C_1^2 \omega_0^2 \tag{14.49}$$

In a differential sampling switch, the even-order harmonics are suppressed.

14.1.6 Linearization Techniques

While amplifiers using "global" feedback (e.g., the switched-capacitor topologies of Chapter 13) can achieve a high linearity, stability and settling issues of feedback circuits limit their usage in high-speed applications. For this reason, many other techniques have been invented to linearize amplifiers with less compromise in speed.

The principle behind linearization is to reduce the dependence of the circuit's gain upon the input level. This usually translates into making the gain relatively independent of the transistor bias currents.

The simplest linearization method is source degeneration by means of a linear resistor. As shown in Fig. 14.11 for a common-source stage and revealed by the observations in the previous section, degeneration reduces the signal swing applied between the gate and the source of the transistor, thereby

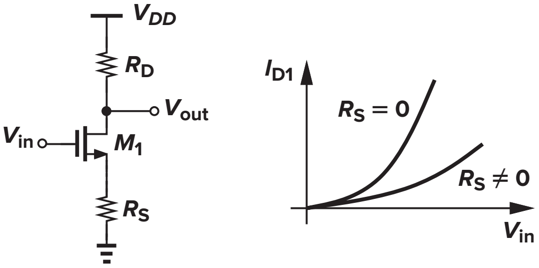


Figure 14.11 Common-source stage with resistive degeneration.

making the input/output characteristic more linear. From another point of view, neglecting body effect, we can write the overall transconductance of the stage as

$$G_m = \frac{g_m}{1 + g_m R_S} \tag{14.50}$$

which for large $g_m R_S$ approaches $1/R_S$, an input-independent value.

Note that the amount of linearization depends on $g_m R_S$ rather than on R_S alone. With a relatively constant G_m, the voltage gain, $G_m R_D$, is also relatively independent of the input and the amplifier is linearized.

▶ **Example 14.2**

A common-source stage biased at a current I_1 experiences an input voltage swing that varies the drain current from $0.75I_1$ to $1.25I_1$. Calculate the variation of the small-signal voltage gain (a) with no degeneration and (b) with degeneration such that $g_m R_S = 2$, where g_m denotes the transconductance at $I_D = I_1$.

Solution

Assuming square-law behavior, we have $g_m \propto \sqrt{I_D}$. For the case of no degeneration,

$$\frac{g_{m,high}}{g_{m,low}} = \sqrt{\frac{1.25}{0.75}} \tag{14.51}$$

With $g_m R_S = 2$,

$$\frac{G_{m,high}}{G_{m,low}} = \frac{\dfrac{\sqrt{1.25}g_m}{1 + \sqrt{1.25}g_m R_S}}{\dfrac{\sqrt{0.75}g_m}{1 + \sqrt{0.75}g_m R_S}} \tag{14.52}$$

$$= \sqrt{\frac{1.25}{0.75}} \cdot \frac{1 + 2\sqrt{0.75}}{1 + 2\sqrt{1.25}} \tag{14.53}$$

$$= 0.84\sqrt{\frac{1.25}{0.75}} \tag{14.54}$$

Thus, degeneration decreases the variation of the small-signal gain by approximately 16% in this case. ◀

Resistive degeneration presents trade-offs among linearity, noise, power dissipation, and gain. For large input voltage swings (e.g., 0.5 V_{pp}), it may be difficult to achieve even a voltage gain of 2 in a common-source stage if the nonlinearity is to remain below 1%.

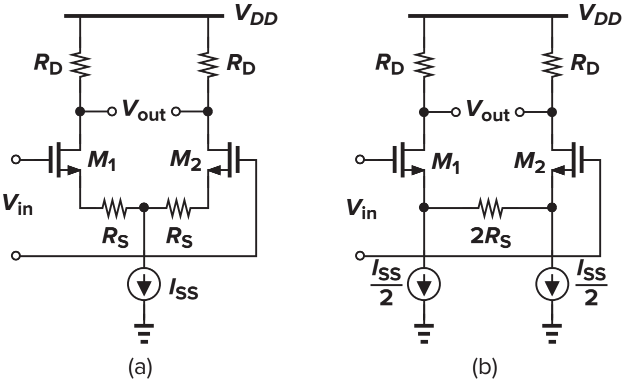


Figure 14.12 Source degeneration applied to a differential pair.

A differential pair can be degenerated as shown in Figs. 14.12(a) and (b). In Fig. 14.12(a), I_{SS} flows through the degeneration resistors, thereby consuming a voltage headroom of $I_{SS}R_S/2$, an important issue if a high level of degeneration is required. The circuit of Fig. 14.12(b), on the other hand, does not involve this issue, but it suffers from a slightly higher noise (and offset voltage) because the two tail current sources introduce some differential error. The reader can prove that if the output noise current of each current source is equal to $\overline{I_n^2}$, then the input-referred noise voltage of the circuit of Fig. 14.12(b) is higher than that of Fig. 14.12(a) by $2\overline{I_n^2}R_S^2$.

As depicted in Fig. 14.13, the resistor can be replaced by a MOSFET operating in the deep triode region. However, for large input swings, M_3 may not remain in the deep triode region, thereby experiencing substantial change in its on-resistance. Furthermore, V_b must track the input common-mode level so that R_{on3} can be defined accurately.

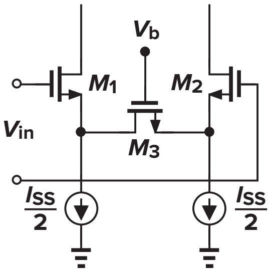


Figure 14.13 Differential pair degenerated by a MOSFET operating in the deep triode region.

A more practical variant of the above idea is illustrated in Fig. 14.14 [1]. Here, M_3 and M_4 are in the deep triode region if $V_{in} = 0$. As the gate voltage of M_1 becomes more positive than the gate voltage of M_2, transistor M_3 stays in the triode region because $V_{D3} = V_{G3} - V_{GS1}$, whereas M_4 eventually enters

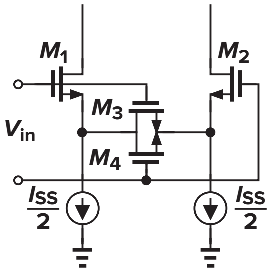


Figure 14.14 Differential pair degenerated by two MOSFETs operating in the triode region.

the saturation region because its drain voltage rises and its gate and source voltages fall. Thus, the circuit remains relatively linear even if one degeneration device goes into saturation. For the widest linear region, [1] suggests that $(W/L)_{1,2} \approx 7(W/L)_{3,4}$.

▶ Example 14.3

Using the offset mechanism introduced in Fig. 4.19, devise another linearization technique.

Solution

From Example 4.6, we know that the mismatch between the transistors' widths shifts the characteristics horizontally. Let us create a negative shift and a positive shift (equal in magnitude) in two differential pairs [Fig. 14.15(a)], observing that their G_m plots are offset by equal and opposite amounts. We now add the output currents by simply shorting the corresponding drains as shown in Fig. 14.15(b). The G_m plots also add (why?), yielding a result that is relatively constant for a *wider* range of $V_{in1} - V_{in2}$ and represents a more linear circuit. The ratio of 2 between the two widths is merely an example to illustrate the technique and may need to be modified for optimum linearity.

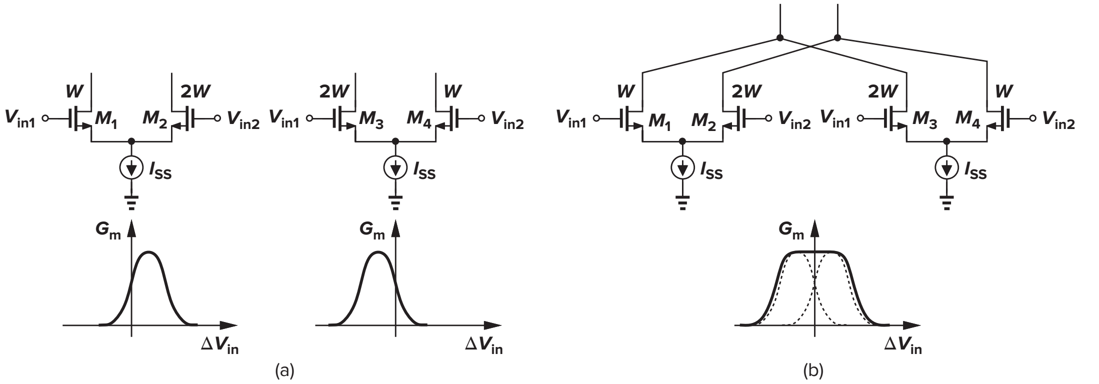


Figure 14.15

A linearization technique that avoids the use of resistors is based on the observation that a MOSFET operating in the triode region can provide a linear I_D/V_{GS} characteristic if its drain-source voltage is held constant: $I_D = (1/2)\mu C_{ox}(W/L)[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2]$. Illustrated in Fig. 14.16, the technique employs amplifiers A_1 and A_2 along with cascode devices M_3 and M_4 to force V_X and V_Y to be equal to V_b for varying input levels.

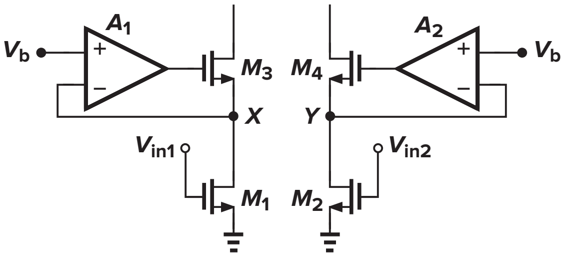


Figure 14.16 Differential pair using input devices operating in the triode region.

This circuit suffers from several drawbacks. First, the transconductance of M_1 and M_2, equal to $\mu_n C_{ox}(W/L)V_{DS}$, is relatively small because V_{DS} must be low enough to ensure that each input transistor remains in the triode region. Second, the input common-mode level must be tightly controlled, and it must track V_b so as to define I_{D1} and I_{D2}. Third, M_3, M_4, and the two auxiliary amplifiers contribute substantial noise to the output.

Another approach to linearizing voltage amplifiers is to perform "post correction." Illustrated in Fig. 14.17, the idea is to view the amplifier as a voltage-to-current (V/I) converter followed by a current-to-voltage (I/V) converter. If the V/I converter can be described as $I_{out} = f(V_{in})$ and the I/V converter as $V_{out} = f^{-1}(I_{in})$, then V_{out} is a linear function of V_{in}. That is, the second stage corrects the nonlinearity introduced by the first stage. As an example, recall from Chapter 4 that for the circuit shown in Fig. 14.18(a), we have

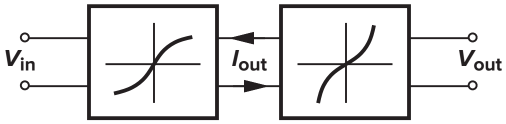


Figure 14.17 Voltage amplifier viewed as a cascade of two nonlinear stages.

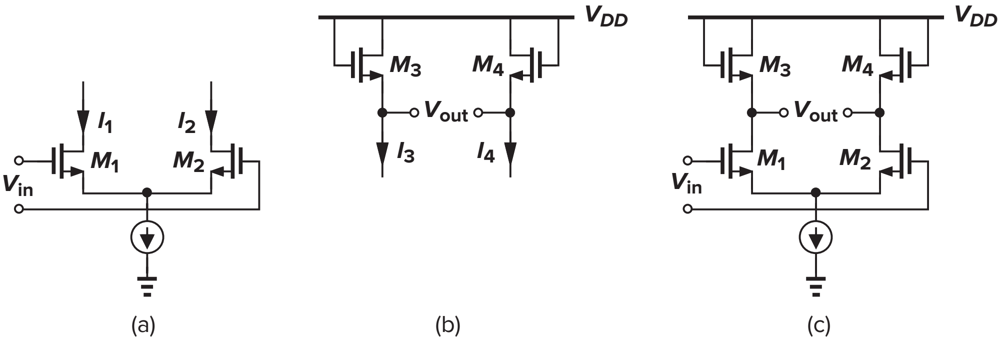


Figure 14.18 (a) Differential pair with nonlinear I/V characteristic; (b) diode-connected devices with nonlinear V/I characteristic; (c) circuit having linear input/output characteristic.

$$V_{in1} - V_{in2} = V_{GS1} - V_{GS2} \tag{14.55}$$

$$= \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\dfrac{W}{L}\right)_{1,2}}} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\dfrac{W}{L}\right)_{1,2}}} \tag{14.56}$$

We also note that for the circuit shown in Fig. 14.18(b),

$$V_{out} = V_{GS3} - V_{GS4} \tag{14.57}$$

$$= \sqrt{\frac{2I_3}{\mu_n C_{ox}\left(\dfrac{W}{L}\right)_{3,4}}} - \sqrt{\frac{2I_4}{\mu_n C_{ox}\left(\dfrac{W}{L}\right)_{3,4}}} \tag{14.58}$$

where channel-length modulation and body effect are neglected. It follows that for the circuit shown in Fig. 14.18(c),

$$V_{out} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4}}} \tag{14.59}$$

$$= \frac{1}{\sqrt{\left(\frac{W}{L}\right)_{3,4}}}(V_{in1} - V_{in2})\, sqrt\left(\frac{W}{L}\right)_{1,2} \tag{14.60}$$

Thus, as derived in Chapter 4, the voltage gain is equal to

$$A_v = \sqrt{\frac{\left(\frac{W}{L}\right)_{1,2}}{\left(\frac{W}{L}\right)_{3,4}}} \tag{14.61}$$

a quantity independent of the bias currents of the transistors.

In practice, body effect and other nonidealities in short-channel devices give rise to nonlinearity in this circuit. Furthermore, as the differential input level increases, driving M_1 or M_2 into the subthreshold region, Eqs. (14.56) and (14.58) no longer hold and the gain drops sharply.

It is possible to add local feedback to a degenerated differential pair to linearize it further. Illustrated in Fig. 14.19(a), the idea is to sense the output voltage of the differential pair by means of M_3 and M_4 and return a proportional current to the *sources* of M_1 and M_2. The reader can readily prove that the feedback is negative. We assume that the circuit is symmetric and $I_1 = \cdots = I_4$.

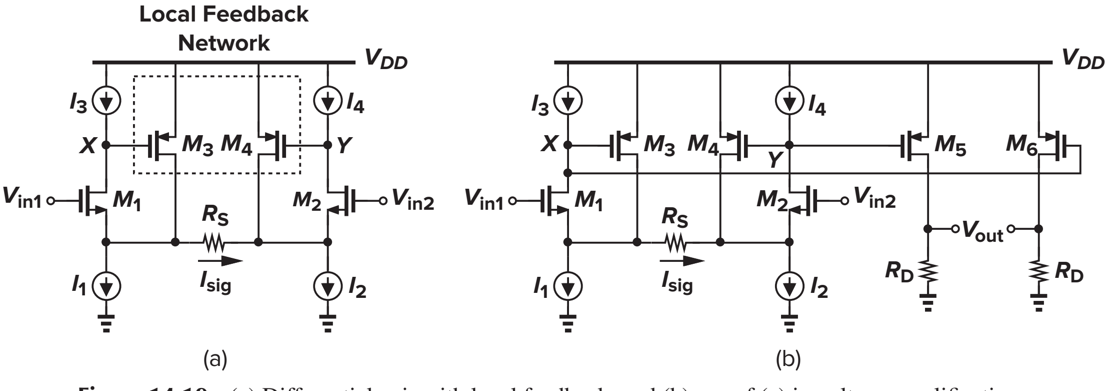


Figure 14.19 (a) Differential pair with local feedback, and (b) use of (a) in voltage amplification.

If channel-length modulation and body effect are neglected, we observe that $I_{D1} = I_3$ and $I_{D2} = I_4$ *regardless* of the input signal. Thus, the input transistors maintain a constant V_{GS} as $V_{in} = V_{in1} - V_{in2}$ varies. Moreover, since $I_1 = I_3 = I_{D1}$ and $I_2 = I_4 = I_{D2}$, the current flowing through R_S must be provided by only M_3 and M_4. Denoting this current by I_{sig}, we have

$$V_{in} = V_{GS1} + I_{sig}R_S - V_{GS2} \tag{14.62}$$

$$= I_{sig}R_S \tag{14.63}$$

Interestingly, the currents produced by M_3 and M_4 are linearly proportional to V_{in} because the feedback from their drains to their sources guarantees a constant V_{GS}. Note, however, that $V_X - V_Y$ is not.

The reader may wonder where this topology's output is! As depicted in Fig. 14.19(b), we copy the PMOS currents onto M_5 and M_6 and allow the results to flow through (linear) resistors. Since I_{D3} and I_{D4} are equal and opposite, it follows from (14.63) that

$$V_{out} = \frac{2R_D}{R_S} V_{in} \tag{14.64}$$

where the PMOS devices are assumed identical. The circuit excluding the R_D's operates as a linear voltage-to-current converter ("transconductor").

The above topology entails two issues. First, the large number of devices in the signal path produces substantial noise. In addition to M_1–M_4, the top and bottom current sources also contribute differential noise. Second, due to the dependence of r_O upon V_{DS} in short-channel devices (Chapter 17), the output stage introduces some nonlinearity.

14.2 ■ Mismatch

Our study of differential amplifiers in the previous chapters has mostly assumed that the circuits are perfectly symmetric, i.e., the two sides exhibit identical properties and bias currents. In reality, however, nominally-identical devices suffer from a finite mismatch due to uncertainties in each step of the manufacturing process. For example, as illustrated in Fig. 14.20, the gate dimensions of MOSFETs suffer from random, microscopic variations, introducing mismatches between the equivalent lengths and widths of two transistors that are identically laid out. Also, MOS devices exhibit threshold voltage mismatch because, from Eq. (2.1), V_{TH} is a function of the doping levels in the channel and the gate, and these levels vary randomly from one device to another.

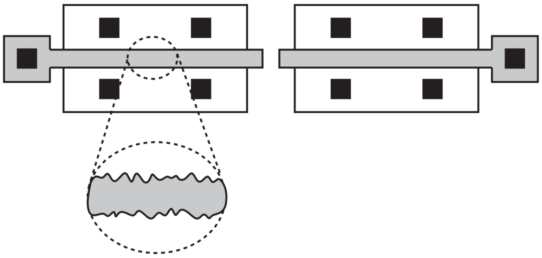


Figure 14.20 Random mismatches due to microscopic variations in device dimensions.

Study of mismatch consists of two steps: (1) identify and formulate the mechanisms that lead to mismatch between devices; and (2) analyze the effect of device mismatches upon the performance of circuits. Unfortunately, the first step is quite complex and heavily dependent on the fabrication technology and the layout, often requiring actual measurements of mismatches. For example, the achievable mismatch between capacitors is typically quoted to be 0.1%, but this value is not derived from any fundamental quantities. Layout techniques for minimum mismatch are described in Chapter 19.

Expressing the characteristics of a MOSFET in saturation as $I_D = (1/2)\mu C_{ox}(W/L)(V_{GS} - V_{TH})^2$, we observe that mismatches between μ, C_{ox}, W, L, and V_{TH} result in mismatches between drain currents (for a given V_{GS}) or gate-source voltages (for a given drain current) of two nominally-identical transistors. Intuitively, we expect that as W and L increase, their relative mismatches, $\Delta W/W$ and $\Delta L/L$, respectively, decrease, i.e., larger devices exhibit smaller mismatches. A more important observation is that all of the mismatches decrease as the *area* of the transistor, WL, increases. For example, increasing W

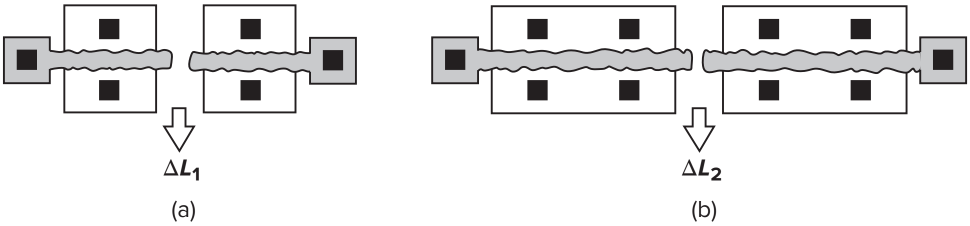


Figure 14.21 Reduction of length mismatch as a result of increasing the width.

reduces both $\Delta W/W$ *and* $\Delta L/L$. This is because as WL increases, random variations experience greater "averaging," thereby falling in magnitude. For the case depicted in Fig. 14.21, $\Delta L_2 < \Delta L_1$ because, if the device is viewed as many small parallel transistors (Fig. 14.22), each having a width of W_0, then we can write the equivalent length as $L_{eq} \approx (L_1 + L_2 + \cdots + L_n)/n$. The overall variation is therefore given by

$$\Delta L_{eq} \approx \left(\Delta L_1^2 + \Delta L_2^2 + \cdots + \Delta L_n^2\right)^{1/2}/n \tag{14.65}$$

$$= \frac{\left(n\Delta L_0^2\right)^{1/2}}{n} \tag{14.66}$$

$$= \frac{\Delta L_0}{\sqrt{n}} \tag{14.67}$$

where ΔL_0 is the statistical variation of the length for a transistor with width W_0. Equation (14.67) reveals that for a given W_0, as n increases, the variation of L_{eq} decreases.

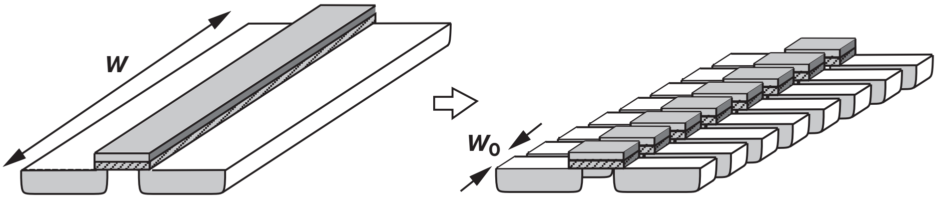


Figure 14.22 Wide MOSFET viewed as a parallel combination of narrow devices.

The above result can be extended to other device parameters as well. For example, we postulate that μC_{ox} and V_{TH} suffer from less mismatch if the device area increases. Illustrated in Fig. 14.23, the reason is that a large transistor can be decomposed into a series and parallel combination of small unit transistors with dimensions W_0 and L_0, each exhibiting $(\mu C_{ox})_j$ and V_{THj}. For given W_0 and L_0, as the number of unit transistors increases, μC_{ox} and V_{TH} experience greater averaging, leading to smaller mismatch between two large transistors.

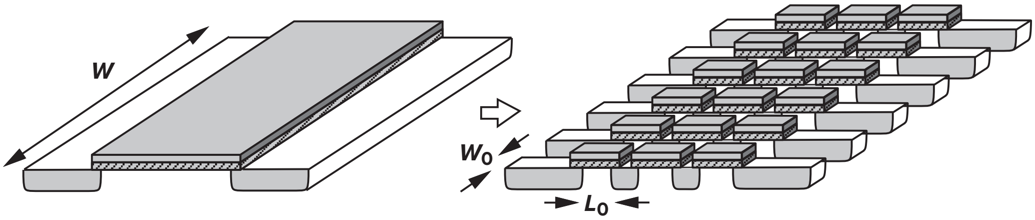


Figure 14.23 Large MOSFET viewed as a combination of small devices.

The foregoing qualitative observations have been verified mathematically and experimentally [2, 3]. Here, we state without proof that

$$\Delta V_{TH} = \frac{A_{VTH}}{\sqrt{WL}} \tag{14.68}$$

$$\Delta\left(\mu C_{OX}\frac{W}{L}\right) = \frac{A_K}{\sqrt{WL}} \tag{14.69}$$

where A_{VTH} and A_K are proportionality factors and obtained from measurements.

▶ **Example 14.4**

A differential pair incorporates transistors having a length of 40 nm. If $A_{VTH} = 4$ mV $\cdot$ μm for 40-nm technology, what is the minimum device width that guarantees $\Delta V_{TH} \leq 2$ mV?

Solution

We write

$$W = \frac{A_{VTH}^2}{L\Delta V_{TH}^2} \tag{14.70}$$

$$= 100\ \mu\text{m} \tag{14.71}$$

We observe the very large W/L necessary for low offsets in nanometer technologies. ◀

Since the channel capacitance is proportional to WLC_{ox}, we note that ΔV_{TH} and the channel capacitance bear a trade-off.

14.2.1 Effect of Mismatch

We now study the effect of device mismatch upon the performance of circuits. Mismatches lead to three significant phenomena: dc offsets, finite even-order distortion, and lower common-mode rejection. The last phenomenon was studied in Chapter 4.

DC Offsets Consider the differential pair shown in Fig. 14.24(a). With $V_{in} = 0$ and perfect symmetry, $V_{out} = 0$, but in the presence of mismatches, $V_{out} \neq 0$. We say that the circuit suffers from a dc "offset" equal to the observed value of V_{out} when V_{in} is set to zero. In practice, it is more meaningful to specify the input-referred offset voltage, defined as the input level that forces the output voltage to go to zero

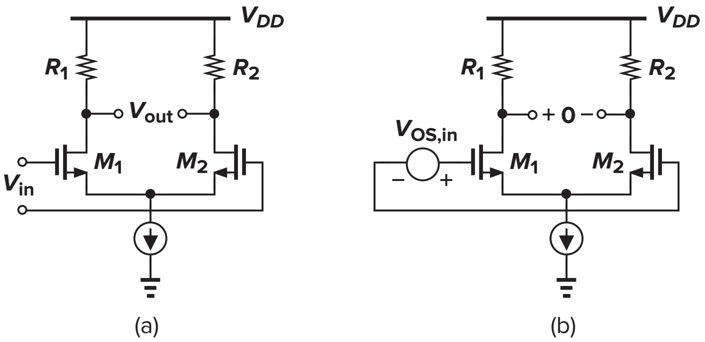


Figure 14.24 (a) Differential pair with offset measured at the output; (b) circuit of (a) with its offset referred to the input.

[Fig. 14.24(b)]. Note that $|V_{OS,in}| = |V_{OS,out}|/A_v$. As with random noise, the polarity of random offsets is unimportant.

How does offset limit the performance? Suppose the differential pair of Fig. 14.24 is to amplify a small input voltage. Then, as depicted in Fig. 14.25, the output contains amplified replicas of both the signal and the offset. In a cascade of direct-coupled amplifiers, the dc offset may experience so much gain that it drives the latter stages into nonlinear operation.

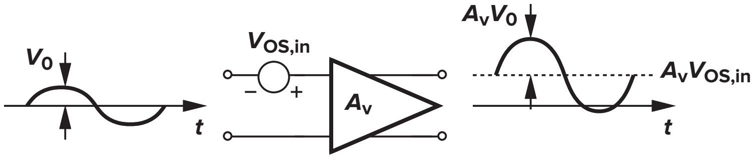


Figure 14.25 Effect of offset in an amplifier.

A more important effect of offset is the limitation on the precision with which signals can be measured. For example, if an amplifier is used to determine whether the input signal is greater or less than a reference, V_{REF} (Fig. 14.26), then the input-referred offset imposes a lower bound on the minimum $V_{in} - V_{REF}$ that can be detected reliably.

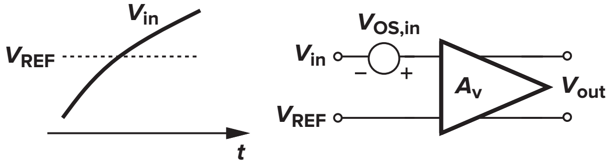


Figure 14.26 Accuracy limitation of an amplifier due to offset.

Let us now calculate the offset voltage of a differential pair, assuming that both the input transistors and the load resistors suffer from mismatch. As illustrated in Fig. 14.24(b), our objective is to find the value of $V_{OS,in}$ such that $V_{out} = 0$. The device mismatches are incorporated as $V_{TH1} = V_{TH}$, $V_{TH2} = V_{TH} + \Delta V_{TH}$; $(W/L)_1 = W/L$, $(W/L)_2 = W/L + \Delta(W/L)$; $R_1 = R_D$, $R_2 = R_D + \Delta R$. For simplicity, $\lambda = \gamma = 0$, and mismatches in $\mu_n C_{ox}$ are neglected. For $V_{out} = 0$, we must have $I_{D1}R_1 = I_{D2}R_2$, concluding that I_{D1} cannot be equal to I_{D2}. Thus, we assume that $I_{D1} = I_D$, $I_{D2} = I_D + \Delta I_D$.

Since $V_{OS,in} = V_{GS1} - V_{GS2}$, we have

$$V_{OS,in} = \sqrt{\frac{2I_{D1}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_1}} + V_{TH1} - \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}\left(\frac{W}{L}\right)_2}} - V_{TH2} \tag{14.72}$$

$$= \sqrt{\frac{2}{\mu_n C_{ox}}}\left[\sqrt{\frac{I_D}{\frac{W}{L}}} - \sqrt{\frac{I_D + \Delta I_D}{\frac{W}{L} + \Delta\left(\frac{W}{L}\right)}}\right] - \Delta V_{TH} \tag{14.73}$$

$$= \sqrt{\frac{2}{\mu_n C_{ox}}}\sqrt{\frac{I_D}{W/L}}\left[1 - \sqrt{\frac{1 + \frac{\Delta I_D}{I_D}}{1 + \Delta\left(\frac{W}{L}\right)\Big/\left(\frac{W}{L}\right)}}\right] - \Delta V_{TH} \tag{14.74}$$

Assuming that $\Delta I_D/I_D$ and $\Delta(W/L)/(W/L) \ll 1$, and noting that for $\epsilon \ll 1$ we can write $\sqrt{1+\epsilon} \approx 1+\epsilon/2$ and $(\sqrt{1+\epsilon})^{-1} \approx 1-\epsilon/2$, we reduce (14.74) to

$$V_{OS,in} = \sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left\{1-\left(1+\frac{\Delta I_D}{2I_D}\right)\left[1-\frac{\Delta(W/L)}{2(W/L)}\right]\right\} - \Delta V_{TH} \tag{14.75}$$

$$= \sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left[\frac{-\Delta I_D}{2I_D}+\frac{\Delta(W/L)}{2(W/L)}\right] - \Delta V_{TH} \tag{14.76}$$

where the product of two small quantities is neglected. Recall that $I_{D1}R_1 = I_{D2}R_2$, and hence $I_D R_D = (I_D+\Delta I_D)(R_D+\Delta R_D) \approx I_D R_D + R_D\Delta I_D + I_D\Delta R_D$. Consequently, $\Delta I_D/I_D \approx -\Delta R_D/R_D$, and

$$V_{OS,in} = \frac{1}{2}\sqrt{\frac{2I_D}{\mu_n C_{ox}\left(\frac{W}{L}\right)}}\left[\frac{\Delta R_D}{R_D}+\frac{\Delta(W/L)}{(W/L)}\right] - \Delta V_{TH} \tag{14.77}$$

We also recognize that the square-root quantity is approximately equal to the equilibrium overdrive voltage of each transistor, $V_{GS}-V_{TH}$, and

$$V_{OS,in} = \frac{V_{GS}-V_{TH}}{2}\left[\frac{\Delta R_D}{R_D}+\frac{\Delta(W/L)}{(W/L)}\right] - \Delta V_{TH} \tag{14.78}$$

Equation (14.78) is an important result, revealing the dependence of $V_{OS,in}$ on device mismatches and bias conditions. We note that (1) the contribution of load resistor mismatch and transistor dimension mismatch *increases* with the equilibrium overdrive, and (2) the threshold voltage mismatch is directly referred to the input. Thus, it is desirable to minimize $V_{GS}-V_{TH}$ by lowering the tail current or increasing the transistor widths. In reality, since mismatches are independent statistical variables, we express (14.78) as[2]

$$V_{OS,in}^2 = \left(\frac{V_{GS}-V_{TH}}{2}\right)^2\left\{\left(\frac{\Delta R_D}{R_D}\right)^2+\left[\frac{\Delta(W/L)}{(W/L)}\right]^2\right\} + \Delta V_{TH}^2 \tag{14.79}$$

where squared quantities represent standard deviations.

To gain more insight into the effect of offset, let us establish an analogy between offset and *noise*. If the two inputs of a differential pair are shorted, the output voltage exhibits a finite noise, that is, a voltage that varies with time. We may therefore say that the offset voltage of a differential pair resembles a very low-frequency noise component, varying so slowly that it appears constant in our measurements. Viewed as such, offsets can be incorporated as noise sources, allowing us to utilize the analysis techniques developed in Chapter 7. To this end, we represent the offset of two nominally-identical transistors by a voltage source equal to (14.79) in series with the gate of one of the transistors.

[2] As mentioned earlier, ΔV_{TH} does depend on W, an effect that can be added as a cross-correlation term. We neglect this term here for simplicity.

▶ **Example 14.5**

Calculate the input-referred offset voltage of the circuit shown in Fig. 14.27(a). Assume all of the transistors operate in saturation.

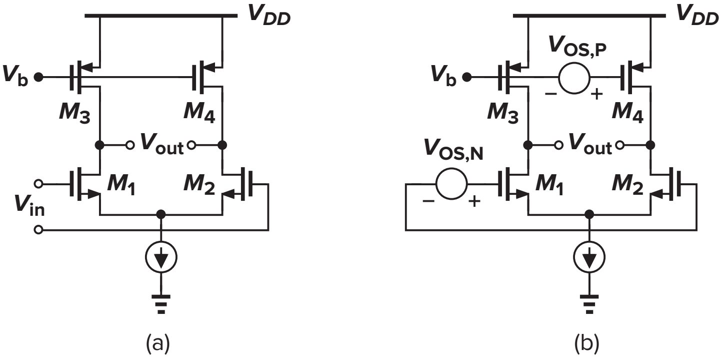


Figure 14.27

Solution

We insert the offsets of the NMOS and PMOS pairs as in Fig. 14.27(b). To obtain $I_{D1} = I_{D2}$ and $I_{D3} = I_{D4}$, we have from (14.78)

$$V_{OS,N} = \frac{(V_{GS} - V_{TH})_N}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_N + \Delta V_{TH,N} \tag{14.80}$$

$$V_{OS,P} = \frac{|V_{GS} - V_{TH}|_P}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_P + \Delta V_{TH,P} \tag{14.81}$$

From the noise analysis in Chapter 7, $V_{OS,P}$ is amplified by a gain of $g_{mP}(r_{ON}||r_{OP})$ and divided by $g_{mN}(r_{ON}||r_{OP})$ when referred to the main input. As a result,

$$\begin{aligned}V_{OS,in} = &\left\{\frac{|V_{GS} - V_{TH}|_P}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_P + \Delta V_{TH,P}\right\}\frac{g_{mP}}{g_{mN}}\\ &+ \frac{(V_{GS} - V_{TH})_N}{2}\left[\frac{\Delta(W/L)}{W/L}\right]_N + \Delta V_{TH,N}\end{aligned} \tag{14.82}$$

In practice, we add the "power" of these terms, as exemplified by (14.79). Note that, as with noise, the contribution of the offset of the PMOS pair is proportional to g_{mP}/g_{mN}.

◀

The foregoing example can be better understood if we study the offset behavior of current sources. Consider the nominally-identical current sources M_1 and M_2 in Fig. 14.28. Neglecting channel-length modulation, we determine the total mismatch between I_{D1} and I_{D2} by calculating the total differential.

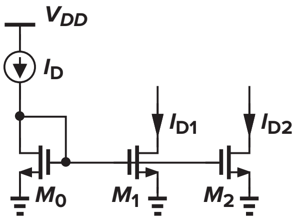


Figure 14.28 Mismatch between two current sources.

Recall from calculus that if $y = f(x_1, x_2, \ldots)$, then the total differential is given by

$$\Delta y = \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots \tag{14.83}$$

Equation (14.83) simply means that each mismatch component Δx_j is weighted by the corresponding sensitivity $\partial f/\partial x_j$ as it contributes to the total mismatch. Since $I_D = (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$, we have

$$\Delta I_D = \frac{\partial I_D}{\partial (W/L)}\Delta\left(\frac{W}{L}\right) + \frac{\partial I_D}{\partial (V_{GS} - V_{TH})}\Delta(V_{GS} - V_{TH}) \tag{14.84}$$

where mismatches in $\mu_n C_{ox}$ are neglected. It follows that

$$\Delta I_D = \frac{1}{2}\mu_n C_{ox}(V_{GS} - V_{TH})^2\Delta\left(\frac{W}{L}\right) - \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})\Delta V_{TH} \tag{14.85}$$

Unlike the input-referred offset *voltage*, current mismatch is usually normalized to the average value to allow a meaningful comparison:

$$\frac{\Delta I_D}{I_D} = \frac{\Delta(W/L)}{W/L} - 2\frac{\Delta V_{TH}}{V_{GS} - V_{TH}} \tag{14.86}$$

This result suggests that, to minimize current mismatch, the overdrive voltage must be *maximized*, a trend opposite to that in (14.78). This is because as $V_{GS} - V_{TH}$ increases, threshold mismatch has a lesser effect on the device currents.

The dependence of offset voltage and current mismatches upon the overdrive voltage is similar to our observations in Chapter 7 for corresponding noise quantities. For a given current, the input noise voltage of a differential pair increases as the overdrive increases because $g_m = 2I_D/(V_{GS} - V_{TH})$. Also, the output noise current of current sources is proportional to g_m and hence proportional to $V_{GS} - V_{TH}$.

Even-Order Distortion Our study of nonlinearity in Sec. 14.1 implies that, by virtue of odd symmetry, differential circuits are free from even-order distortion. In reality, however, mismatches degrade the symmetry, thereby introducing a finite even-order nonlinearity.

Analysis of the even-order distortion in the presence of mismatches is generally quite complex, often necessitating simulations. Here, we consider a simple case to gain some insight. Suppose the two signal paths in a differential circuit are represented by $y_1 \approx \alpha_1 x_1 + \alpha_2 x_1^2 + \alpha_3 x_1^3$ and $y_2 \approx \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3$ (Fig. 14.29). The differential output is given by

$$y_1 - y_2 = (\alpha_1 x_1 - \beta_2 x_2) + \left(\alpha_2 x_1^2 - \beta_2 x_2^2\right) + \left(\alpha_3 x_1^3 - \beta_3 x_2^3\right) \tag{14.87}$$

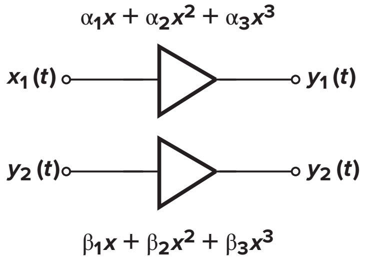


Figure 14.29 Effect of mismatch on second-order distortion.

which, for $x_1 = -x_2$, reduces to

$$y_1 - y_2 = (\alpha_1 + \beta_1)x_1 + (\alpha_2 - \beta_2)x_1^2 + (\alpha_3 + \beta_3)x_1^3 \tag{14.88}$$

If $x_1(t) = A\cos\omega t$, then the second harmonic has an amplitude equal to $(\alpha_2 - \beta_2)A^2/2$, i.e., proportional to the mismatch between the second-order coefficients of the input/output characteristic.

We should also mention that since at high frequencies, signals experience considerable phase shift, even-order distortion may arise from *phase* mismatch. This point is considered in Problem 14.14.1.

In circuits dissipating a high power, thermal gradients across the chip may create asymmetries. For example, if one transistor of a differential pair is closer to a high-power output stage than the other transistor, then mismatches arise between the threshold voltages and the mobilities of the two transistors.

14.2.2 Offset Cancellation Techniques

As mentioned above, the threshold voltage mismatch of MOSFETS trades with the channel capacitance. For example, a threshold mismatch of 1 mV translates to roughly 300 fF of channel capacitance for each transistor in a 0.6-μm technology. If many differential pairs are connected in parallel (e.g., in an A/D converter), the input capacitance becomes prohibitively large, severely degrading the speed and/or demanding high power dissipation in the preceding stage. Another difficulty is that mechanical stress may increase the offset voltages after a circuit is packaged. For these reasons, many high-precision systems require electronic cancellation of the offsets. As explained below, offset cancellation can also reduce $1/f$ noise of amplifiers considerably.

As our first step toward understanding the principle of offset cancellation, let us consider the circuit of Fig. 14.30(a), where a differential amplifier having an input-referred offset voltage V_{OS} is followed by two series capacitors. Now suppose, as shown in Fig. 14.30(b), the inputs are shorted together, driving the amplifier output to $V_{out} = A_v V_{OS}$. Furthermore, assume that during this period, nodes X and Y are shorted together as well. We note that when all of the node voltages are settled and $A_v V_{OS}$ is stored across C_1 and C_2, a zero differential input results in a zero difference between V_X and V_Y. Thus, after S_1 and S_2 turn off, the circuit consisting of the amplifier and C_1 and C_2 exhibits a zero offset voltage, amplifying only *changes* in the differential input voltage. In practice, the inputs and outputs must be shorted to proper common-mode voltages [Fig. 14.30(c)].

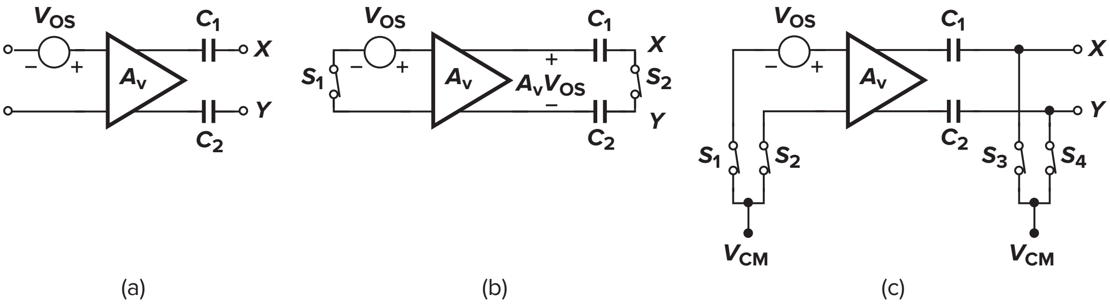


Figure 14.30 (a) Simple amplifier with capacitive coupling at the output; (b) circuit of (a) with its inputs and outputs shorted; (c) proper setting of the common-mode level during offset cancellation.

In summary, this type of offset cancellation "measures" the offset by setting the differential input to zero and stores the result on capacitors in series with the output. The circuit therefore requires a dedicated offset cancellation period, during which the actual input is disabled. Figure 14.31 depicts the final topology, where CK denotes the offset cancellation command. Called "output offset storage," this

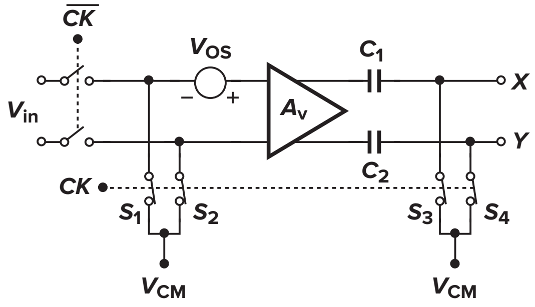


Figure 14.31 Control of amplification and offset cancellation modes by a clock.

technique reduces the overall offset to zero if S_3-S_4 exhibit no charge injection mismatch. Note, however, that if A_v is large, $A_v V_{OS}$ may "saturate" the amplifier output. For this reason, A_v is typically chosen to be less than roughly 10.

In applications where a high voltage gain is required, the topology of Fig. 14.32(a) may be employed. Called "input offset storage," this approach incorporates two series capacitors at the input and places the amplifier in a unity-gain negative-feedback loop during offset cancellation. Thus, from Fig. 14.32(b), $V_{out} = V_{XY}$ and $(V_{out} - V_{OS})(-A_v) = V_{out}$. That is

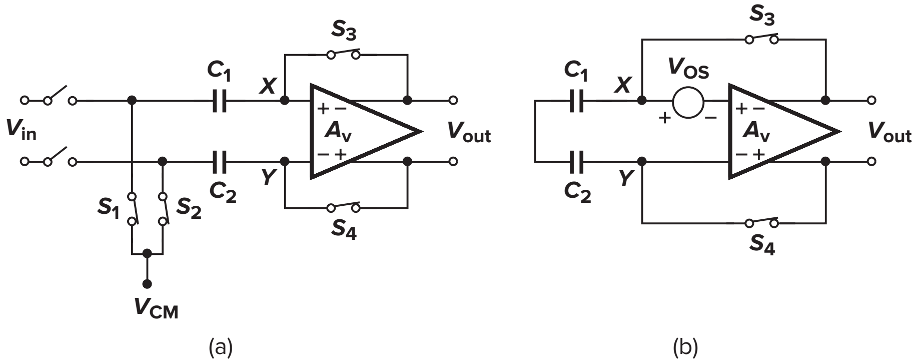


Figure 14.32 (a) Input offset storage; (b) circuit of (a) in the offset cancellation mode.

$$V_{out} = \frac{A_v}{1 + A_v} V_{OS} \tag{14.89}$$

$$\approx V_{OS} \tag{14.90}$$

In essence, the circuit reproduces the amplifier's offset at nodes X and Y, storing the result on C_1 and C_2. Note that for a zero differential input, the differential output is equal to V_{OS}. Therefore, the input-referred offset voltage of the overall circuit (after S_3 and S_4 turn off) equals V_{OS}/A_v if S_3 and S_4 match perfectly (and the input capacitance of the amplifier is much less than C_1 and C_2). In reality, however, when S_3 and S_4 turn off, their charge injection mismatch may saturate the amplifier if A_v is very large.

The general drawback of input and output storage techniques is that they introduce capacitors in the signal path, a particularly serious issue in op amps and feedback systems. The bottom-plate parasitic of the capacitors may reduce the magnitude of the poles in the circuit, thereby degrading the phase margin. Even in open-loop amplifiers, this parasitic may limit the settling speed, intensifying the speed-power trade-off.

To resolve the above issues, the offset cancellation scheme can isolate the signal path from the offset storage capacitors though the use of an "auxiliary" amplifier. Consider the topology shown in Fig. 14.33, where A_{aux} amplifies the differential voltage V_1 stored across C_1 and C_2 and subtracts the result from the output of A_1. We note that if $V_{OS1}A_1 = V_1 A_{aux}$, then for $V_{in} = 0$, $V_{out} = 0$, and the circuit is free from offsets. The key point here is that C_1 and C_2 do not appear in the signal path.

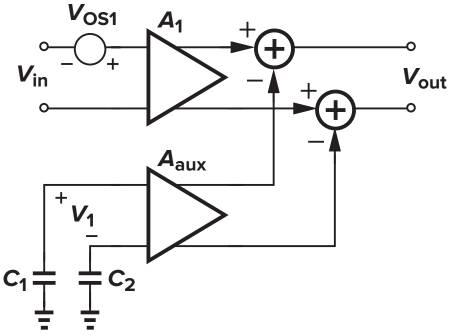


Figure 14.33 Addition of an auxiliary stage to remove the offset of an amplifier.

How is V_1 generated in Fig. 14.33? This is accomplished as illustrated in Fig. 14.34. Here, a second stage, A_2, is added and its output is sensed by A_{aux} during offset cancellation. To understand the operation, suppose that first only S_1 and S_2 are on, yielding $V_{out} = V_{OS1}A_1A_2$. Now, assume that S_3 and S_4 turn on, placing A_2 and A_{aux} in a negative-feedback loop. The reader can show that V_{out} then drops by a factor approximately equal to the loop gain: $V_{OS1}A_1A_2/(A_2A_{aux}) = V_{OS1}A_1/A_{aux}$. Stored across C_1 and C_2, this value is indeed the required V_1 in Fig. 14.33 because $(V_{OS1}A_1/A_{aux})A_{aux} = V_{OS1}A_1$.

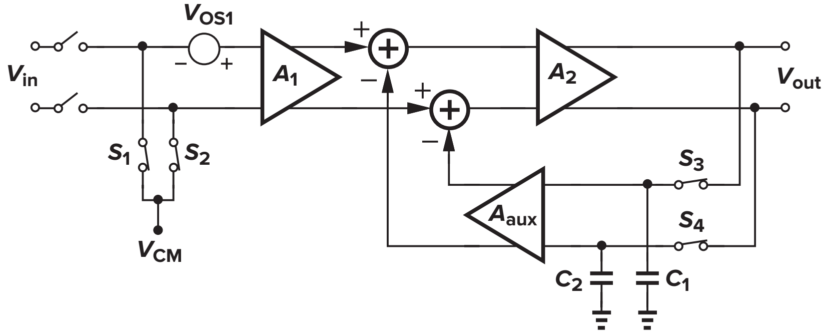


Figure 14.34 Auxiliary amplifier placed in a feedback loop during offset cancellation.

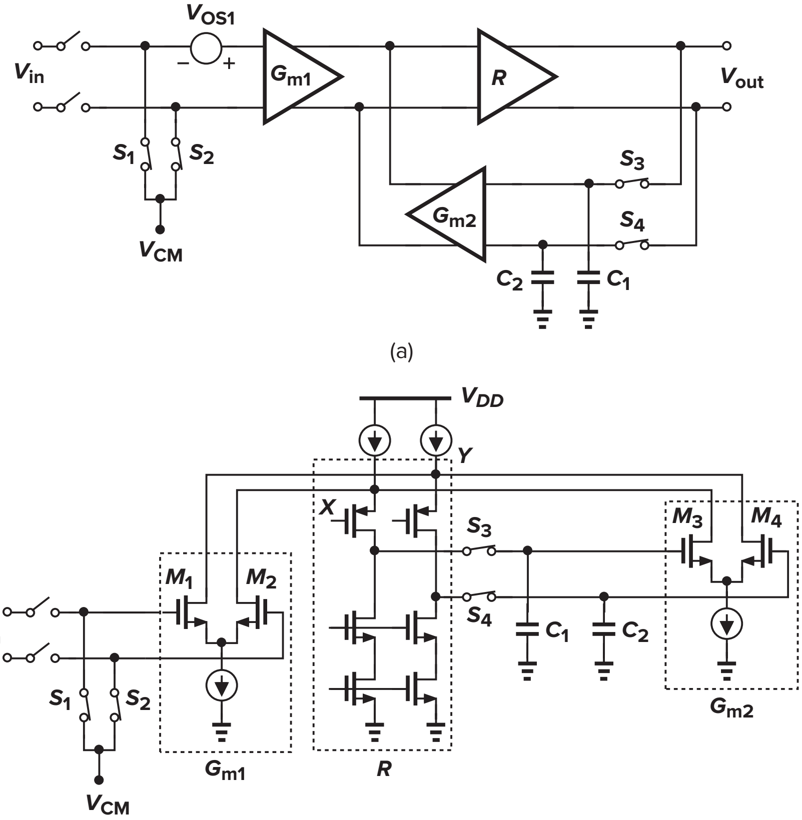
The topology of Fig. 14.34 suffers from two drawbacks. First, two voltage gain stages in the signal path may not be desirable in a high-speed op amp. Second, addition of the output voltages of A_1 and A_2 is quite difficult. For these reasons, the technique is usually realized as shown in Fig. 14.35(a), where each G_m stage is simply a differential pair and the R stage represents a transimpedance amplifier. As exemplified by Fig. 14.35(b), G_{m1} and R may in fact constitute a one-stage op amp, while G_{m2} adds an offset correction current at the low-impedance nodes X and Y.

Let us now examine the offset cancellation in Fig. 14.35(a) carefully, taking the offset voltage of G_{m2} into account as well. As depicted in Fig. 14.36, we can write

$$[G_{m1}V_{OS1} - G_{m2}(V_{out} - V_{OS2})]R = V_{out} \tag{14.91}$$

Thus,

$$V_{out} = \frac{G_{m1}RV_{OS1} + G_{m2}RV_{OS2}}{1 + G_{m2}R} \tag{14.92}$$



(b)

Figure 14.35 (a) Circuit of Fig. 14.34 using G_m and R stages; (b) realization of (a) in a folded-cascode op amp.

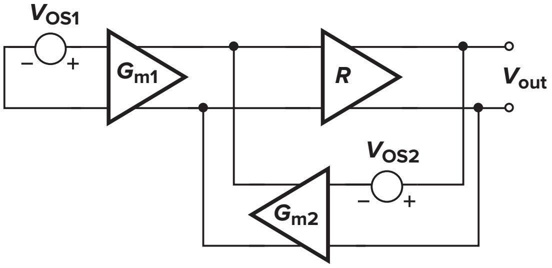


Figure 14.36 Circuit of Fig. 14.35(a) including offset of G_{m2}.

This voltage is stored on C_1 and C_2 after S_3 and S_4 turn off. The offset voltage referred to the main input is therefore given by

$$V_{OS,tot} = \frac{V_{out}}{G_{m1}R} \tag{14.93}$$

$$= \frac{V_{OS1}}{1+G_{m2}R} + \frac{G_{m2}}{G_{m1}}\frac{V_{OS2}}{1+G_{m2}R} \tag{14.94}$$

$$\approx \frac{V_{OS1}}{G_{m2}R} + \frac{V_{OS2}}{G_{m1}R} \tag{14.95}$$

where we have assumed that $G_{m2}R \gg 1$. If $G_{m2}R$ and $G_{m1}R$ are large, as in the op amp of Fig. 14.35(b), then $V_{OS,tot}$ is very small.

The offset cancellation of Fig. 14.35 warrants a cautionary note. Upon turning off, S_3 and S_4 may inject slightly unequal charges onto C_1 and C_2, respectively, creating an error voltage that is *not* corrected because the feedback loop is opened. The reader can prove that for a differential injection-induced error voltage of ΔV, the resulting input-referred offset voltage equals $(G_{m2}/G_{m1})\Delta V$. For this reason, G_{m2} is usually chosen to be on the order of $0.1G_{m1}$.

We should also mention that the unity-gain and precision multiply-by-two circuits described in Chapter 13 cancel the offset of the op amp as well. The proof is left to the reader.[3]

It is important to note that the offset cancellation techniques studied here require periodic refreshing because the junction and subthreshold leakage of the switches eventually corrupts the correction voltage stored across the capacitors. In a typical design, the offset must be refreshed at a rate of at least a few kilohertz.

14.2.3 Reduction of Noise by Offset Cancellation

Recall from previous sections that the offset of a differential amplifier can be viewed as a noise component having a very low frequency. We therefore expect that periodic offset cancellation can potentially reduce the (low-frequency) noise of the circuit as well.

Consider a simple differential amplifier that is to be used in the front end of a sampling system [Fig. 14.37(a)]. Here, the noise of A_1 directly corrupts V_{in}. The $1/f$ noise of A_1 proves especially problematic if the signal spectrum extends from zero to only a few megahertz, because the $1/f$ noise corner frequency is typically around 500 kHz to 1 MHz.

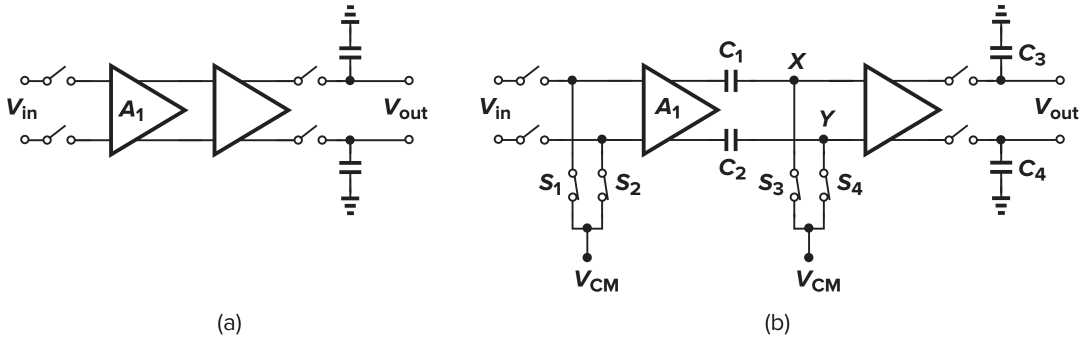


Figure 14.37 (a) Front end of a sampler; (b) circuit of (a) with offset cancellation applied to the first stage.

Now suppose the amplifier undergoes offset cancellation before *every* sampling operation [Fig. 14.37(b)]. That is, as depicted in Fig. 14.38, the input is disabled; the offset of A_1 is stored on C_1 and C_2; the input is enabled and amplified by A_1 and A_2 and stored on C_3 and C_4; and finally the sampling switches are turned off. How does the noise of A_1 affect the final output? Denoting the time elapsed from the end of offset cancellation to the end of sampling by $\Delta t = t_2 - t_1$, we recall that at $t = t_1$, $V_{XY} = 0$. Thus, from t_1 to t_2, only *high-frequency* noise components of A_1, on the order of $> 1/\Delta t$, change V_{XY} significantly. In other words, offset cancellation suppresses noise frequencies below roughly $1/\Delta t$.

To better understand this concept, let us consider a numerical example. Assuming that $\Delta t = 10$ ns, we examine two noise components, one at 1 MHz and another at 10 MHz, approximating each with a sinusoid (Fig. 14.39). For a sinusoid of amplitude A and frequency f, the maximum slew rate is equal to $2\pi f A$,

[3] If, as shown in Fig. 13.35, an equalizing switch is added to the circuit, then the op amp offset may not be removed.

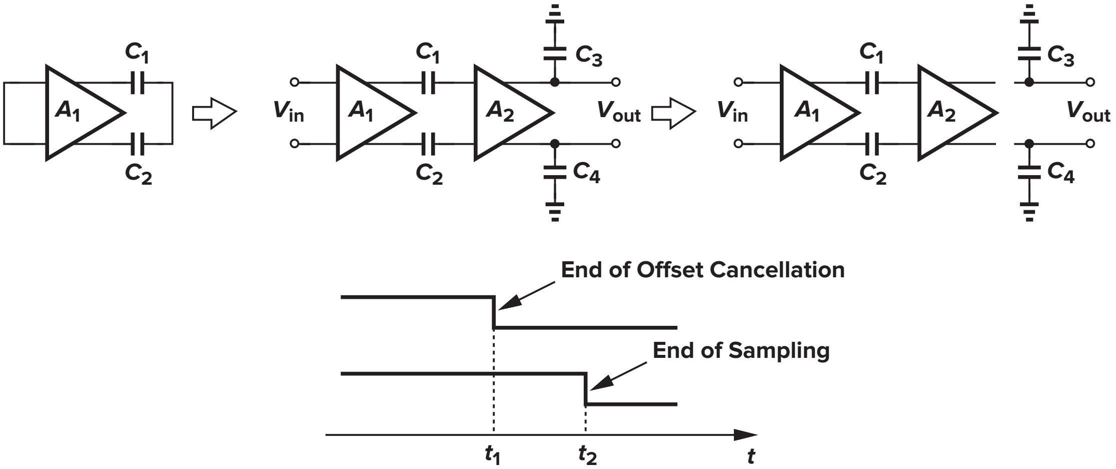


Figure 14.38 Sequence of operations in the sampler.

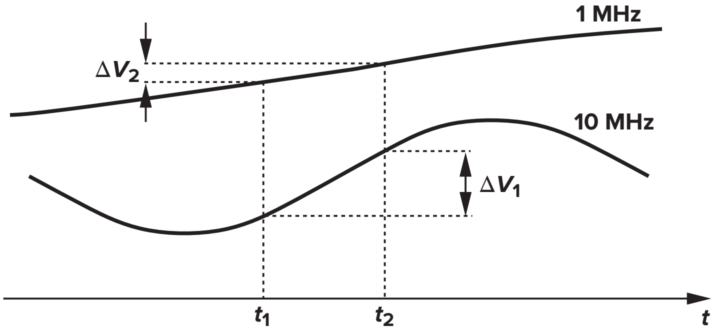


Figure 14.39 Variation of 1-MHz and 10-MHz noise components in a time interval of 10 ns.

and hence the maximum variation in Δt seconds is $2\pi f A \Delta t$. Normalizing this value to the amplitude, we obtain the change for the 1-MHz and 10-MHz components as $\Delta V_1/A = 6.3\%$ and $\Delta V_2/A = 63\%$, respectively. We therefore conclude that noise frequencies below a few megahertz do not have sufficient time to change if the sampling occurs only 10 ns after the end of offset cancellation.

Originally utilized in charge-coupled devices (CCDs), the foregoing property of offset cancellation is called "correlated double sampling" (CDS) because it involves two consecutive sampling operations (the first being offset storage) that are so tightly spaced in time that they do not allow (low-frequency) noise components to vary significantly. A powerful technique, CDS finds wide usage in suppressing the $1/f$ noise of MOS circuits. Nonetheless, it leads to aliasing of wideband noise [5].

14.2.4 Alternative Definition of CMRR

Recall from Chapter 4 that common-mode rejection is represented by the change in the differential output divided by the change in the input common-mode level, and the CMRR is defined as the differential gain divided by this quantity. We also noted that in fully differential circuits, the finite output impedance of the tail current source and asymmetries limit the common-mode rejection.

Now consider a differential circuit sensing an input CM change, $\Delta V_{in,CM}$. If the differential output voltage changes by ΔV_{out} while the differential input voltage is zero, we can say that the output *offset* voltage of the circuit has changed by ΔV_{out}. In other words, common-mode rejection can be viewed as the change in the output offset divided by the change in the input CM level. Following the notation in

Chapter 4, we write

$$A_{CM-DM} = \frac{\Delta V_{OS,out}}{\Delta V_{CM,in}} \tag{14.96}$$

Since CMRR $= A_{DM}/A_{CM-DM}$, we have

$$\text{CMRR} = \frac{A_{DM}}{\dfrac{\Delta V_{OS,out}}{\Delta V_{CM,in}}} \tag{14.97}$$

$$= \frac{\Delta V_{CM,in}}{\dfrac{\Delta V_{OS,out}}{A_{DM}}} \tag{14.98}$$

Noting that $\Delta V_{OS,out}/A_{DM}$ is in fact the input-referred offset voltage, we have

$$\text{CMRR} = \frac{\Delta V_{CM,in}}{\Delta V_{OS,in}} \tag{14.99}$$

The above result proves useful in analyzing the behavior of circuits. For example, suppose an op amp incorporates a PMOS differential pair at the input. Which one of the topologies shown in Fig. 14.40 yields a higher CMRR? In Fig. 14.40(a), body effect is eliminated and the threshold voltages of M_1 and M_2 are independent of the input CM level. In Fig. 14.40(b), on the other hand, M_1 and M_2 experience body effect and if they suffer from mismatches in their body effect coefficients, then the difference between V_{TH1} and V_{TH2}, i.e., the input offset voltage, *varies* with the input CM level, degrading the common-mode rejection.

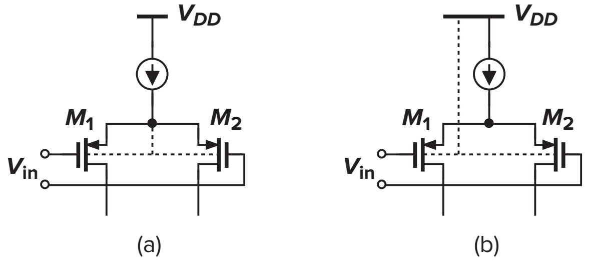


Figure 14.40 PMOS differential pair (a) without and (b) with body effect.

References

[1] F. Krummenacher and N. Joehl, "A 4-MHz CMOS Continuous-Time Filter with On-Chip Automatic Tuning," *IEEE J. of Solid-State Circuits*, vol. 23, pp. 750–758, June 1988.

[2] K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatches in MOS Transistors for Precision Analog Design," *IEEE J. of Solid-State Circuits*, vol. 21, pp. 1057–1066, December 1986.

[3] M. J. M. Pelgrom, A. C. J. Duinmaiger, and A. P. G. Welbers, "Matching Properties of MOS Transistors," *IEEE J. of Solid-State Circuits*, vol. SC-24, pp. 1433–1439, October 1989.

[4] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor Matching in Analog CMOS Applications," *IEDM Dig. of Tech. Papers*, pp. 34.1.1–34.1.4, December 1998.

[5] C. C. Enz and G. C. Temes, "Circuit Techniques for Reducing the Effects of Op-Amp Imperfections: Autozeroing, Correlated Double Sampling, and Chopper Stabilization," *Proc. IEEE*, vol. 84, pp. 1584–1614, November 1996.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

14.1. The input-output characteristic of an amplifier is approximated as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t)$ in the range $x = [0 \;\; x_{max}]$.
(a) What is the maximum nonlinearity?
(b) What is the THD for $x(t) = (x_{max} \cos \omega t + x_{max})/2$.

14.2. In the circuits of Fig. 14.6, $W/L = 20/0.5$ and $I = 0.5$ mA. Calculate the harmonic distortion in each circuit if the input signal has a peak amplitude of 100 mV. How do the results change if we double W/L or I?

14.3. For the circuits of Fig. 14.6, plot the THD and the input-referred thermal noise as a function of **(a)** W/L, **(b)** I. Identify the trade-offs among noise, linearity, and power dissipation.

14.4. In Fig. 14.6, *two* effects lead to a trade-off between nonlinearity and voltage gain. Describe these effects.

14.5. The circuit of Fig. 14.6(a) is designed with $W/L = 50/0.5$, $I = 1$ mA, and $R_D = 2$ kΩ. The circuit is placed in a feedback loop similar to that of Fig. 14.7 with $\beta = 0.2$ and senses an input sinusoid with a peak amplitude of 10 mV. Calculate the THD at the output.

14.6. Suppose that in Fig. 14.16, A_1 and A_2 have an input-referred noise voltage V_n. Neglecting other sources of noise, calculate the input-referred noise voltage of the overall circuit.

14.7. Equation 14.36 suggests that if the open-loop gain, α_1, increases while other parameters remain constant, then the harmonic distortion drops sharply. Repeat Problem 14.14.5 with $W/L = 200/0.5$ to achieve a higher open-loop gain and explain the results.

14.8. Equation 14.36 suggests that if $\beta\alpha_1 \gg 1$, then $b/a \propto \beta^{-2}$. Repeat Problem 14.14.5 with $\beta = 0.4$.

14.9. Suppose the nonlinear feedforward amplifier in Fig. 14.7 is characterized by $y(t) = \alpha_1 x(t) + \alpha_3 x^3(t)$. Estimate the magnitude of the third harmonic at the output of the overall system.

14.10. As mentioned in Chapter 2, MOS devices operating in the subthreshold region exhibit an exponential behavior: $I_D = I_0 \exp[V_{GS}/(\zeta V_T)]$. Suppose both of the circuits shown in Fig. 14.6 operate in the subthreshold region. Derive expressions for the harmonic amplitudes if the input signal is much less than ζV_T. For the differential pair, first prove that $I_{D1} - I_{D2} \propto \tanh[V_{in}/(2\zeta V_T)]$ and then write the Taylor expansion of the hyperbolic tangent.

14.11. The mobility of MOSFETs is in fact a function of the gate-source voltage and expressed as $\mu = \mu_0/[1 + \theta(V_{GS} - V_{TH})]$, where θ is an empirical factor (Chapter 17). Assuming that $\theta(V_{GS} - V_{TH}) \ll 1$ and using the relationship $(1+\epsilon)^{-1} \approx 1 - \epsilon$ for $\epsilon \ll 1$, calculate the third harmonic in the circuit of Fig. 14.6(a).

14.12. The input devices of a differential pair have an effective length of 0.5 μm.
(a) Assuming that $\Delta V_{TH} = 0.1 t_{ox}/\sqrt{WL}$ and neglecting other mismatches, determine the minimum width of the transistors such that $V_{OS} \leq 5$ mV.
(b) If the tail current is 1 mA, what is the maximum input swing that gives a THD of 1%?

14.13. Repeat Problem 14.14.12 if the tolerable input offset is 2 mV and compare the results.

14.14. Determine the dimensions of M_1 and M_2 in Fig. 14.28 such that $I_{D1} \approx I_{D2} = 0.5$ mA, $\Delta I_D/I_D = 2\%$, and $V_{GS} - V_{TH} = 0.5$ V. Assume that $\Delta V_{TH} = 0.1 t_{ox}/\sqrt{WL}$ and neglect other mismatches.

14.15. Source degeneration can improve the matching between current sources if resistor mismatches are small. Prove that in the circuit of Fig. 14.41,

$$\frac{\Delta I_D}{I_D} = \frac{1}{1 + g_m R_S}\left[\frac{\Delta(\mu_n C_{ox})}{\mu_n C_{ox}} + \frac{\Delta(W/L)}{(W/L)} - \frac{2\Delta V_{TH}}{V_{GS} - V_{TH}} - g_m \Delta R_S\right] \tag{14.100}$$

where ΔR_S denotes the mismatch between R_{S1} and R_{S2}. Note that for an appreciable reduction of $\Delta I/I_D$, R_S must be greater than $1/g_m$.

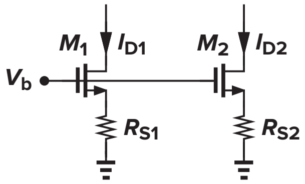


Figure 14.41

14.16. In the circuit of Fig. 14.29, assume that $\alpha_j = \beta_j$ but $x_1(t) = A\cos\omega t$ and $x_2(t) = A\cos(\omega t + \theta)$, where θ denotes a small phase mismatch. Calculate the magnitude of the second harmonic at the output.

14.17. In the circuit of Fig. 14.42, M_3 and M_4 suffer from a threshold mismatch of ΔV_{TH} and the circuit is otherwise symmetric. Assuming that $\lambda \neq 0$ but $\gamma = 0$, calculate the input-referred offset voltage. What happens as $R_D \to \infty$?

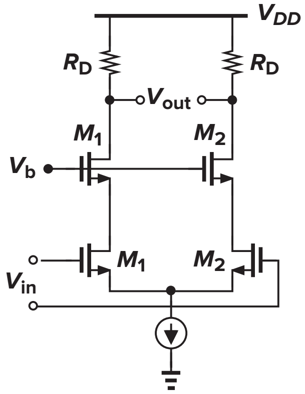


Figure 14.42

14.18. In the circuit of Fig. 14.32, the amplifier has an input capacitance (between X and Y) equal to C_{in}. Calculate the input offset voltage after offset compensation.

14.19. The circuit of Fig. 14.32 is designed for an input offset voltage of 1 mV. If the width of the transistors in the input differential pair of the amplifier is doubled, what is the overall input offset voltage? (Neglect the input capacitance of the amplifier.)

14.20. Explain why the circuit of Fig. 14.27 suffers from a trade-off between the input offset and the output voltage swing (for a given tail current).

CHAPTER

15

Oscillators

Oscillators are an integral part of many electronic systems. Applications range from clock generation in microprocessors to carrier synthesis in cellular telephones, requiring vastly different oscillator topologies and performance parameters. Robust, high-performance oscillator design in CMOS technology continues to pose interesting challenges. As described in Chapter 16, oscillators are usually embedded in a phase-locked system.

This chapter deals with the analysis and design of CMOS oscillators, more specifically, voltage-controlled oscillators (VCOs). Beginning with a general study of oscillation in feedback systems, we introduce ring oscillators and LC oscillators along with methods of varying the frequency of oscillation. We then describe a mathematical model of VCOs that will be used in the analysis of PLLs in Chapter 16.

15.1 ■ General Considerations

A simple oscillator produces a periodic output, usually in the form of voltage. As such, the circuit has no input while sustaining the output indefinitely. How can a circuit oscillate? Recall from Chapter 10 that negative-feedback systems may oscillate, i.e., an oscillator is a badly-designed feedback amplifier![1] Consider the unity-gain negative-feedback circuit shown in Fig. 15.1, where

$$\frac{V_{out}}{V_{in}}(s) = \frac{H(s)}{1+H(s)} \tag{15.1}$$

As mentioned in Chapter 10, if the amplifier itself experiences so much phase shift at high frequencies that the overall feedback becomes positive, then oscillation may occur. More accurately, if for $s = j\omega_0$, $H(j\omega_0) = -1$, then the closed-loop gain approaches infinity at ω_0. Under this condition, the circuit amplifies its own noise components at ω_0 indefinitely. In fact, as conceptually illustrated in Fig. 15.2, a noise component at ω_0 experiences a total gain of unity and a phase shift of 180°, returning to the subtractor as a negative replica of the input. Upon subtraction, the input and the feedback signals give a larger difference. Thus, the circuit continues to "regenerate," allowing the component at ω_0 to grow.

[1] It is said, "In the high-frequency world, amplifiers oscillate and oscillators don't."

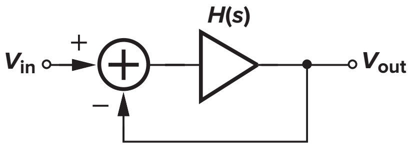


Figure 15.1 Feedback system.

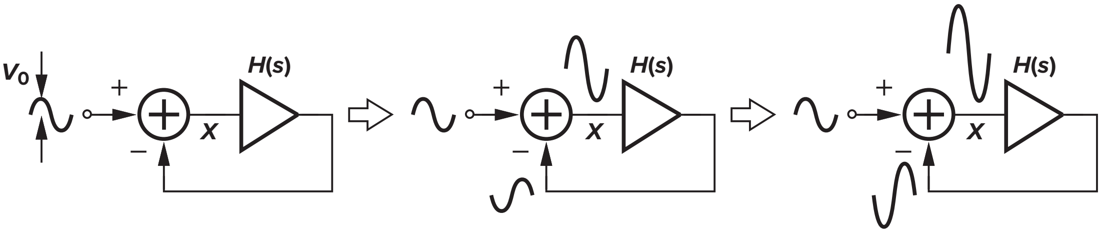


Figure 15.2 Evolution of oscillatory system with time.

For the oscillation to begin, a loop gain of unity or greater is necessary. This can be seen by following the signal around the loop over many cycles and expressing the amplitude of the subtractor's output in Fig. 15.2 as a geometric series (if $\angle H(j\omega_0) = 180°$):

$$V_X = V_0 + |H(j\omega_0)|V_0 + |H(j\omega_0)|^2 V_0 + |H(j\omega_0)|^3 V_0 + \cdots \tag{15.2}$$

If $|H(j\omega_0)| > 1$, the above summation diverges, whereas if $|H(j\omega_0)| < 1$, then

$$V_X = \frac{V_0}{1 - |H(j\omega_0)|} < \infty \tag{15.3}$$

In summary, if a negative-feedback circuit has a loop gain that satisfies two conditions:

$$|H(j\omega_0)| \geq 1 \tag{15.4}$$

$$\angle H(j\omega_0) = 180° \tag{15.5}$$

then the circuit may oscillate at ω_0. Called "Barkhausen criteria," these conditions are necessary but not sufficient [1].[2] In order to ensure oscillation in the presence of temperature and process variations, we typically choose the loop gain to be at least twice or three times the required value.

We may state the second Barkhausen criterion as $\angle H(j\omega) = 180°$ or a *total* phase shift of 360°. This should not be confusing: if the system is designed to have low-frequency negative feedback, it already produces 180° of phase shift in the signal traveling around the loop (as represented by the subtractor in Fig. 15.1), and $\angle H(j\omega) = 180°$ denotes an additional *frequency-dependent* phase shift that, as illustrated in Fig. 15.2, ensures that the feedback signal *enhances* the original signal. Thus, the three cases illustrated in Fig. 15.3 are equivalent in terms of the second criterion. We say that the system of Fig. 15.3(a) exhibits a frequency-dependent phase shift of 180° (denoted by the arrow) and a dc phase shift of 180°. The difference between Figs. 15.3(b) and (c) is that the open-loop amplifier in the former contains enough stages with proper polarities to provide a total phase shift of 360° at ω_0, whereas that in the latter produces *no* phase shift at ω_0. Examples of these topologies are presented later in this chapter.

[2] We only know that, if the gain crossover frequency is less than the phase crossover frequency, then the system is stable.

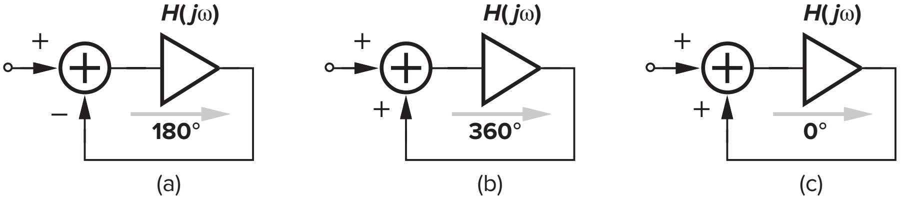


Figure 15.3 Various views of oscillatory feedback system.

CMOS oscillators in today's technology are typically implemented as "ring oscillators" or "LC oscillators." We study each type in the following sections.

15.2 ■ Ring Oscillators

A ring oscillator consists of a number of gain stages in a loop. To arrive at the actual implementation, we begin by attempting to make a single-stage feedback circuit oscillate.

▶ Example 15.1

Explain why a single common-source stage does not oscillate if it is placed in a unity-gain loop.

Solution

From Fig. 15.4, it is seen that the open-loop circuit contains only one pole, thereby providing a maximum frequency-dependent phase shift of 90° (at a frequency of infinity). Since the common-source stage exhibits a dc phase shift of 180° due to the signal inversion from the gate to the drain, the maximum total phase shift is 270°. The loop therefore fails to sustain oscillation growth.

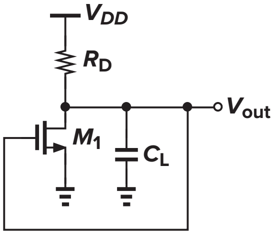


Figure 15.4

The above example suggests that oscillation may occur if the circuit contains multiple stages and hence multiple poles. Indeed, such a topology was considered *undesirable* in Chapter 10 because it led to inadequate phase margin in op amps. We therefore surmise that if the circuit of Fig. 15.4 is modified as shown in Fig. 15.5, then two significant poles appear in the signal path, allowing the frequency-dependent

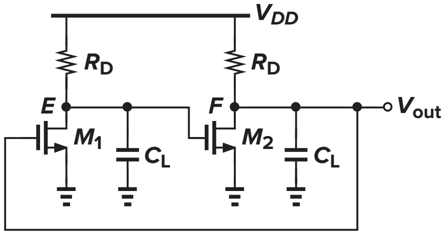


Figure 15.5 Two-pole feedback system.

phase shift to approach 180°. Unfortunately, this circuit exhibits *positive* feedback near zero frequency due to the signal inversion through each common-source stage. As a result, it simply "latches up" rather than oscillates. That is, if V_E rises, V_F falls, thereby turning M_1 off and allowing V_E to rise further. This may continue until V_E reaches V_{DD} and V_F drops to near zero, a state that will remain indefinitely.

To gain more insight into the oscillation conditions, let us assume that an ideal inverting stage (with zero phase shift at all frequencies) is inserted in the loop of Fig. 15.5, providing *negative* feedback near zero frequency and eliminating the problem of latch-up (Fig. 15.6). Does this circuit oscillate? We note that the loop contains only two poles: one at E and another at F. The frequency-dependent phase shift can therefore reach 180°, but at a frequency of infinity. Since the loop gain vanishes at very high frequencies, we observe that the circuit does not satisfy both of Barkhausen's criteria at the same frequency (Fig. 15.7), and thus fails to oscillate.

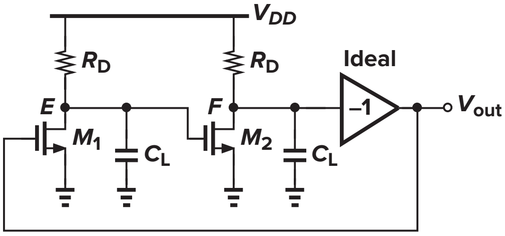


Figure 15.6 Two-pole feedback system with additional signal inversion.

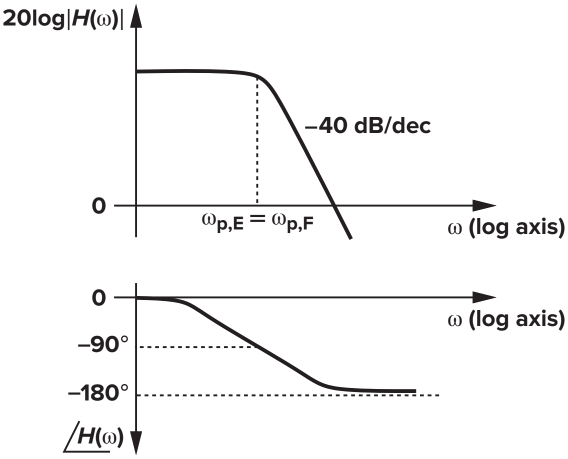


Figure 15.7 Loop gain characteristics of a two-pole system.

The foregoing discussion points to the need for greater phase shift around the loop, suggesting the possibility of oscillation if the third inverting stage in Fig. 15.6 contains a pole that contributes significant phase. We then arrive at the topology depicted in Fig. 15.8. If the three stages are identical, the total phase shift around the loop, ϕ, reaches $-135°$ at $\omega = \omega_{p,E}(= \omega_{p,F} = \omega_{p,G})$ and $-270°$ at $\omega = \infty$. Consequently, ϕ equals $-180°$ at $\omega < \infty$, where the loop gain can still be greater than or equal to unity. This circuit indeed oscillates if the loop gain is sufficient and it is an example of a ring oscillator.

It is instructive to calculate the minimum voltage gain per stage in Fig. 15.8 that is necessary for oscillation. Neglecting the effect of the gate-drain overlap capacitance and denoting the transfer function of each stage by $-A_0/(1 + s/\omega_0)$, we have for the loop gain

$$H(s) = -\frac{A_0^3}{\left(1 + \frac{s}{\omega_0}\right)^3} \tag{15.6}$$

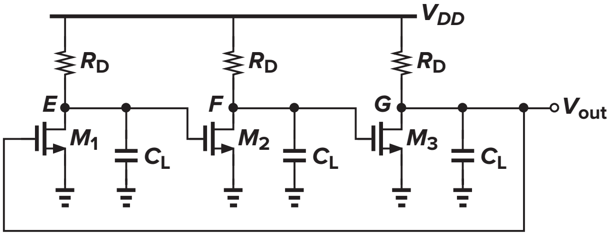


Figure 15.8 Three-stage ring oscillator.

The circuit oscillates only if the frequency-dependent phase shift equals 180°, i.e., if each stage contributes 60°. The frequency at which this occurs is given by

$$\tan^{-1}\frac{\omega_{osc}}{\omega_0} = 60^\circ \tag{15.7}$$

and hence

$$\omega_{osc} = \sqrt{3}\omega_0 \tag{15.8}$$

The minimum voltage gain per stage must be such that the magnitude of the loop gain at ω_{osc} is equal to unity:

$$\frac{A_0^3}{\left[\sqrt{1+(\frac{\omega_{osc}}{\omega_0})^2}\right]^3} = 1 \tag{15.9}$$

It follows from (15.8) and (15.9) that

$$A_0 = 2 \tag{15.10}$$

In summary, a three-stage ring oscillator requires a low-frequency gain of 2 per stage, and it oscillates at a frequency of $\sqrt{3}\omega_0$, where ω_0 is the 3-dB bandwidth of each stage.

Let us now examine the waveforms at the three nodes of the oscillator of Fig. 15.8. Since each stage contributes a frequency-dependent phase shift of 60° as well as a low-frequency signal inversion, the waveform at each node is 240° (or 120°) out of phase with respect to its neighboring nodes (Fig. 15.9). The ability to generate multiple phases is a very useful property of ring oscillators.

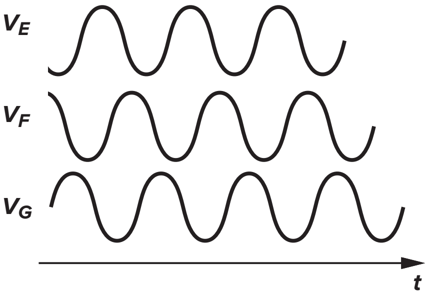


Figure 15.9 Waveforms of a three-stage ring oscillator.

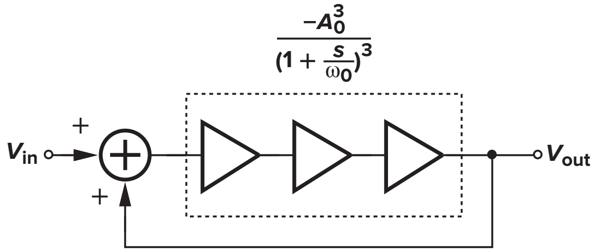


Figure 15.10 Linear model of three-stage ring oscillator.

Amplitude Limiting The natural question at this point is—What happens if in the three-stage ring of Fig. 15.8, $A_0 \neq 2$? We know from Barkhausen's criteria that if $A_0 < 2$, the circuit fails to oscillate, but what if $A_0 > 2$? To answer this question, we first model the oscillator by a linear feedback system, as depicted in Fig. 15.10. Note that the feedback is positive (i.e., V_{out} is *added* to V_{in}) because $H(s)$ in Eq. (15.6) already includes the negative polarity resulting from three inversions in the signal path. The closed-loop transfer function is

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{\dfrac{-A_0^3}{(1+s/\omega_0)^3}}{1+\dfrac{A_0^3}{(1+s/\omega_0)^3}} \tag{15.11}$$

$$= \frac{-A_0^3}{(1+s/\omega_0)^3 + A_0^3} \tag{15.12}$$

The denominator of (15.12) can be expanded as

$$\left(1+\frac{s}{\omega_0}\right)^3 + A_0^3 = \left(1+\frac{s}{\omega_0}+A_0\right)\left[\left(1+\frac{s}{\omega_0}\right)^2 - \left(1+\frac{s}{\omega_0}\right)A_0 + A_0^2\right] \tag{15.13}$$

Thus, the closed-loop system exhibits three poles:

$$s_1 = (-A_0 - 1)\omega_0 \tag{15.14}$$

$$s_{2,3} = \left[\frac{A_0(1 \pm j\sqrt{3})}{2} - 1\right]\omega_0 \tag{15.15}$$

Since A_0 itself is positive, the first pole leads to a decaying exponential term: $\exp[(-A_0-1)\omega_0 t]$, which can be neglected in the steady state. Figure 15.11 illustrates the locations of the poles for different values of A_0, revealing that for $A_0 > 2$, the two complex poles exhibit a positive real part and hence give rise to a growing sinusoid. Neglecting the effect of s_1, we express the output waveform as

$$V_{out}(t) = a \exp\left(\frac{A_0 - 2}{2}\omega_0 t\right) \cos\left(\frac{A_0\sqrt{3}}{2}\omega_0 t\right) \tag{15.16}$$

Thus, if $A_0 > 2$, the exponential envelope grows to infinity.

In practice, as the oscillation amplitude increases, the stages in the signal path experience nonlinearity and eventually "saturation," limiting the maximum amplitude. We may say that the poles begin in the right half plane and eventually move to the imaginary axis to stop the growth. If the small-signal loop

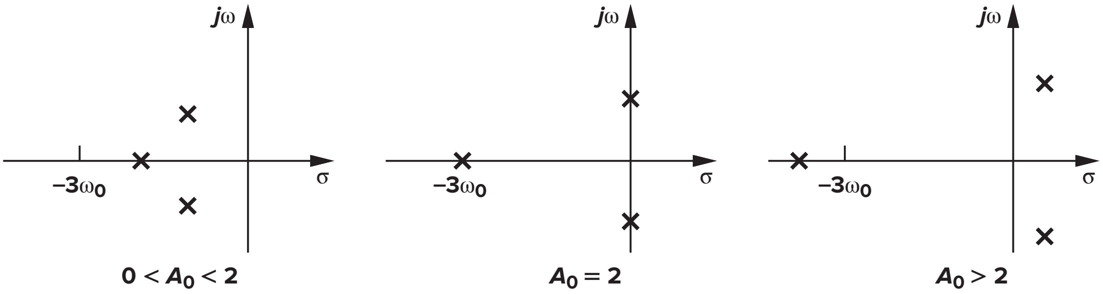


Figure 15.11 Poles of three-stage ring oscillator for various values of gain.

gain is greater than unity, the circuit must spend enough time in saturation so that the "average" loop gain is still equal to unity.[3]

▶ Example 15.2

Shown in Fig.15.12 is a differential implementation of the oscillator of Fig. 15.8. What is the maximum voltage swing of each stage?

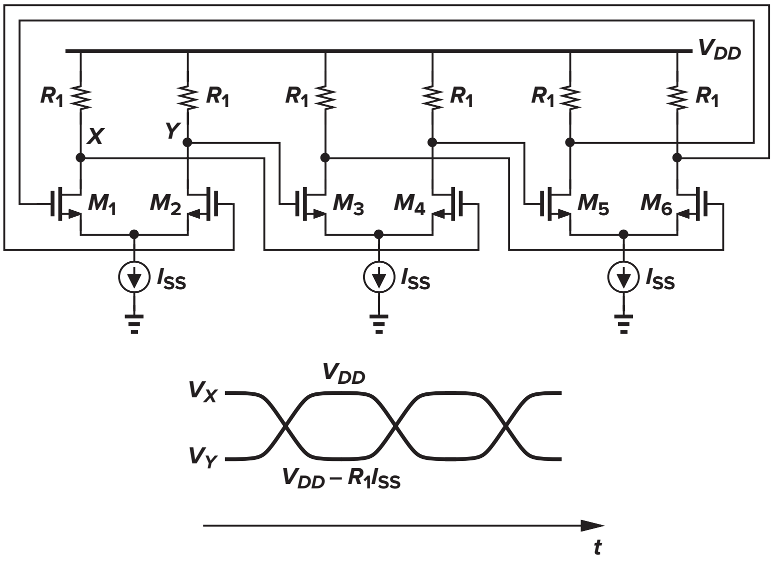


Figure 15.12

Solution

If the gain per stage is well above 2, then the amplitude grows until each differential pair experiences complete switching, that is, until I_{SS} is completely steered to one side every half cycle. As a result, the swing at each node is equal to $I_{SS}R_1$. From the waveforms shown in Fig. 15.12, we also observe that each stage is in its high-gain region for only a fraction of the period, (e.g., when $|V_X - V_Y|$ is small).

[3] While intuitive, these statements are not rigorous. The concepts of transfer function, poles, and loop gain are difficult to apply to a nonlinear circuit.

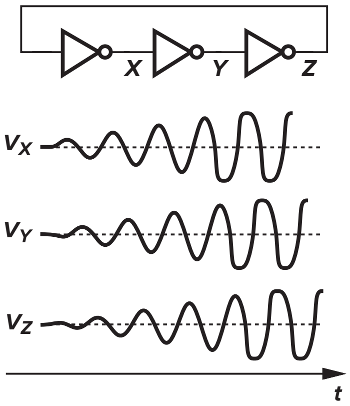


Figure 15.13 Ring oscillator using CMOS inverters.

A simple implementation of ring oscillators that does not require resistors is depicted in Fig. 15.13. Suppose the circuit is released with an initial voltage at each node equal to the trip point of the inverters, V_{trip}.[4] With identical stages and no noise in the devices, the circuit would remain in this state indefinitely,[5] but noise components disturb each node voltage, yielding a growing waveform. The signal eventually exhibits rail-to-rail swings.

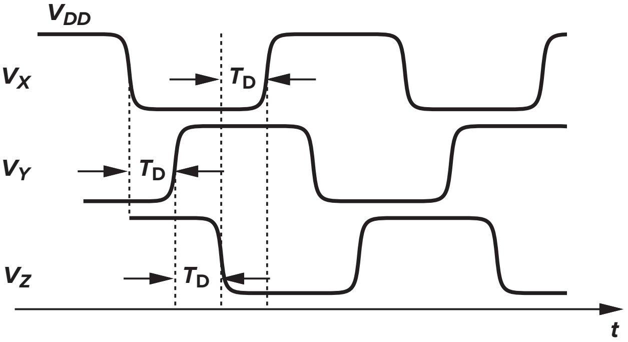


Figure 15.14 Waveforms of ring oscillator when one node is initialized at V_{DD}.

Let us now assume that the circuit of Fig. 15.13 begins with $V_X = V_{DD}$ (Fig. 15.14). Under this condition, $V_Y = 0$ and $V_Z = V_{DD}$. Thus, when the circuit is released, V_X begins to fall to zero (because the first inverter senses a high input), forcing V_Y to rise to V_{DD} after one inverter delay, T_D, and V_Z to fall to zero after another inverter delay. The circuit therefore oscillates with a delay of T_D between consecutive node voltages, yielding a period of $6T_D$.

The above small-signal and large-signal analyses raise an interesting question. While the small-signal oscillation frequency is given by $A_0\sqrt{3}\omega_0/2$ [from Eq. (15.16)], the large-signal value is $1/(6T_D)$. Are these two values equal? Not necessarily. After all, ω_0 is determined by the small-signal output resistance and capacitance of each inverter near the trip point, whereas T_D results from the large-signal, nonlinear

[4] The trip point of an inverter is the input voltage that results in an equal output voltage.

[5] This is indeed how SPICE predicts the circuit's behavior. To start the oscillation in SPICE, one of the nodes must be initialized at a different voltage.

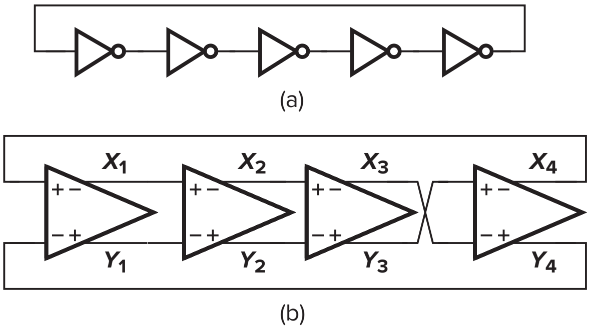


Figure 15.15 (a) Five-stage single-ended ring oscillator; (b) four-stage differential ring oscillator.

current drive and capacitances of each stage. In other words, when the circuit is released with all inverters at their trip point, the oscillation begins with a frequency of $\sqrt{3}A_0\omega_0/2$, but, as the amplitude grows and the circuit becomes nonlinear, the frequency shifts to $1/(6T_D)$ (which is a lower value).

Ring oscillators employing more than three stages are also feasible. The total number of inversions in the loop must be odd so that the circuit does not latch up. For example, as shown in Fig. 15.15(a), a ring can incorporate five inverters, providing a frequency of $1/(10T_D)$. On the other hand, the differential implementation can utilize an *even* number of stages by simply configuring one stage such that it does not invert. Illustrated in Fig. 15.15(b), this flexibility demonstrates another advantage of differential circuits over their single-ended counterparts.

▶ Example 15.3

What is the minimum required voltage gain per stage in the four-stage oscillator of Fig. 15.15(b)? How many signal phases are provided by the circuit?

Solution

Using a notation similar to that for Fig. 15.8, we have

$$H(s) = -\frac{A_0^4}{\left(1+\dfrac{s}{\omega_0}\right)^4} \tag{15.17}$$

For the circuit to oscillate, each stage must contribute a frequency-dependent phase shift of $180°/4 = 45°$. The frequency at which this occurs is given by $\tan^{-1}\omega_{osc}/\omega_0 = 45°$ and hence $\omega_{osc} = \omega_0$. The minimum voltage gain is therefore derived as

$$\frac{A_0}{\sqrt{1+\left(\dfrac{\omega_{osc}}{\omega_0}\right)^2}} = 1 \tag{15.18}$$

That is, $A_0 = \sqrt{2}$. As expected, this value is lower than that required in a three-stage ring.

With 45° of phase shift per stage, the oscillator provides four phases and their complements. This is illustrated in Fig. 15.16.

◀

The number of stages in a ring oscillator is determined by various requirements, including speed, power dissipation, noise immunity, etc. In most applications, three to five stages provide optimum performance (for differential implementations).

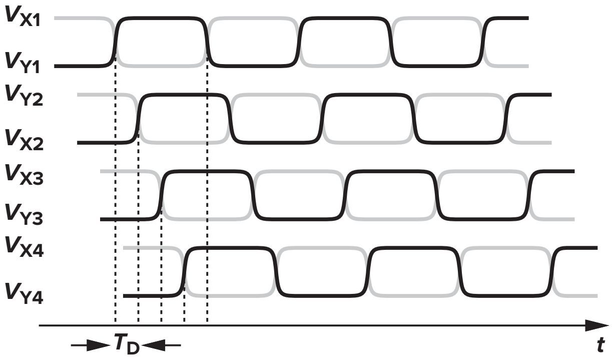


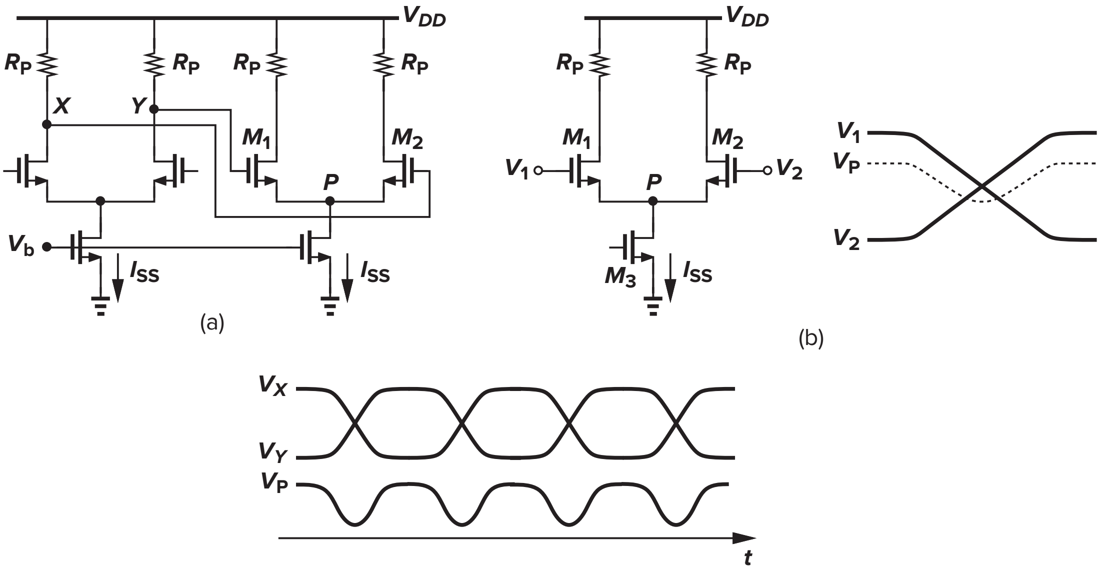
Figure 15.16

▶ Example 15.4

Determine the maximum voltage swings and the minimum supply voltage of a ring oscillator incorporating differential pairs with resistive loads (e.g., as in Fig. 15.12) if no transistor must enter the triode region. Assume that each stage experiences complete switching.

Solution

Figure 15.17(a) shows two stages in cascade. If each stage experiences complete switching, then each drain voltage, e.g., V_X or V_Y, varies between V_{DD} and $V_{DD} - I_{SS}R_P$. Thus, when M_1 is fully on, its gate and drain voltages are equal to V_{DD} and $V_{DD} - I_{SS}R_P$, respectively. For this transistor to remain in saturation, we have $I_{SS}R_P \leq V_{TH}$, i.e., the peak-to-peak swing at each drain must not exceed V_{TH}.



(a)

(b)

(c)

Figure 15.17

How is the minimum supply voltage determined? If V_{DD} is lowered, the voltage at the common source node of each differential pair, e.g., V_P in Fig. 15.17(a), falls, eventually driving the tail transistor into the triode region. We must therefore calculate V_P for the worst case, noting that V_P does vary with time because M_1 and M_2 carry unequal currents when the input difference becomes large.

Now consider the stand-alone circuit of Fig. 15.17(b), assuming that the inputs vary between V_{DD} and $V_{DD} - I_{SS}R_P$. How does V_P vary? When the gate voltage of M_1, V_1, is equal to V_{DD} and M_1 carries all of I_{SS},

$$V_P = V_{DD} - \sqrt{\frac{2I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} - V_{TH} \tag{15.19}$$

As V_1 falls and V_2 rises, so does V_P because, so long as M_2 is off, M_1 operates as a source follower. When the difference between V_1 and V_2 reaches $\sqrt{2}(V_{GS,eq} - V_{TH})$, where $V_{GS,eq}$ denotes the equilibrium overdrive of each transistor, M_2 turns on. To calculate V_P after this point, we note that $I_{D1} + I_{D2} = I_{SS}$, $V_{GS1} = V_1 - V_P$, and $V_{GS2} = V_2 - V_P$. Thus,

$$\frac{1}{2}\mu_n C_{ox}\Big(\frac{W}{L}\Big)_{1,2}(V_1 - V_P - V_{TH})^2 + \frac{1}{2}\mu_n C_{ox}\Big(\frac{W}{L}\Big)_{1,2}(V_2 - V_P - V_{TH})^2 = I_{SS} \tag{15.20}$$

Expanding the quadratic terms and rearranging the result, we have

$$2V_P^2 - 2(V_1 - V_{TH} + V_2 - V_{TH})V_P + (V_1 - V_{TH})^2 + (V_2 - V_{TH})^2 - \frac{2I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}} = 0 \tag{15.21}$$

It follows that

$$V_P = \frac{1}{2}\Big[V_1 + V_2 - 2V_{TH} \pm \sqrt{-(V_1 - V_2)^2 + \frac{4I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}}\Big] \tag{15.22}$$

If V_1 and V_2 vary differentially, they can be expressed as $V_1 = V_{CM} + \Delta V$ and $V_2 = V_{CM} - \Delta V$, where $V_{CM} = V_{DD} - I_{SS}R_P/2$, yielding

$$V_P = V_{CM} - V_{TH} \pm \frac{1}{2}\sqrt{-(2\Delta V)^2 + \frac{4I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} \tag{15.23}$$

This expression reveals why node P is considered a virtual ground in small-signal operation: if $|\Delta V|$ is much less than the maximum overdrive voltage, then V_P is relatively constant. Since the term under the square root reaches a maximum for $\Delta V = 0$ (equilibrium condition),

$$V_{P,min} = V_{CM} - V_{TH} - \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} \tag{15.24}$$

As expected, the last term in (15.24) represents the overdrive voltage of each transistor in equilibrium (where $I_{D1} = I_{D2} = I_{SS}/2$).

Figure 15.17(c) shows typical waveforms in the oscillator. Note that V_P varies at twice the oscillation frequency. This property is sometimes exploited in "frequency doublers."

To determine the minimum supply voltage, we write $V_{P,min} \geq V_{ISS}$, where V_{ISS} denotes the minimum required voltage across I_{SS}. Thus,

$$V_{DD} - \frac{R_P I_{SS}}{2} - V_{TH} - \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} \geq V_{ISS} \tag{15.25}$$

and

$$V_{DD} \geq V_{ISS} + V_{TH} + \sqrt{\frac{I_{SS}}{\mu_n C_{ox}(W/L)_{1,2}}} + \frac{R_P I_{SS}}{2} \tag{15.26}$$

The terms on the right are the voltage headroom consumed by a current source, one threshold voltage, the equilibrium overdrive, and half of the swing at each node.

◀

In CMOS technologies lacking high-quality resistors, the implementation of Fig. 15.17(a) must be modified. While a PMOS transistor operating in the deep triode region can serve as the load [Fig. 15.18(a)], the gate voltage must be set so as to define the on-resistance accurately. Alternatively, a diode-connected load can be utilized [Fig. 15.18(b)], but at the cost of one threshold voltage in the headroom. Figure 15.18(c) shows a more efficient load where an NMOS source follower is inserted between the drain and gate of each PMOS transistor. With the output sensed at nodes X and Y, M_3 and M_4 consume only a voltage headroom equal to $|V_{DS3,4}|$. If $V_{GS5} \approx V_{TH3}$, then M_3 operates at the edge of the triode region and the small-signal resistance of the load is roughly equal to $1/g_{m3}$ (with the assumption that $\lambda = \gamma = 0$) (Problem 15.4).

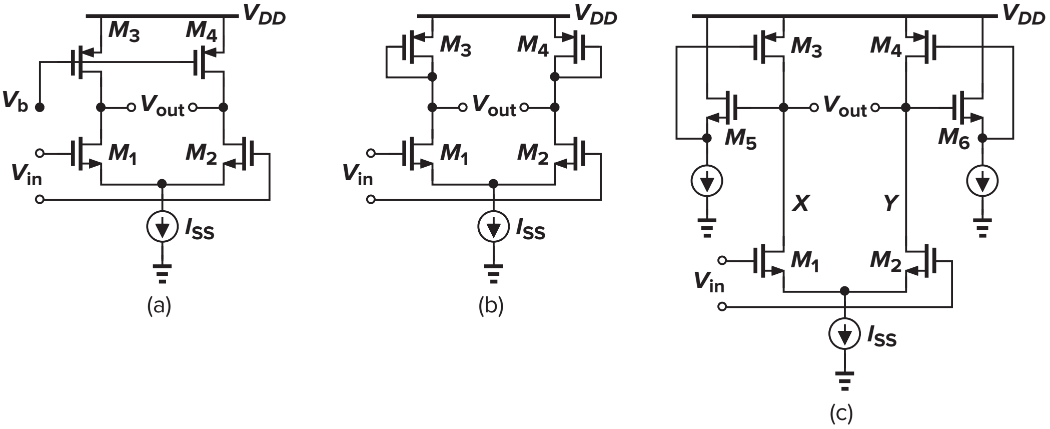


Figure 15.18 Differential stages using PMOS loads.

The load of Fig. 15.18(c) exhibits another interesting property as well. Since the gate-source capacitance of M_3 is driven by the source follower, the time constant associated with the load is smaller than that of a diode-connected transistor. Also, the finite output resistance of the follower may yield an inductive behavior for the load (Problem 15.5).

15.3 ■ LC Oscillators

Monolithic inductors have become common in CMOS technologies, making it possible to design oscillators based on passive resonant circuits. Before delving into such oscillators, it is instructive to review the basic properties of RLC circuits.

15.3.1 Basic Concepts

As shown in Fig. 15.19(a), an inductor L_1 placed in parallel with a capacitor C_1 resonates at a frequency $\omega_{res} = 1/\sqrt{L_1 C_1}$. At this frequency, the impedances of the inductor, $jL_1\omega_{res}$, and the capacitor, $1/(jC_1\omega_{res})$, are equal and opposite, thereby yielding an infinite impedance. We say that the circuit has an infinite quality factor, Q. In practice, inductors (and capacitors) suffer from resistive components. For example, the series resistance of the metal wire used in the inductor can be modeled as shown in Fig. 15.19(b). We define the Q of the inductor as $L_1\omega/R_S$. For this circuit, the reader can show that the equivalent impedance is given by

$$Z_{eq}(s) = \frac{R_S + L_1 s}{1 + L_1 C_1 s^2 + R_S C_1 s} \tag{15.27}$$

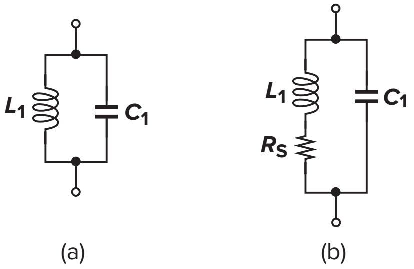


Figure 15.19 (a) Ideal and (b) realistic LC tanks.

and hence,

$$|Z_{eq}(s = j\omega)|^2 = \frac{R_S^2 + L_1^2\omega^2}{(1 - L_1C_1\omega^2)^2 + R_S^2C_1^2\omega^2} \tag{15.28}$$

That is, the impedance does not go to infinity at any $s = j\omega$. We say that the circuit has a finite Q. The magnitude of Z_{eq} in (15.28) reaches a peak in the vicinity of $\omega = 1/\sqrt{L_1C_1}$, but the actual resonance frequency has some dependency on R_S.

The circuit of Fig. 15.19(b) can be transformed to an equivalent topology that more easily lends itself to analysis and design. To this end, we first consider the series combination shown in Fig. 15.20(a). For a narrow frequency range, it is possible to convert the circuit to the parallel configuration of Fig. 15.20(b).

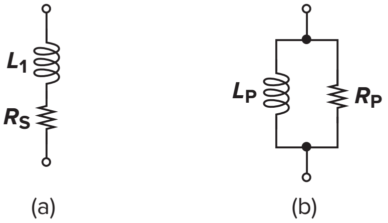


Figure 15.20 Conversion of a series combination to a parallel combination.

For the two impedances to be equivalent,

$$L_1s + R_S = \frac{R_PL_Ps}{R_P + L_Ps} \tag{15.29}$$

Considering only the steady-state response, we assume that $s = j\omega$ and rewrite (15.29) as

$$(L_1R_P + L_PR_S)j\omega + R_SR_P - L_1L_P\omega^2 = R_PL_Pj\omega \tag{15.30}$$

This relationship must hold for all values of ω (in a narrow range), dictating that

$$L_1R_P + L_PR_S = R_PL_P \tag{15.31}$$

$$R_SR_P - L_1L_P\omega^2 = 0 \tag{15.32}$$

Calculating R_P from the latter and substituting in the former, we have

$$L_P = L_1\left(1 + \frac{R_S^2}{L_1^2\omega^2}\right) \tag{15.33}$$

Recall that $L_1\omega/R_S = Q$, a value typically greater than 3 for monolithic inductors. Thus,

$$L_P \approx L_1 \tag{15.34}$$

and

$$R_P \approx \frac{L_1^2\omega^2}{R_S} \tag{15.35}$$

$$\approx Q^2 R_S \tag{15.36}$$

In other words, the parallel network has the same reactance but a resistance Q^2 times the series resistance. This concept holds valid for a first-order RC network as well if the Q of the series combination is defined as $1/(C\omega)/R_S$.

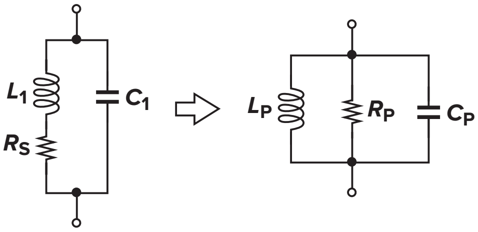


Figure 15.21 Conversion of a tank to three parallel components.

The above transformation allows the conversion illustrated in Fig. 15.21, where $C_P = C_1$. The equivalence of course breaks down as ω departs substantially from the resonance frequency. The insight gained from the parallel combination is that at $\omega_1 = 1/\sqrt{L_p C_p}$, the tank reduces to a simple resistor; i.e., the phase difference between the voltage and current of the tank drops to zero. Plotting the magnitude of the tank impedance versus frequency [Fig. 15.22(a)], we note that the behavior is inductive for $\omega < \omega_1$ and capacitive for $\omega > \omega_1$. We then surmise that the phase of the impedance is positive for $\omega < \omega_1$ and negative for $\omega > \omega_1$ [Fig. 15.22(b)]. These observations prove useful in studying LC oscillators. (Why do we expect the phase shift to approach +90° at very low frequencies and −90° at very high frequencies?)

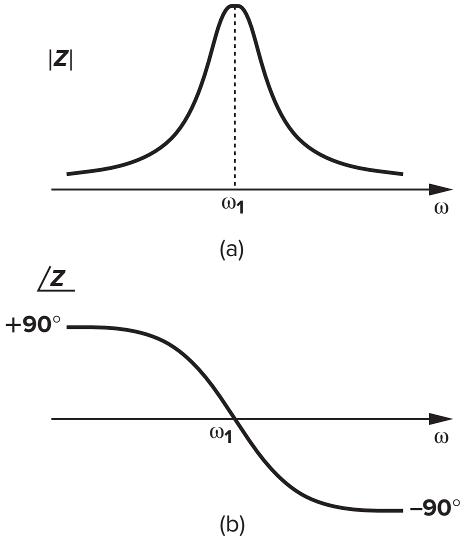


Figure 15.22 (a) Magnitude and (b) phase of the impedance of an LC tank as a function of frequency.

Let us now consider the "tuned" stage of Fig. 15.23(a), where an LC tank operates as the load. At resonance, $jL_p\omega = 1/(jC_p\omega)$ and the voltage gain equals $-g_{m1}R_P$. (Note that the gain of the circuit is very small at frequencies near zero.) Does this circuit oscillate if the output is connected to the input [Fig. 15.23(b)]? At resonance, the total phase shift around the loop is equal to 180° (rather than 360°). Also, from Fig. 15.22(b), the frequency-dependent phase shift of the tank never reaches 180°. Thus, the circuit does not oscillate.

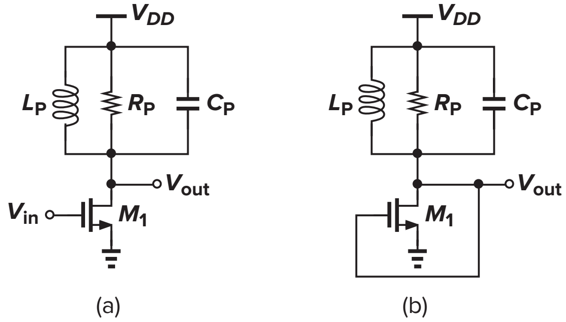


Figure 15.23 (a) Tuned gain stage; (b) stage of (a) in feedback.

Before modifying the circuit for oscillatory behavior, let us observe another interesting property of the gain stage of Fig. 15.23(a) that distinguishes it from a common-source topology using a resistive load. Suppose, as shown in Fig. 15.24, the stage is biased at a drain current I_1. If the series resistance of L_p is small, the dc level of V_{out} is close to V_{DD}. How does V_{out} vary if a small sinusoidal voltage at the resonance frequency is applied to the input? We expect V_{out} to be an inverted sinusoid with an average value near V_{DD} because the inductor cannot sustain a large dc drop. In other words, if the average value of V_{out} deviates significantly from V_{DD}, then the inductor series resistance must carry an average current greater than I_1. Thus, the peak output level in fact *exceeds* the supply voltage, an important and often useful attribute of the LC load. For example, with proper design, the output peak-to-peak swing can be larger than V_{DD}.

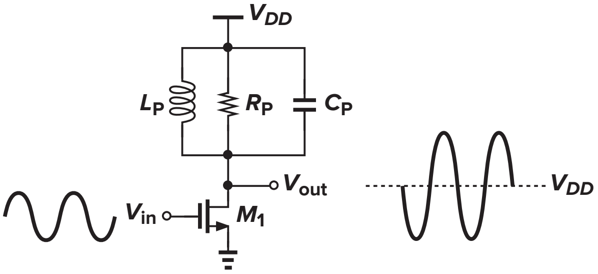


Figure 15.24 Output signal levels in a tuned stage.

We now study two types of LC oscillators.

15.3.2 Cross-Coupled Oscillator

Suppose we place two stages of Fig. 15.23(a) in a cascade, as depicted in Fig. 15.25. While similar to the topology of Fig. 15.5, this configuration does not latch up because its low-frequency gain is very small. Furthermore, at resonance, the total phase shift around the loop is zero because each stage contributes zero frequency-dependent phase shift. That is, if $g_{m1}R_Pg_{m2}R_P \geq 1$, then the loop oscillates. Note that V_X and V_Y are differential waveforms. (Why?)

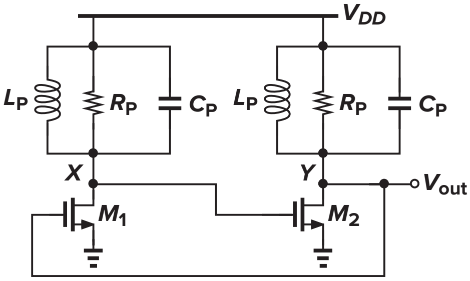


Figure 15.25 Two tuned stages in a feedback loop.

▶ **Example 15.5**

Sketch the open-loop voltage gain and phase of the circuit shown in Fig. 15.25. Neglect transistor capacitances.

Solution

The magnitude of the transfer function has a shape similar to that in Fig. 15.22(a), but with a sharper rise and fall because it results from the *product* of those of the two stages. The total phase at low frequencies is given by signal inversion by each common-source stage plus a 90° phase shift due to each tank. A similar behavior occurs at high frequencies. The gain and phase are sketched in Fig. 15.26. From these plots, the reader can prove that the circuit cannot oscillate at any other frequency.

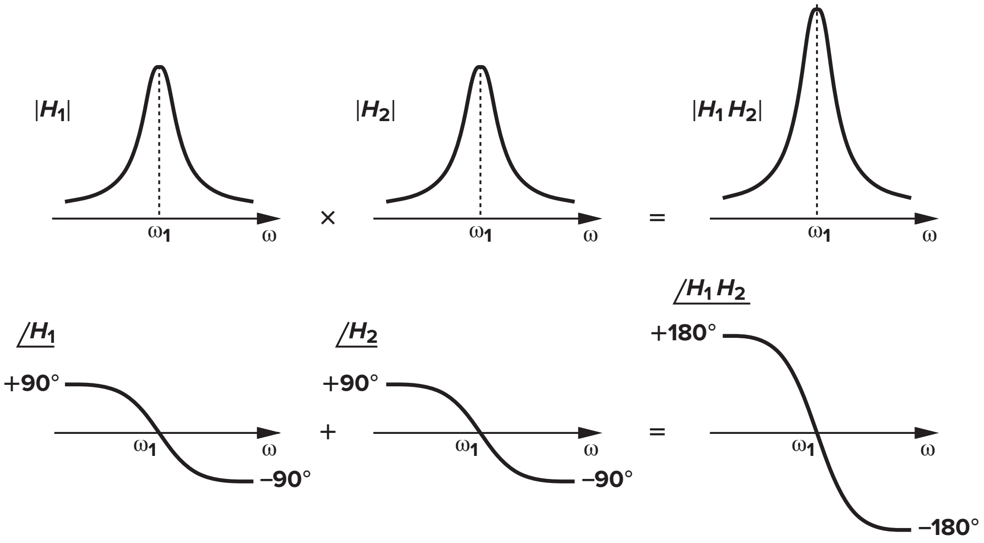


Figure 15.26 Loop gain characteristics of the circuit shown in Fig. 15.25.

The circuit of Fig. 15.25 serves as the core of many LC oscillators and is sometimes drawn as in Fig. 15.27(a) or (b). However, the drain currents of M_1 and M_2, and hence the output swings, heavily depend on the supply voltage. Since the waveforms at X and Y are differential, the drawing in Fig. 15.27(b) suggests that M_1 and M_2 can be converted to a differential pair as depicted in Fig. 15.27(c), where the total bias current is defined by I_{SS}.

▶ **Example 15.6**

For the circuit of Fig. 15.27(c), plot V_X and V_Y and I_{D1} and I_{D2} as the oscillation begins.

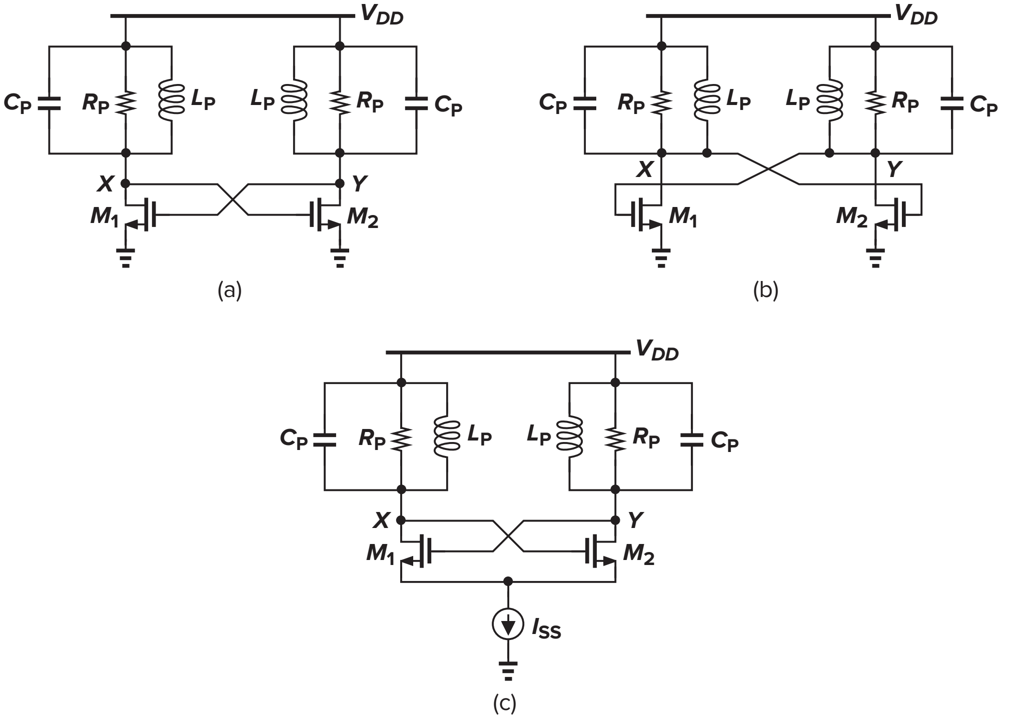


Figure 15.27 (a) Redrawing of the oscillator shown in Fig. 15.25; (b) another redrawing of the circuit; (c) addition of tail current source to lower supply sensitivity.

Solution

If the circuit begins with zero difference between V_X and V_Y, then $V_X = V_Y \approx V_{DD}$. The two transistors share the tail current equally. If $(g_{m1,2}R_P)^2 \geq 1$, where R_P is the equivalent parallel resistance of the tank at resonance, then noise components at the resonance frequency are amplified by M_1 and M_2, allowing the oscillation to grow. The drain currents of M_1 and M_2 vary according to the instantaneous value of $V_X - V_Y$ (as in a differential pair).

As shown in Fig. 15.28, the oscillation amplitude grows until the loop gain drops at the peaks. In fact, if $g_{m1,2}R_P$ is large enough, the difference between $V_X - V_Y$ reaches a level that steers the entire tail current to one transistor, turning the other off. Thus, in the steady state, I_{D1} and I_{D2} vary between zero and I_{SS}.

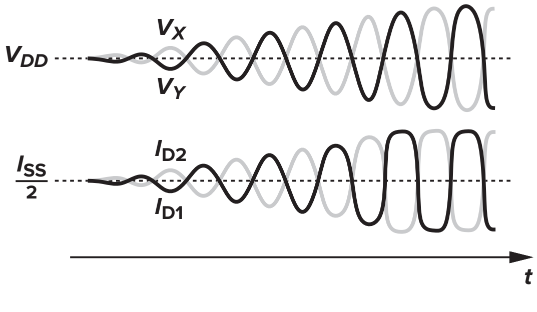


Figure 15.28

The oscillator of Fig. 15.27(c) is constructed in fully differential form. The supply sensitivity of the circuit, however, is nonzero even with perfect symmetry. This is because the drain junction capacitances of M_1 and M_2 vary with the supply voltage. We return to this issue in Example 15.9.

15.3.3 Colpitts Oscillator

An LC oscillator may be realized with only one transistor in the signal path. Consider the gain stage of Fig. 15.23(a) again and recall that the drain voltage cannot be applied to the gate because the overall phase shift at resonance equals 180° rather than 360°. Also, recall that in a common-gate stage, the phase shift from the source to the drain is zero. We then surmise that if, as shown in Fig. 15.29(a), the drain voltage is returned to the source rather than the gate, the circuit may oscillate. The coupling must incorporate a capacitor to avoid disturbing the bias point of M_1.

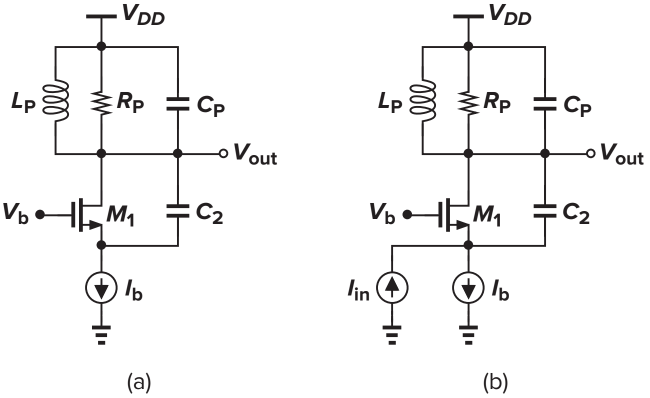


Figure 15.29 (a) Tuned stage with feedback applied from drain to source; (b) addition of input current to calculate closed-loop gain.

Unfortunately, owing to insufficient loop gain, the circuit of Fig. 15.29(a) does not oscillate. To prove this point, we invoke the view of Fig. 15.1, where an oscillator is considered a feedback system with infinite closed-loop gain. Applying an input current as depicted in Fig. 15.29(b) and neglecting transistor parasitics, we obtain the closed-loop gain as

$$\frac{V_{out}}{I_{in}} = L_P s \Big\| \frac{1}{C_P s} \Big\| R_P \tag{15.37}$$

because M_1 and C_2 directly conduct the input current to the tank. Since the closed-loop gain cannot be equal to infinity at any frequency, the circuit fails to oscillate.

▶ **Example 15.7**

The reader may wonder why the input to the feedback system is realized as a current source applied to the source of the transistor rather than a voltage source applied to its gate. Perform the analysis with the latter stimulus.

Solution

From Fig. 15.30, we note that with a finite variation of V_{in}, the change in I_b is still zero if the bias current source is ideal. Thus, if the source-bulk junction capacitance of M_1 is neglected, the change in the tank current is zero, yielding $V_{out}/V_{in} = 0$. Interestingly, V_X does vary with V_{in}, but M_1 generates a small-signal current that cancels that through C_2. The reader can prove that $V_X/V_{in} = g_m/(g_m + C_2 s)$. ◀

The above example reveals two important points. First, to excite a circuit into oscillation, the stimulus can be applied at different points. (That is, the noise of any device in the loop can initiate the

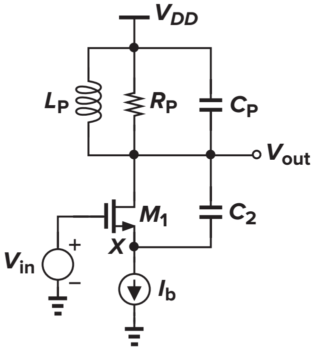
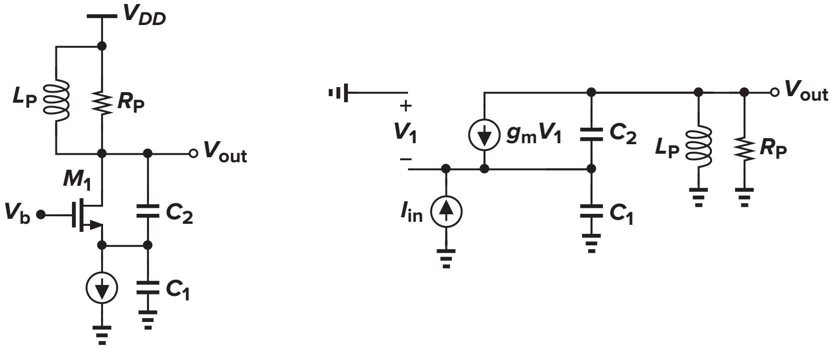


Figure 15.30

oscillation.[6]) Second, in Fig. 15.30, V_{out}/V_{in} is zero because the impedance connected between the source of M_1 and ground is infinity. We then add a capacitor from this node to ground as shown in Fig. 15.31(a), seeking conditions of oscillation. Note that the capacitor in parallel with L_P is removed. The reason will become clear later.



(a) (b)

Figure 15.31 (a) Colpitts oscillator; (b) equivalent circuit of (a) with input stimulus.

Approximating M_1 by a single voltage-dependent current source, we construct the equivalent circuit of Fig. 15.31(b). Since the current through the parallel combination of L_P and R_P is given by $V_{out}/(L_Ps) + V_{out}/R_P$, the total current through C_1 is equal to $I_{in} - V_{out}/(L_Ps) - V_{out}/R_P$, yielding

$$V_1 = -\left(I_{in} - \frac{V_{out}}{L_Ps} - \frac{V_{out}}{R_P}\right)\frac{1}{C_1s} \tag{15.38}$$

Writing the current through C_2 as $(V_{out} + V_1)C_2s$, we sum all of the currents at the output node:

$$-g_m\left(I_{in} - \frac{V_{out}}{L_Ps} - \frac{V_{out}}{R_P}\right)\frac{1}{C_1s} + \left[V_{out} - \left(I_{in} - \frac{V_{out}}{L_Ps} - \frac{V_{out}}{R_P}\right)\frac{1}{C_1s}\right]C_2s + \frac{V_{out}}{L_Ps} + \frac{V_{out}}{R_P} = 0 \tag{15.39}$$

It follows that

$$\frac{V_{out}}{I_{in}} = \frac{R_PL_Ps(g_m + C_2s)}{R_PC_1C_2L_Ps^3 + (C_1 + C_2)L_Ps^2 + [g_mL_P + R_P(C_1 + C_2)]s + g_mR_P} \tag{15.40}$$

[6] This is because the natural frequencies of a linear (observable) system do not depend on the location of the stimulus. Of course, the type of stimulus (voltage or current) must be chosen such that when it is set to zero, the circuit returns to its original topology. For example, driving the gate of M_1 in Fig. 15.30 by a current changes the natural frequencies of the circuit.

Note that, as expected, (15.40) reduces to $(L_P s || R_P)$ if $C_1 = 0$. The circuit oscillates if the closed-loop transfer function goes to infinity at an imaginary value of s, $s_R = j\omega_R$. Consequently, both the real and imaginary parts of the denominator must drop to zero at this frequency:

$$-R_P C_1 C_2 L_P \omega_R^3 + [g_m L_P + R_P(C_1 + C_2)]\omega_R = 0 \tag{15.41}$$

$$-(C_1 + C_2) L_P \omega_R^2 + g_m R_P = 0 \tag{15.42}$$

Since with typical values, $g_m L_P \ll R_P(C_1 + C_2)$, Eq. (15.41) yields,

$$\omega_R^2 = \frac{1}{L_P \dfrac{C_1 C_2}{C_1 + C_2}} \tag{15.43}$$

and Eq. (15.42) results in

$$g_m R_P = \frac{(C_1 + C_2)^2}{C_1 C_2} \tag{15.44}$$

$$= \frac{C_1}{C_2}\left(1 + \frac{C_2}{C_1}\right)^2 \tag{15.45}$$

Recognizing that $g_m R_P$ is the voltage gain from the source of M_1 to the output (if $g_{mb} = 0$), we determine the ratio C_1/C_2 for the minimum required gain. The reader can prove that the minimum occurs for $C_1/C_2 = 1$, requiring

$$g_m R_P \geq 4 \tag{15.46}$$

Equation (15.46) demonstrates an important disadvantage of the Colpitts oscillator with respect to the cross-coupled topology of Fig. 15.27(c). The former demands a voltage gain of at least 4 at resonance, and the latter, only unity. This issue is critical if the inductor suffers from a low Q and hence a small R_P, a common situation in CMOS technologies. As a consequence, the cross-coupled scheme is used more widely.

The foregoing analysis neglected the capacitance that appears in parallel with the inductor. As suggested in Problem 15.10, if this capacitance, C_P, is included in the equivalent circuit, Eq. (15.43) is modified as

$$\omega_R^2 = \frac{1}{L_P\left(C_P + \dfrac{C_1 C_2}{C_1 + C_2}\right)} \tag{15.47}$$

whereas (15.46) remains unchanged. Thus, C_P is simply included in parallel with the series combination of C_1 and C_2.

15.3.4 One-Port Oscillators

Our development of oscillators thus far has been based on feedback systems. An alternative view that provides more insight into the oscillation phenomenon employs the concept of "negative resistance." To arrive at this view, let us first consider a simple tank that is stimulated by a current impulse [Fig. 15.32(a)]. The tank responds with a decaying oscillatory behavior because, in every cycle, some of the energy that reciprocates between the capacitor and the inductor is lost in the form of heat in the resistor. Now suppose a resistor equal to $-R_P$ is placed in parallel with R_P and the experiment is repeated [Fig. 15.32(b)].

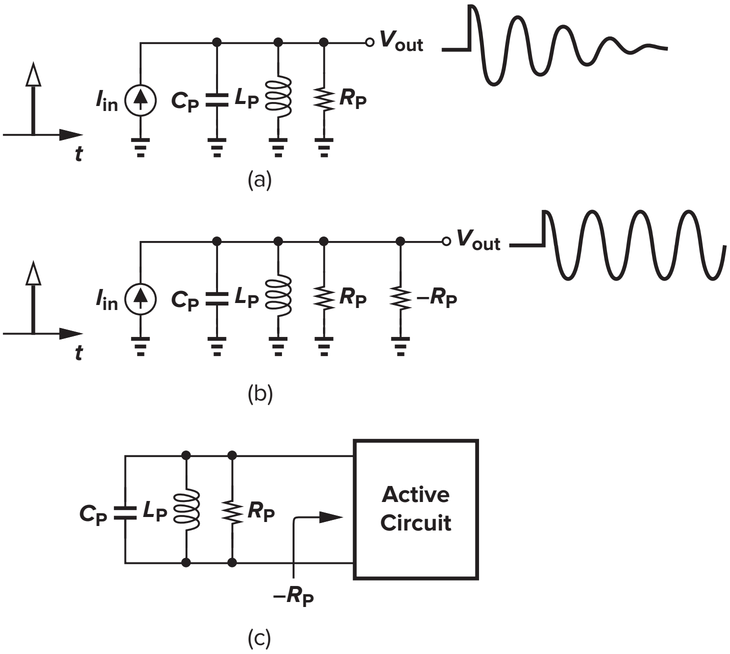


Figure 15.32 (a) Decaying impulse response of a tank; (b) addition of negative resistance to cancel loss in R_P; (c) use of an active circuit to provide negative resistance.

Since $R_P||(-R_P) = \infty$, the tank oscillates indefinitely. Thus, if a one-port circuit exhibiting a negative resistance is placed in parallel with a tank [Fig. 15.32(c)], the combination may oscillate. Such a topology is called a one-port oscillator.

How can a circuit provide a negative resistance? Recall that feedback multiplies or divides the input and output impedances of circuits by a factor equal to one plus the loop gain. Thus, if the loop gain is sufficiently *negative* (i.e., the feedback is sufficiently positive), a negative resistance is achieved. As a simple example, let us apply positive feedback around a source follower. The follower introduces no signal inversion, and neither must the feedback network. As depicted in Fig. 15.33(a), we implement the feedback by a common-gate stage and add the current source I_b to provide the bias current of M_2.[7] From the equivalent circuit in Fig. 15.33(b) (where channel-length modulation and body effect are neglected),

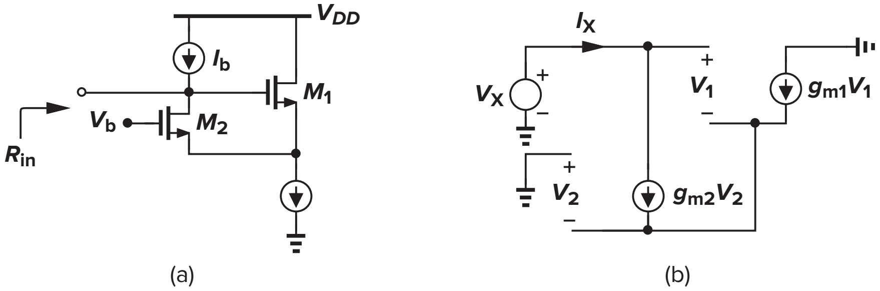


Figure 15.33 (a) Source follower with positive feedback to create negative input impedance; (b) equivalent circuit of (a) to calculate the input impedance.

[7] This circuit can also be viewed as a CG stage with the source follower providing feedback.

we have

$$I_X = g_{m2}V_2 = -g_{m1}V_1 \tag{15.48}$$

and

$$V_X = V_1 - V_2 \tag{15.49}$$

$$= -\frac{I_X}{g_{m1}} - \frac{I_X}{g_{m2}} \tag{15.50}$$

Thus,

$$\frac{V_X}{I_X} = -(\frac{1}{g_{m1}} + \frac{1}{g_{m2}}) \tag{15.51}$$

and, if $g_{m1} = g_{m2} = g_m$, then

$$\frac{V_X}{I_X} = \frac{-2}{g_m} \tag{15.52}$$

Negative resistance becomes more intuitive if we bear in mind that it is an *incremental* quantity; that is, negative resistance indicates that if the applied voltage *increases*, the current drawn by the circuit *decreases*. In Fig. 15.33(a), for example, if the input voltage increases, so does the source voltage of M_1, decreasing the drain current of M_2 and allowing part of I_b to flow to the input source.

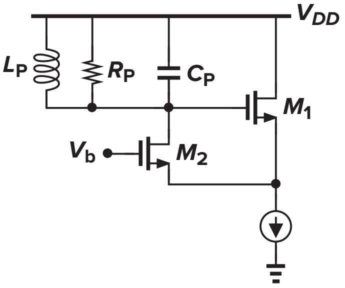


Figure 15.34 Oscillator using the negative input resistance of a source follower with positive feedback.

With a negative resistance available, we can now construct an oscillator as illustrated in Fig. 15.34. Here, R_P denotes the equivalent parallel resistance of the tank and, for oscillation build-up, $R_P - 2/g_m \geq 0$. Note that the inductor provides the bias current of M_2, obviating the need for a current source. If the small-signal resistance presented by M_1 and M_2 to the tank is less negative than $-R_P$, then the circuit experiences large swings such that each transistor is nearly off for part of the period, thereby yielding an "average" resistance of $-R_P$.

The circuit of Fig. 15.34 is similar to the stage of Fig. 15.29(a), but with the feedback capacitor replaced by a source follower. More interestingly, the circuit can be redrawn as in Fig. 15.35(a), bearing a resemblance to Fig. 15.27(c). In fact, if the drain current of M_1 flows through a tank and the resulting voltage is applied to the gate of M_2, the topology of Fig. 15.35(b) is obtained. Ignoring bias paths and merging the two tanks into one (Fig. 15.36), we note that the cross-coupled pair must provide a negative resistance of $-R_P$ between nodes X and Y to enable oscillation. The reader can prove that this resistance is equal to $-2/g_m$ and hence it is necessary that $R_P \geq 1/g_m$. Thus, the circuit can be viewed as either a feedback system or a negative resistance in parallel with a lossy tank. This topology is also called a "negative-G_m oscillator."

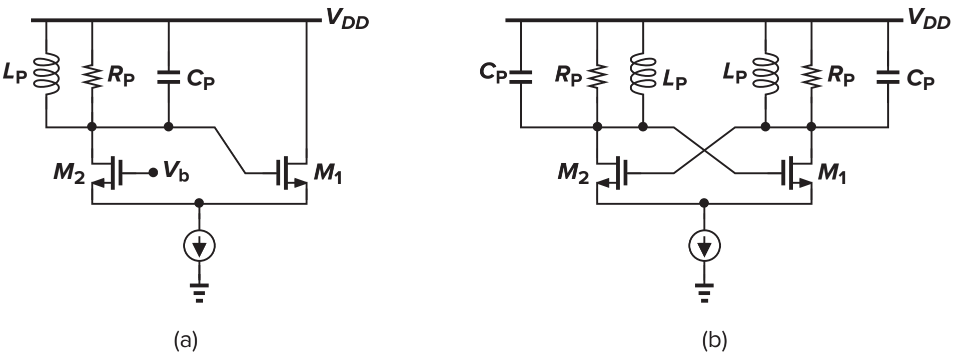


Figure 15.35 (a) Redrawing of the topology shown in Fig. 15.34; (b) differential version of (a).

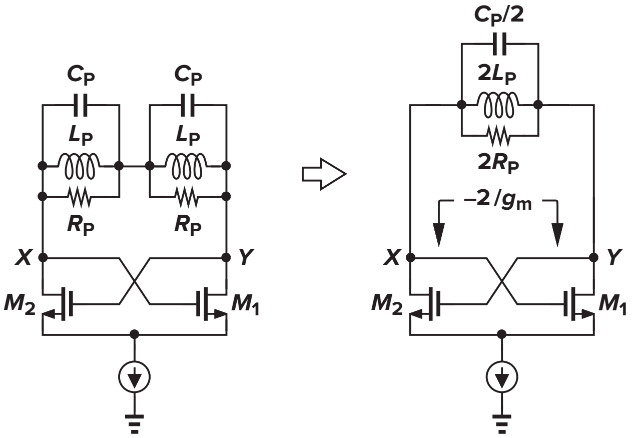


Figure 15.36 Equivalent circuit of Fig. 15.35(b).

As another method of creating negative resistance, consider the topology depicted in Fig. 15.37(a), where none of the nodes is grounded and channel-length modulation, body effect, and transistor capacitances are neglected. Since the drain current of M_1 is equal to $(-I_X/C_1s)g_m$, we have

$$V_X = \left(I_X - \frac{-I_X}{C_1s}g_m\right)\frac{1}{C_2s} + \frac{I_X}{C_1s} \tag{15.53}$$

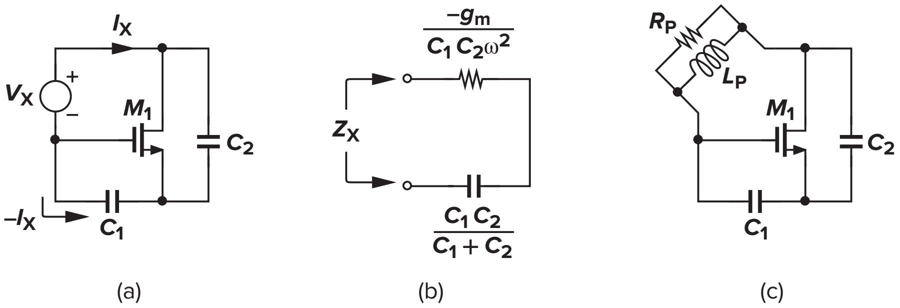


Figure 15.37 (a) Circuit topology providing negative resistance; (b) equivalent circuit of (a); (c) oscillator using (a).

and hence

$$\frac{V_X}{I_X} = \frac{g_m}{C_1 C_2 s^2} + \frac{1}{C_2 s} + \frac{1}{C_1 s} \tag{15.54}$$

For $s = j\omega$, this impedance consists of a negative resistance equal to $-g_m/(C_1 C_2 \omega^2)$ in series with the series combination of C_1 and C_2 [Fig. 15.37(b)]. Thus, as shown in Fig. 15.37(c), if an inductor is placed between the gate and drain of M_1, the circuit may oscillate. Of the three nodes in the circuit, one can be an ac ground, resulting in the three different topologies illustrated in Fig. 15.38. The circuit of Fig. 15.38(a) is in fact based on a source follower, whose input impedance was found in Chapter 6 to contain a negative real part. The configuration of Fig. 15.38(b) is a Colpitts oscillator.

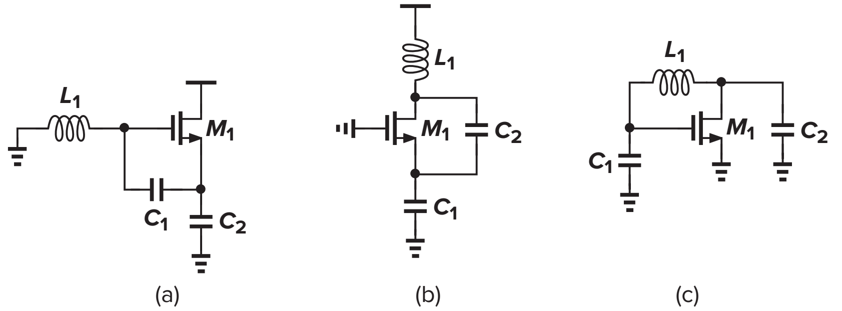


Figure 15.38 Oscillator topologies derived from the circuit of Fig. 15.37(c).

Example 15.8

Redraw the circuits of Fig. 15.38 with proper biasing.

Solution

The circuits are redrawn in Fig. 15.39.

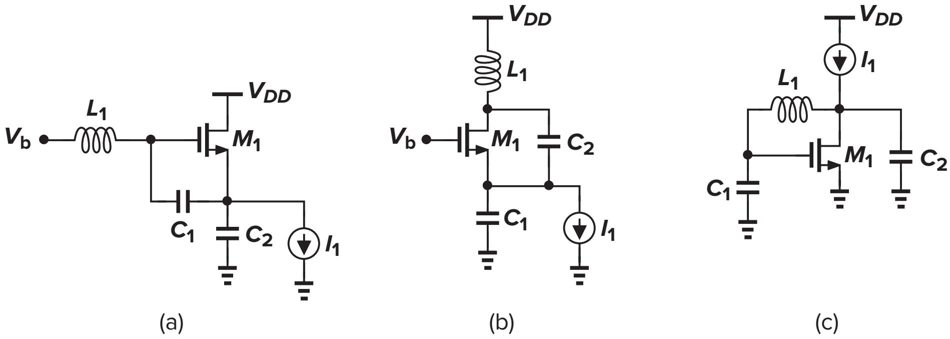


Figure 15.39

15.4 ■ Voltage-Controlled Oscillators

Most applications require that oscillators be “tunable,” i.e., that their output frequency be a function of a control input, usually a voltage. An ideal voltage-controlled oscillator is a circuit whose output frequency

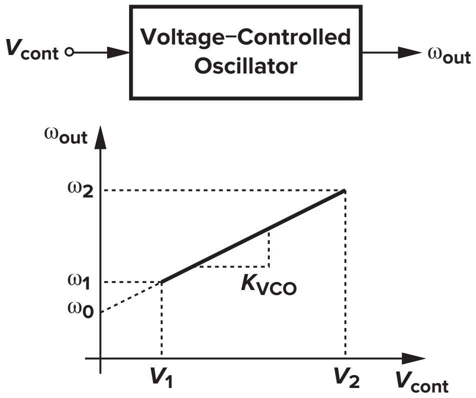


Figure 15.40 Definition of a VCO.

is a linear function of its control voltage (Fig. 15.40):

$$\omega_{out} = \omega_0 + K_{VCO}V_{cont} \tag{15.55}$$

Here, ω_0 represents the intercept corresponding to $V_{cont} = 0$ and K_{VCO} denotes the "gain" or "sensitivity" of the circuit (expressed in rad/s/V).[8] The achievable range, $\omega_2 - \omega_1$, is called the "tuning range."

Example 15.9

In the negative-G_m oscillator of Fig. 15.27(c), assume that $C_P = 0$, consider only the drain junction capacitance, C_{DB}, of M_1 and M_2, and explain why V_{DD} can be viewed as the control voltage. Calculate the gain of the VCO.

Solution

Since C_{DB} varies with the drain-bulk voltage, if V_{DD} changes, so does the resonance frequency of the tank. Noting that the average voltage across C_{DB} is approximately equal to V_{DD}, we write

$$C_{DB} = \frac{C_{DB0}}{\left(1 + \dfrac{V_{DD}}{\phi_B}\right)^m} \tag{15.56}$$

and

$$K_{VCO} = \frac{\partial\omega_{out}}{\partial V_{DD}} \tag{15.57}$$

$$= \frac{\partial\omega_{out}}{\partial C_{DB}} \cdot \frac{\partial C_{DB}}{\partial V_{DD}} \tag{15.58}$$

With $\omega_{out} = 1/\sqrt{L_P C_{DB}}$, we have

$$K_{VCO} = \frac{-1}{2\sqrt{L_P C_{DB}}C_{DB}} \cdot \frac{-mC_{DB}}{\phi_B\left(1 + \dfrac{V_{DD}}{\phi_B}\right)} \tag{15.59}$$

$$= \frac{m}{2\phi_B\left(1 + \dfrac{V_{DD}}{\phi_B}\right)} \cdot \omega_{out} \tag{15.60}$$

Note that the relationship between ω_{out} and V_{cont} is nonlinear because K_{VCO} varies with V_{DD} and ω_{out}.

[8] A more familiar unit is Hz/V, but one must be careful with the dimension of K_{VCO} in the context of phase-locked loops.

Before modifying the oscillators studied in the previous sections for tunability, we summarize the important performance parameters of VCOs.

Center Frequency The center frequency (i.e., the midrange value in Fig. 15.40) is determined by the environment in which the VCO is used. For example, in the clock generation network of a microprocessor, the VCO may be required to run at the clock rate or even twice that. Today's CMOS VCOs achieve center frequencies as high as hundreds of gigahertz.

Tuning Range The required tuning range is dictated by two parameters: (1) the variation of the VCO center frequency with process and temperature, and (2) the frequency range necessary for the application. The center frequency of some CMOS oscillators may vary by a factor of two at the extremes of process and temperature, thus mandating a sufficiently wide ($\geq 2\times$) tuning range to guarantee that the VCO output frequency can be driven to the desired value. Also, some applications incorporate clock frequencies that must vary by one to two orders of magnitude depending on the mode of operation, demanding a proportionally wide tuning range.

An important concern in the design of VCOs is the disturbance of the output phase and frequency as a result of noise on the control line. For a given noise amplitude, the noise in the output frequency is proportional to K_{VCO} because $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$. Thus, to minimize the effect of noise in V_{cont}, the VCO gain must be *minimized*, a constraint in direct conflict with the required tuning range. In fact, if, as shown in Fig. 15.40, the allowable range of V_{cont} is from V_1 to V_2 (e.g., from 0 to V_{DD}) and the tuning range must span at least ω_1 to ω_2, then K_{VCO} must satisfy the following requirement:

$$K_{VCO} \geq \frac{\omega_2 - \omega_1}{V_2 - V_1} \tag{15.61}$$

Note that, for a given tuning range, K_{VCO} increases as the supply voltage decreases, making the oscillator more sensitive to noise on the control line.

Tuning Linearity As exemplified by Eq. (15.60), the tuning characteristics of VCOs exhibit nonlinearity, i.e., their gain, K_{VCO}, is not constant. As explained in Chapter 16, such nonlinearity degrades the settling behavior of phase-locked loops. For this reason, it is desirable to minimize the variation of K_{VCO} across the tuning range.

Actual oscillator characteristics typically exhibit a high-gain region in the middle of the range and a low gain at the two extremes (Fig. 15.41). Compared to a linear characteristic (the gray line), the actual behavior displays a maximum gain *greater* than that predicted by (15.61), implying that, for a given tuning range, nonlinearity inevitably leads to higher sensitivity for some region of the characteristic.

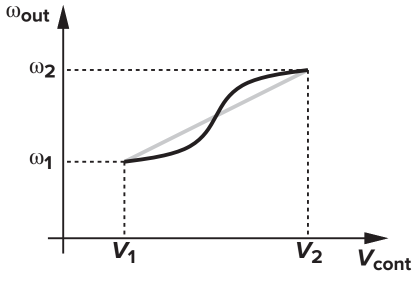


Figure 15.41 Nonlinear VCO characteristic.

Output Amplitude It is desirable to achieve a large output oscillation amplitude, thus making the waveform less sensitive to noise. The amplitude trades with power dissipation, supply voltage, and (as explained in Sec. 15.4.2) even the tuning range. Also, the amplitude may vary across the tuning range, an undesirable effect.

Power Dissipation As with other analog circuits, oscillators suffer from trade-offs among speed, power dissipation, and noise. Typical oscillators drain 1 to 10 mW of power.

Supply and Common-Mode Rejection Oscillators are quite sensitive to noise, especially if they are realized in single-ended form. As seen in Example 15.9, even differential oscillators exhibit supply sensitivity. The design of oscillators for high noise immunity is a difficult challenge.

Output Signal Purity Even with a constant control voltage, the output waveform of a VCO is not perfectly periodic. The electronic noise of the devices in the oscillator and supply noise lead to noise in the output phase and frequency. These effects are quantified by "jitter" and "phase noise" and determined by the requirements of each application.

15.4.1 Tuning in Ring Oscillators

Recall from Sec. 15.2 that the oscillation frequency, f_{osc}, of an N-stage ring equals $(2NT_D)^{-1}$, where T_D denotes the large-signal delay of each stage. Thus, to vary the frequency, T_D can be adjusted.

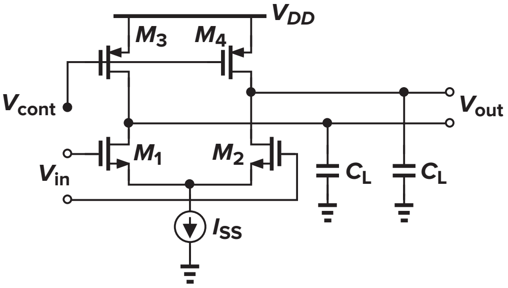


Figure 15.42 Differential pair with variable output time constant.

As a simple example, consider the differential pair of Fig. 15.42 as one stage of a ring oscillator. Here, M_3 and M_4 operate in the triode region, each acting as a variable resistor controlled by V_{cont}. As V_{cont} becomes more positive, the on-resistance of M_3 and M_4 increases, thus raising the time constant at the output, τ_1, and lowering f_{osc}. If M_3 and M_4 remain in the deep triode region,

$$\tau_1 = R_{on3,4}C_L \tag{15.62}$$

$$= \frac{C_L}{\mu_p C_{ox}\left(\frac{W}{L}\right)_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)} \tag{15.63}$$

In the above equation, C_L denotes the total capacitance seen at each output to ground (including the input capacitance of the following stage). The delay of the circuit is roughly proportional to τ_1, yielding

$$f_{osc} \propto \frac{1}{T_D} \tag{15.64}$$

$$\propto \frac{\mu_p C_{ox}\left(\frac{W}{L}\right)_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)}{C_L} \tag{15.65}$$

Interestingly, f_{osc} is linearly proportional to V_{cont}.

▶ **Example 15.10**

For the given device dimensions and bias currents in Fig. 15.42, determine the maximum allowable value of V_{cont}. What happens if M_3 and M_4 enter saturation?

Solution

Let us assume (somewhat arbitrarily) that M_3 and M_4 remain in the deep triode region if $|V_{DS3,4}| \leq 0.2 \times 2|V_{GS3,4} - V_{THP}|$. If each stage in the ring experiences complete switching, then the maximum drain current of M_3 and M_4 is equal to I_{SS}. To satisfy the above condition, we must have $I_{SS}R_{on3,4} \leq 0.4(V_{DD} - V_{cont} - |V_{THP}|)$, and hence

$$\frac{I_{SS}}{\mu_p C_{ox}\left(\frac{W}{L}\right)_{3,4}(V_{DD} - V_{cont} - |V_{THP}|)} \leq 0.4(V_{DD} - V_{cont} - |V_{THP}|) \tag{15.66}$$

It follows that

$$V_{cont} \leq V_{DD} - |V_{THP}| - \sqrt{\frac{I_{SS}}{0.4\mu_p C_{ox}\left(\frac{W}{L}\right)_{3,4}}} \tag{15.67}$$

If V_{cont} exceeds this level by a large margin, M_3 and M_4 eventually enter saturation. Each stage then requires common-mode feedback to produce the output swings around a well-defined CM level.

The differential pair of Fig. 15.42 suffers from a critical drawback: the output swing of the circuit varies considerably across the tuning range. With complete switching, each stage provides a differential output swing of $2I_{SS}R_{on3,4}$. Thus, a tuning range of, say, two to one translates to a twofold variation in the swing.

In order to minimize the swing variation, the tail current can be adjusted by V_{cont} as well such that, as V_{cont} becomes more positive, I_{SS} decreases. The circuit nonetheless requires a means of maintaining $I_{SS}R_{on3,4}$ relatively constant. To this end, let us consider the circuit in Fig. 15.43(a), where M_5 operates in the deep triode region and amplifier A_1 applies negative feedback to the gate of M_5. If the loop gain is sufficiently large, the differential input voltage of A_1 must be small, giving $V_P \approx V_{REF}$ and $|V_{DS5}| \approx V_{DD} - V_{REF}$. Thus, the feedback ensures a relatively constant drain-source voltage even if I_1 varies. In fact, as I_1, say, decreases, A_1 raises the gate voltage of M_5 such that $R_{on5}I_1 \approx V_{DD} - V_{REF}$.

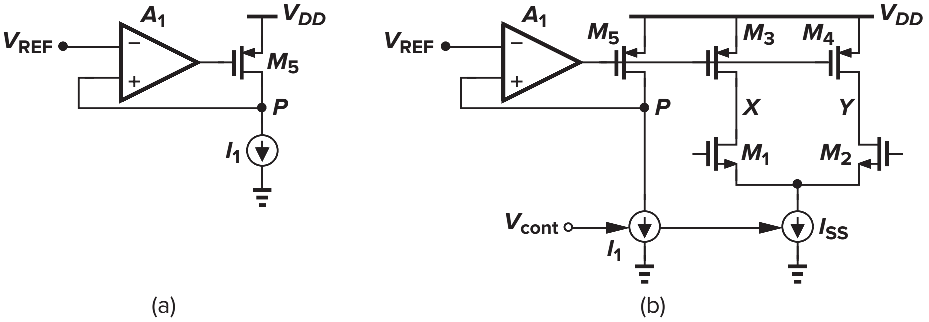


Figure 15.43 (a) Simple feedback circuit defining V_P; (b) replica biasing to define voltage swings in a ring oscillator.

The topology of Fig. 15.43(a) can serve as a "replica circuit" for the stages of a ring oscillator, thereby defining the oscillation amplitude. Illustrated in Fig. 15.43(b), the idea is to "servo" the on-resistance of M_3 and M_4 to that of M_5 and vary the frequency by adjusting I_1 and I_{SS} simultaneously [2]. If M_3 and M_4 are identical to M_5 and I_{SS} to I_1, then V_X and V_Y vary from V_{DD} to $V_{DD} - V_{REF}$ as M_1 and M_2 steer the tail current to one side or the other. Thus, if process and temperature variations, say, decrease I_1 and I_{SS}, then A_1 increases the on-resistance of M_3–M_5, forcing V_P and hence V_X and V_Y (when M_1 or M_2 is fully on) equal to V_{REF}.

The bandwidth of the op amp A_1 in Fig. 15.43(b) is of some concern. If a change in V_{cont} takes a long time to change ω_{out}, then the settling speed of a PLL using this VCO degrades significantly (Chapter 16).

▶ **Example 15.11**

How does the oscillation frequency depend on I_{SS} for a VCO incorporating the stage of Fig. 15.43(b)?

Solution

Noting that $R_{on3,4}I_{SS} \approx V_{DD} - V_{REF}$, we have $R_{on3,4} \approx (V_{DD} - V_{REF})/I_{SS}$, and hence

$$f_{osc} \propto \frac{1}{R_{on3,4}C_L} \tag{15.68}$$

$$\propto \frac{I_{SS}}{(V_{DD} - V_{REF})C_L} \tag{15.69}$$

Thus, the characteristic is relatively linear. ◀

Delay Variation by Positive Feedback To arrive at another tuning technique, recall that a cross-coupled transistor pair such as that of Fig. 15.36 exhibits a negative resistance of $-2/g_m$, a value that can be controlled by the bias current. A negative resistance $-R_N$ placed in parallel with a positive resistance $+R_P$ gives an equivalent value $+R_N R_P/(R_N - R_P)$, which is more positive if $|-R_N| > |+R_P|$. This idea can be applied to each stage of a ring oscillator as illustrated in Fig. 15.44(a). Here, the load of the differential pair consists of resistors R_1 and R_2 ($R_1 = R_2 = R_P$) and the cross-coupled pair M_3–M_4. As I_1 increases, the small-signal differential resistance $-2/g_{m3,4}$ becomes less negative and, from the half circuit of Fig. 15.44(b), the equivalent resistance $R_P||(-1/g_{m3,4}) = R_P/(1 - g_{m3,4}R_P)$ increases, thereby lowering the frequency of oscillation.

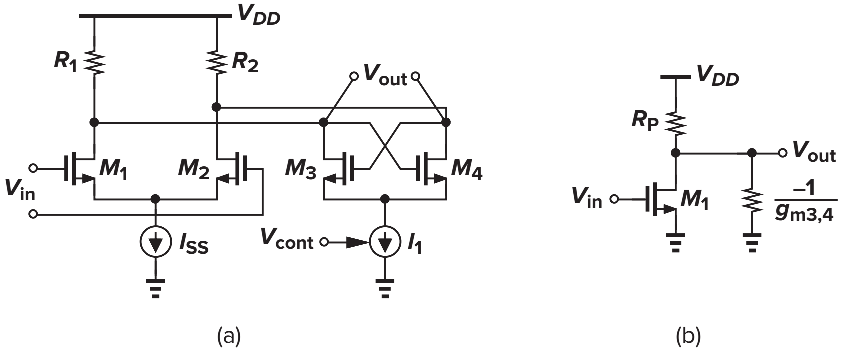


Figure 15.44 (a) Differential stage with variable negative-resistance load; (b) half-circuit equivalent of (a).

An important issue in the circuit of Fig. 15.44(a) is that as I_1 varies, so do the currents steered by M_3 and M_4 to R_1 and R_2. Thus, the output voltage swing is not constant across the tuning range. To minimize this effect, I_{SS} can be varied in the *opposite* direction such that the total current steered between R_1 and R_2 remains constant. In other words, it is desirable to vary I_1 and I_{SS} *differentially* while their sum is fixed, a characteristic provided by a differential pair. Illustrated in Fig. 15.45, the idea is to employ a differential pair M_5–M_6 to steer I_T to M_1–M_2 or M_3–M_4 so that $I_{SS} + I_1 = I_T$. Since I_T must flow through R_1 and R_2, if M_1–M_4 experience complete switching in each cycle of oscillation, then I_T is steered to R_1 (through M_1 and M_3) in half a period and to R_2 (through M_2 and M_4) in the other half, giving a differential swing of $2R_P I_T$.

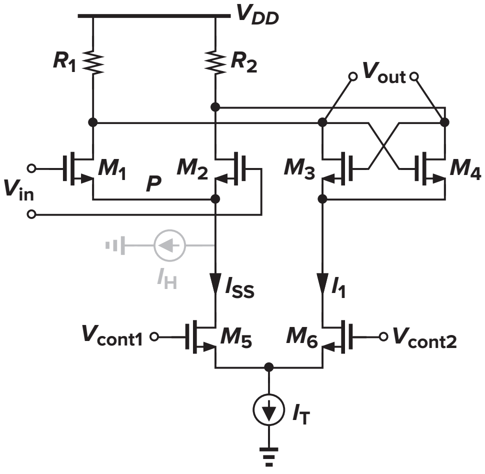


Figure 15.45 Use of a differential pair to steer current between M_1–M_2 and M_3–M_4.

In the circuit of Fig. 15.45, V_{cont1} and V_{cont2} can be viewed as differential control lines if they vary by equal and opposite amounts. Such a topology provides higher noise immunity for the control input than if V_{cont} is single-ended. Now, note that as V_{cont1} decreases and V_{cont2} increases, the cross-coupled pair exhibits a greater transconductance, thereby raising the time constant at the output nodes. But what happens if all of I_T is steered by M_6 to M_3 and M_4? Since M_1 and M_2 carry no current, the gain of the stage falls to zero, prohibiting oscillation. To avoid this effect, a small constant current source, I_H, can be connected from node P to ground, thereby ensuring that M_1 and M_2 always remain on. With typical values, this ring oscillator provides a two-to-one tuning range and reasonable linearity.

▶ Example 15.12

Calculate the minimum value of I_H in Fig. 15.45 to guarantee a low-frequency gain of 2 when all of I_T is steered to the cross-coupled pair.

Solution

The small-signal voltage gain of the circuit equals $g_{m1,2}R_P/(1-g_{m3,4}R_P)$. Assuming square-law devices, we have

$$\sqrt{\mu_n C_{ox}\left(\frac{W}{L}\right)_{1,2} I_H}\,\frac{R_P}{1-\sqrt{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4} I_T R_P}} \geq 2 \tag{15.70}$$

That is

$$I_H \geq \frac{4\left[1-\sqrt{\mu_n C_{ox}\left(\frac{W}{L}\right)_{3,4} I_T R_P}\right]^2}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{1,2} R_P^2} \tag{15.71}$$

◀

An important drawback of using the differential pair M_5–M_6 in the circuit of Fig. 15.45 is the additional voltage headroom that it consumes. As depicted in Fig. 15.46, for M_5 to remain in saturation, V_P must be sufficiently higher than V_N. When $V_{cont1} = V_{cont2}$, the minimum allowable drain-source voltage of M_5 is equal to its equilibrium overdrive voltage, implying that, compared to that calculated in Example 15.4,

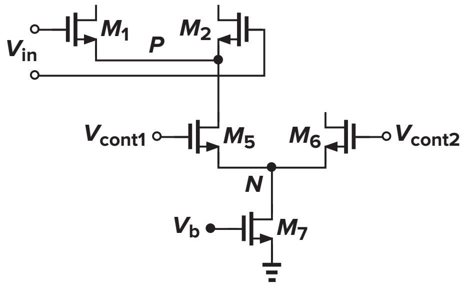


Figure 15.46 Headroom calculation for a current-steering topology.

the supply voltage must be higher by this value. Note also that if V_{cont1} or V_{cont2} is allowed to vary above its equilibrium value by more than V_{TH}, then M_5 or M_6 enters the triode region.

The previous observation reveals a trade-off between voltage headroom and the *sensitivity* of the VCO. In order to minimize the sensitivity with a given tuning range, the transconductance of M_5–M_6 must be *minimized*. (That is, to steer all of the tail current, the differential pair must require a *large* $V_{cont1} - V_{cont2}$.) However, for a given tail current, $g_m = 2I_D/(V_{GS} - V_{TH})$, indicating a large equilibrium overdrive for M_5–M_6 and a correspondingly higher value for the minimum required supply voltage.

We should mention that the pair M_5–M_6 need not remain in complete saturation. If the drain voltages are low enough to drive these transistors into the triode region, then the equivalent transconductance of the differential pair drops, demanding a greater $V_{cont1} - V_{cont2}$ to steer the tail current. This phenomenon in fact translates to a *lower* VCO sensitivity. In practice, careful simulations are required to ensure that the VCO characteristic remains relatively linear across the range of interest.[9]

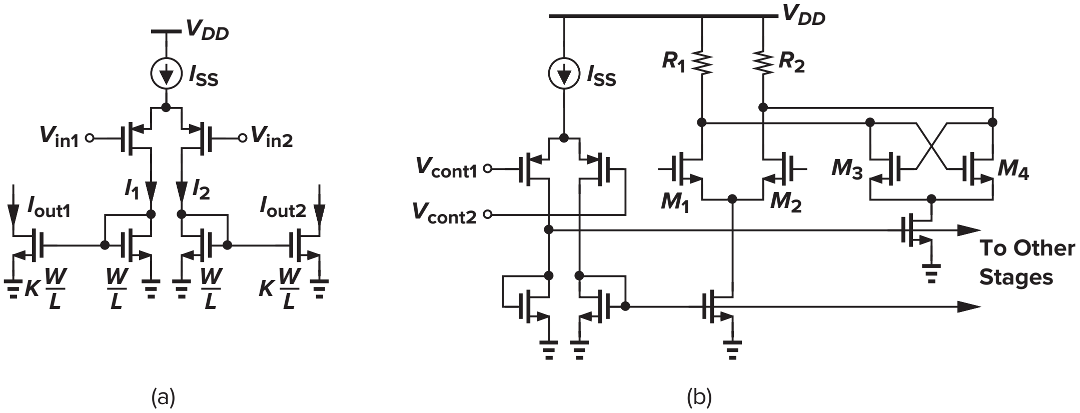


Figure 15.47 (a) Current folding topology; (b) application of current folding to current steering.

At low supply voltages, it is desirable to avoid the voltage headroom consumed by M_5–M_6 in Fig. 15.45. The issue can be resolved by means of “current folding.” Suppose, as illustrated in Fig. 15.47(a), a differential pair drives two current mirrors, generating I_{out1} and I_{out2}. Since $I_1 + I_2 = I_{SS}$, $I_{out1} = KI_1$, and $I_{out2} = KI_2$, we have $I_{out1} + I_{out2} = KI_{SS}$. Thus, as $V_{in1} - V_{in2}$ goes from a very negative value to a very positive value, I_{out1} varies from KI_{SS} to zero and I_{out2} from zero to KI_{SS} while their sum remains constant—a behavior similar to that of a differential pair.

[9] If both M_5 and M_6 are in the triode region and $V_{cont1} \neq V_{cont2}$, then supply voltage variations affect the current steered between the two transistors, introducing noise in the frequency of oscillation.

We now utilize the topology of Fig. 15.47(a) in the gain stage of Fig. 15.44(a). Shown in Fig. 15.47(b), the resulting circuit operates from a low supply voltage. However, the devices in the control path contribute substantial noise, modulating the oscillation frequency.

Delay Variation by Interpolation Another approach to tuning ring oscillators is based on "interpolation" [3, 4]. As illustrated in Fig. 15.48(a), each stage consists of a fast path and a slow path whose outputs are summed and whose gains are adjusted by V_{cont} in opposite directions. At one extreme of the control voltage, only the fast path is on and the slow path is disabled, yielding the maximum oscillation frequency [Fig. 15.48(b)]. Conversely, at the other extreme, only the slow path is on and the fast path is off, providing the minimum oscillation frequency [Fig. 15.48(c)]. If V_{cont} lies between the two extremes, each path is partially on, and the total delay is a weighted sum of their delays.

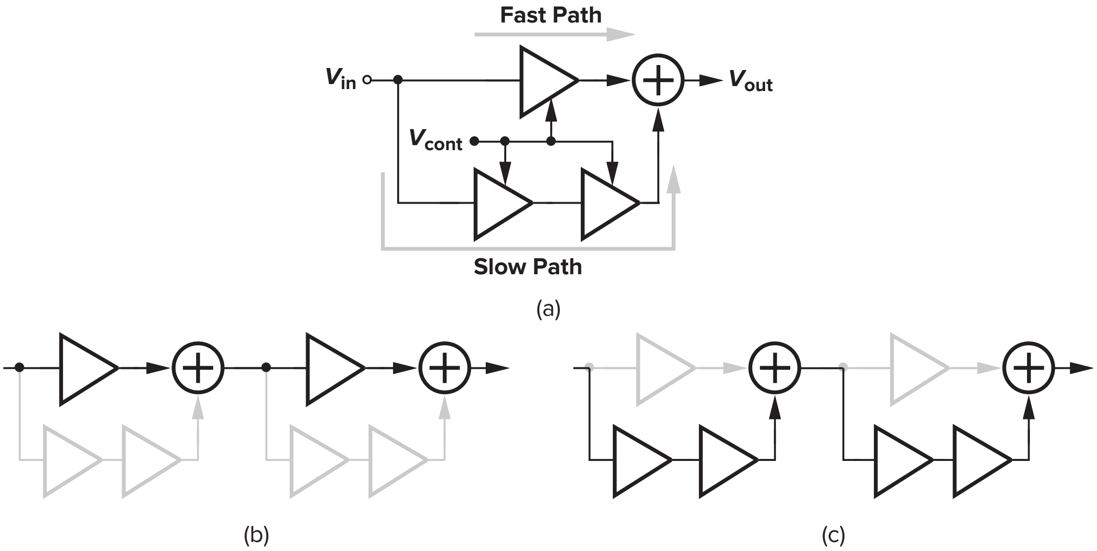


Figure 15.48 (a) Interpolating delay stage; (b) smallest delay; (b) largest delay.

To better understand the concept of interpolation, let us implement the topology of Fig. 15.48(a) at the transistor level. Each stage can be simply realized as a differential pair whose gain is controlled by its tail current. But how are the two outputs summed? Since the two transistors in a differential pair provide output *currents*, the outputs of the two pairs can be added in the current domain. As depicted in Fig. 15.49(a), simply shorting the outputs of two pairs performs the current addition, e.g., for small signals, $I_{out} = g_{m1,2}V_{in1} + g_{m3,4}V_{in2}$. The overall interpolating stage therefore assumes the configuration shown in Fig. 15.49(b), where V_{cont}^{+} and V_{cont}^{-} denote voltages that vary in opposite directions (so that when one path turns on, the other turns off). The output currents of M_1–M_2 and M_3–M_4 are summed at X and Y and flow through R_1 and R_2, producing V_{out}.

In the circuit of Fig. 15.49(b), the gain of each stage is varied by the tail current to achieve interpolation. But it is desirable to maintain constant voltage swings. We also recognize that the gain of the differential pair M_5–M_6 need not be varied because even if only the gain of M_3–M_4 drops to zero, the slow path is fully disabled. We then surmise that if the tail currents of M_1–M_2 and M_3–M_4 vary in opposite directions such that their sum remains constant, we achieve both interpolation between the two paths and constant output swings. Illustrated in Fig. 15.50, the resulting circuit employs the differential pair M_7–M_8 to steer I_{SS} between M_1–M_2 and M_3–M_4. If V_{cont} is very negative, M_8 is off and only the fast path amplifies the input. Conversely, if V_{cont} is very positive, M_7 is off and only the slow path is enabled. Since the

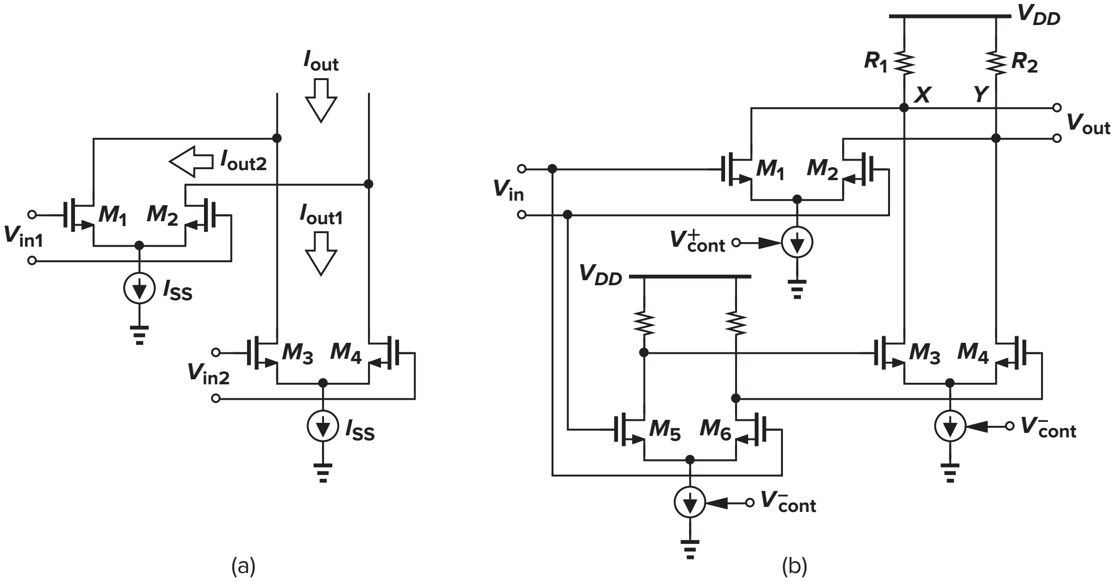


Figure 15.49 (a) Addition of currents of two differential pairs; (b) interpolating delay stage.

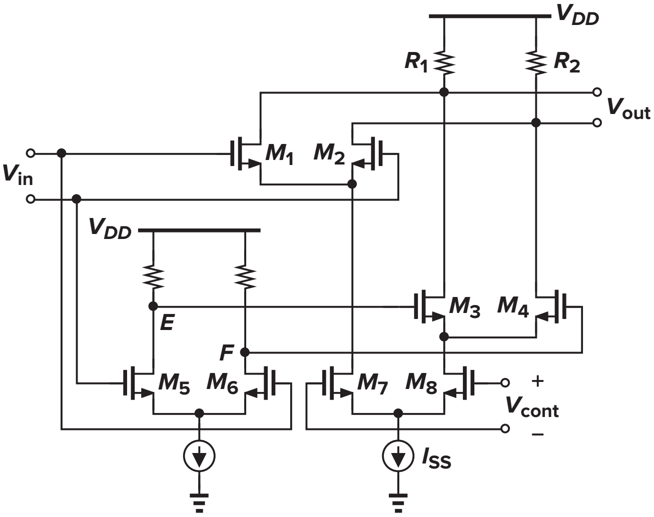


Figure 15.50 Interpolating delay stage with current steering.

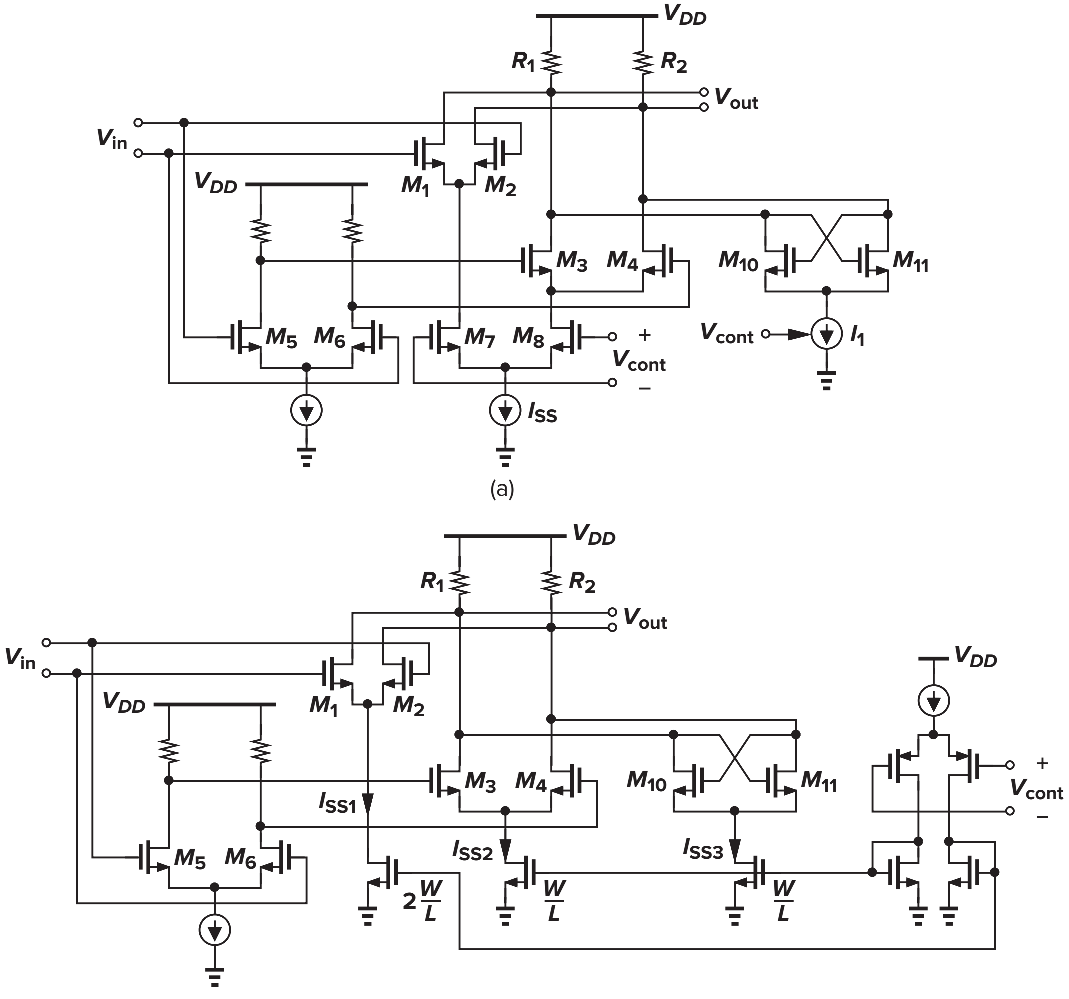
slow path in this case employs one more stage than the fast path, the VCO achieves a tuning range of roughly two to one. For operation with low supply voltages, the control pair M_7–M_8 can be replaced by the current-folding topology of Fig. 15.47(a).

▶ Example 15.13

Combine the tuning techniques of Figs. 15.45 and 15.50 to achieve a wider tuning range.

Solution

We begin with the interpolating stage of Fig. 15.50 and add a cross-coupled pair to the output nodes [Fig. 15.51(a)]. However, in order to obtain constant voltage swings, the total current through the load resistors must remain



(a)

(b)

Figure 15.51

constant. This is accomplished by replacing the control differential pair with the current-folding circuit of Fig. 15.47(a). Depicted in Fig. 15.51(b), the resulting configuration steers the current to M_1–M_2 to speed up the circuit and to M_3–M_4 and M_{10}–M_{11} to slow down the circuit. The tail current source dimensions are chosen such that $I_{SS1} = I_{SS2} + I_{SS3}$.

Wide-Range Tuning Except for the circuit of Fig. 15.43(b), the ring oscillator tuning techniques presented thus far achieve a tuning range of typically no more than three to one. In applications where the frequency must be varied by orders of magntitude, the topology shown in Fig. 15.52 can be used. Driven by the input, the additional PMOS transistors M_5 and M_6 pull each output node to V_{DD}, creating a relatively constant output swing even with large variations in I_{SS}. The oscillation frequency of a ring incorporating this stage can be varied by more than four orders of magnitude with less than a twofold variation in the amplitude.

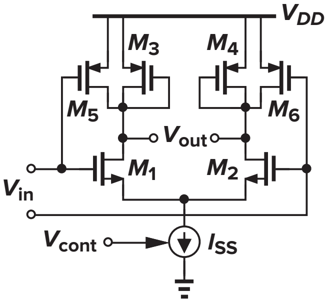


Figure 15.52 Differential stage with wide tuning range.

15.4.2 Tuning in LC Oscillators

The oscillation frequency of LC topologies is equal to $f_{osc} = 1/(2\pi\sqrt{LC})$, suggesting that only the inductor and capacitor values can be varied to tune the frequency, and other parameters such as bias currents and transistor transconductances affect f_{osc} negligibly. Since it is difficult to vary the value of monolithic inductors, we simply change the tank capacitance to tune the oscillator. Voltage-dependent capacitors are called "varactors."[10]

A reverse-biased *pn* junction can serve as a varactor. The voltage dependence is expressed as

$$C_{var} = \frac{C_0}{\left(1 + \frac{V_R}{\phi_B}\right)^m} \tag{15.72}$$

where C_0 is the zero-bias value, V_R the reverse-bias voltage, ϕ_B the built-in potential of the junction, and m a value typically between 0.3 and 0.4.[11] Equation (15.72) reveals an important drawback of LC oscillators: at low supply voltages, V_R has a very limited range, yielding a small range for C_{var} and hence for f_{osc}. We also note that to maximize the tuning range, constant capacitances in the tank must be *minimized*.

▶ **Example 15.14**

Suppose that in Eq. (15.72), $\phi_B = 0.7$ V, $m = 0.35$, and V_R can vary from zero to 2 V. How much tuning range can be achieved?

Solution

For $V_R = 0, C_j = C_0$ and $f_{osc,min} = 1/(2\pi\sqrt{LC_0})$. For $V_R = 2$ V, $C_j \approx 0.62C_0$ and $f_{osc,max} = 1/(2\pi\sqrt{L \times 0.62C_0}) \approx 1.27 f_{osc,min}$. Thus, the tuning range is approximately equal to 27%. As explained later, the parasitic capacitances of the inductor and the transistor(s) further limit this range because they cannot be varied by the control voltage.

◀

Let us now add varactor diodes to a cross-coupled LC oscillator (Fig. 15.53). To avoid forward-biasing D_1 and D_2 significantly, V_{cont} must not exceed V_X or V_Y by more than a few hundred millivolts. Thus, if the peak amplitude at each node is A, then $0 < V_{cont} < V_{DD} - A + 300$ mV, where it is assumed that a forward bias of 300 mV creates negligible current. Interestingly, the circuit suffers from a trade-off between the output swing and the tuning range. This effect appears in most LC oscillators.

[10] The term "varicap" is also used.

[11] Note that $m = 0.5$ for an abrupt junction, but *pn* junctions in CMOS technology are not abrupt.

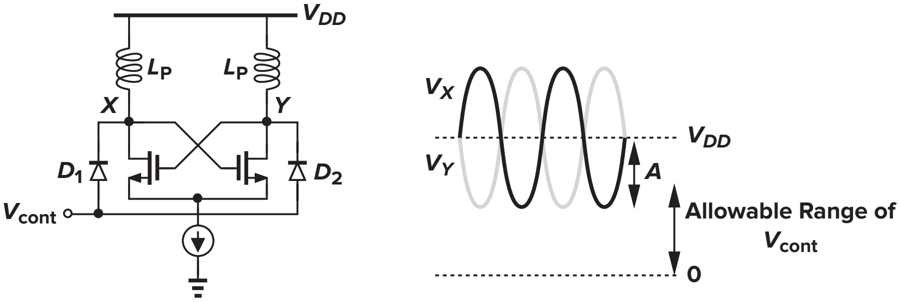


Figure 15.53 LC oscillator using varactor diodes.

Note that, since the swings at X and Y are typically large (e.g., 1 V_{pp} at each node), the capacitance of D_1 and D_2 *varies* with time. Nonetheless, the "average" value of the capacitance is still a function of V_{cont}, providing the tuning range.

How are varactor diodes realized in CMOS technology? Illustrated in Fig. 15.54 are two types of *pn* junctions. In Fig. 15.54(a), the anode is inevitably grounded whereas in Fig. 15.54(b), both terminals are floating. For the circuit of Fig. 15.53, only the floating diode can be used. To increase the capacitance of the junction, the p^+ and n^+ areas (and hence the *n*-well) are enlarged.

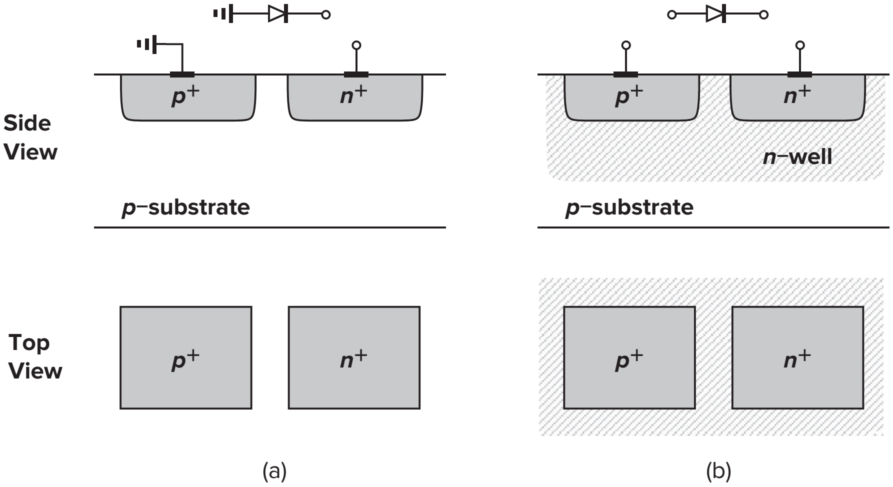


Figure 15.54 Diodes realized in CMOS technology.

Upon closer examination, the structure of Fig. 15.54(b) suffers from a number of drawbacks. First, the *n*-well material has a high resistivity, creating a resistance in series with the reverse-biased diode and lowering the quality factor of the capacitance. Second, the *n*-well displays substantial capacitance to the substrate, contributing a constant capacitance to the tank and limiting the tuning range. The diode is therefore represented as shown in Fig. 15.55, where C_n represents the (voltage-dependent) capacitance between the *n*-well and the substrate.[12]

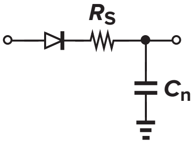


Figure 15.55 Circuit model of the varactor shown in Fig. 15.54(b).

[12] In circuit simulations, C_n is replaced by a diode having proper junction capacitance.

In order to decrease the series resistance of the structure shown in Fig. 15.54(b), the p^+ region can be surrounded by an n^+ ring so that the displacement current flowing through the junction capacitance sees a low resistance in all four directions [Fig. 15.56(a)]. Since a single minimum-size p^+ area has a small capacitance, many of these units can be placed in parallel [Fig. 15.56(b)]. The n-well, however, must accommodate the entire set, exhibiting a large capacitance to the substrate.

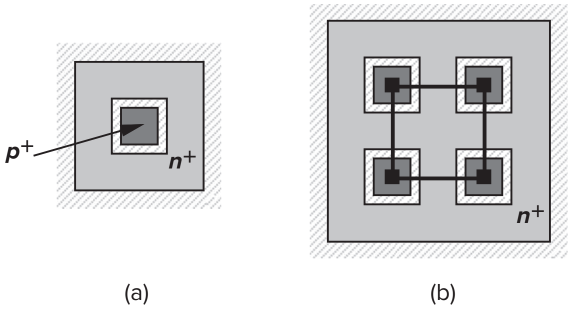


Figure 15.56 (a) Reduction of series resistance by surrounding the p^+ region by an n^+ ring; (b) several diodes in parallel.

It is instructive at this point to examine the unwanted capacitances in the circuit of Fig. 15.53, i.e., the components that are not varied by V_{cont}. We identify three such capacitances: (1) the capacitance between the n-well and the substrate associated with D_1 and D_2; (2) the capacitances contributed by the transistors to each node, i.e., C_{GD}, $2C_{GD}$ (the factor of 2 arising from the Miller effect[13]), and C_{DB}; and (3) the parasitic capacitance of the inductor itself. Monolithic inductors are typically implemented as metal spiral structures (Fig. 15.57) having relatively large dimensions ($S \approx 100-200\ \mu$m).

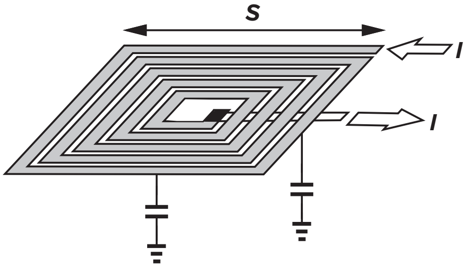


Figure 15.57 Spiral inductor structure.

In Fig. 15.53, it is desirable to connect the anode of the diodes to nodes X and Y, thereby eliminating the parasitic n-well capacitances from the tank. Shown in Fig. 15.58 is a topology allowing such a modification. Here, the cross-coupled pair incorporates PMOS devices, providing swings around the ground potential. The use of PMOS devices also leads to less flicker noise, an important advantage because this noise may be "upconverted," appearing around the oscillation frequency.

In modern LC VCO design, we employ MOS varactors. Recall from Chapter 2 that the gate-channel capacitance of MOSFETs varies with the gate-source voltage [Fig. 15.59(a)]. However, the *nonmonotonic* dependence proves undesirable in VCO design (why?). To resolve this issue, an NMOS transistor can be placed inside an n-well, forming an "accumulation-mode" varactor [Fig. 15.59(b)]. The source, drain, and n-well are ohmically connected and serve as one terminal, and the gate as the other. The capacitance of this structure varies monotonically with V_{GS}, as shown in Fig. 15.59(c).

[13] If the gate and drain voltages vary by equal and opposite amounts, the Miller multiplication factor is equal to 2 regardless of the small-signal gain.

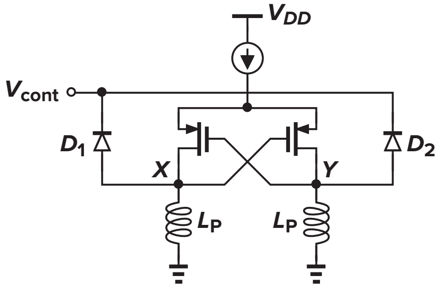


Figure 15.58 Negative-G_m oscillator using PMOS devices to eliminate n-well capacitance from the tanks.

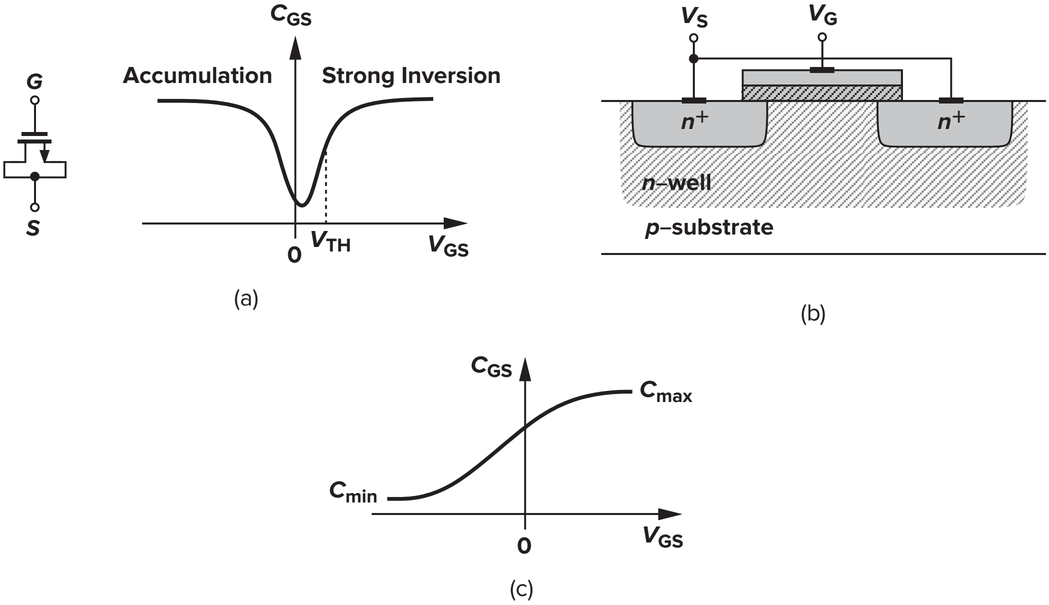


Figure 15.59 (a) Voltage dependence of a MOS gate capacitance, (b) MOS varactor formed as an NFET inside an n-well, and (c) resulting characteristic.

An important advantage of the MOS varactor over the pn junction is that the former does not experience forward bias and can therefore tolerate both positive and negative voltages. The design of LC VCOs entails numerous interesting concepts and issues. The reader is referred to [5] and the vast literature on the subject for details.

15.5 ■ Mathematical Model of VCOs

The definition of the voltage-controlled oscillator given by Eq. (15.55) specifies the relationship between the control voltage and the output frequency. The dependence is "memoryless" because a change in V_{cont} immediately results in a change in ω_{out}. But how is the output signal of the VCO expressed as a function of time? To answer this question, we must review the concepts of phase and frequency.

Consider the waveform $V_0(t) = V_m \sin \omega_0 t$. The argument of the sinusoid is called the "total phase" of the signal. In this example, the phase varies linearly with time, exhibiting a slope equal to ω_0. Note that, as depicted in Fig. 15.60, every time $\omega_0 t$ crosses an integer multiple of π, $V_0(t)$ crosses zero.

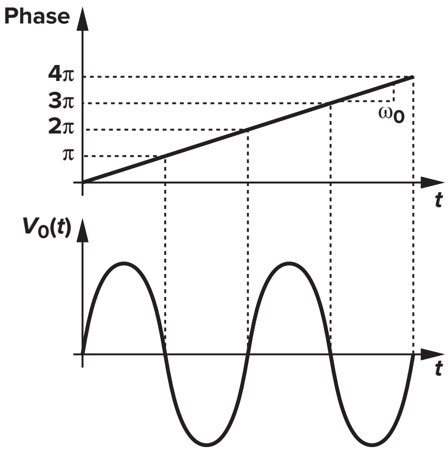


Figure 15.60 Illustration of phase of a signal.

Now consider two waveforms $V_1(t) = V_m \sin[\phi_1(t)]$ and $V_2(t) = V_m \sin[\phi_2(t)]$, where $\phi_1(t) = \omega_1 t$, $\phi_2(t) = \omega_2 t$, and $\omega_1 < \omega_2$. As illustrated in Fig. 15.61, $\phi_2(t)$ crosses integer multiples of π faster than $\phi_1(t)$ does, yielding faster variations in $V_2(t)$. We say that $V_2(t)$ accumulates phase faster.

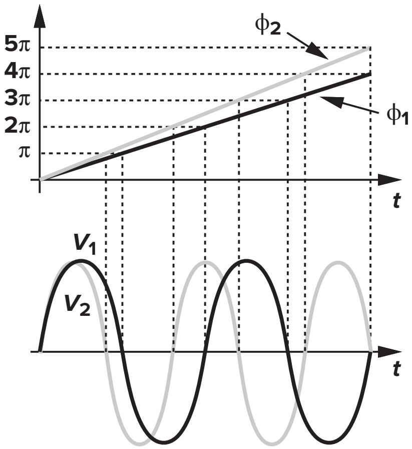


Figure 15.61 Variation of phase for two signals.

The above study reveals that the faster the phase of a waveform varies, the higher the frequency of the waveform, suggesting that the frequency[14] can be defined as the derivative of the phase with respect to time:

$$\omega = \frac{d\phi}{dt} \tag{15.73}$$

▶ Example 15.15

Figure 15.62(a) shows the phase of a sinusoidal waveform with constant amplitude as a function of time. Plot the waveform in the time domain.

[14] The quantity $\omega = 2\pi f$ is called the "radian frequency" (and expressed in rad/s) to distinguish it from f (expressed in Hz). In this book, we call both the frequency, but use ω more often to avoid the factor 2π.

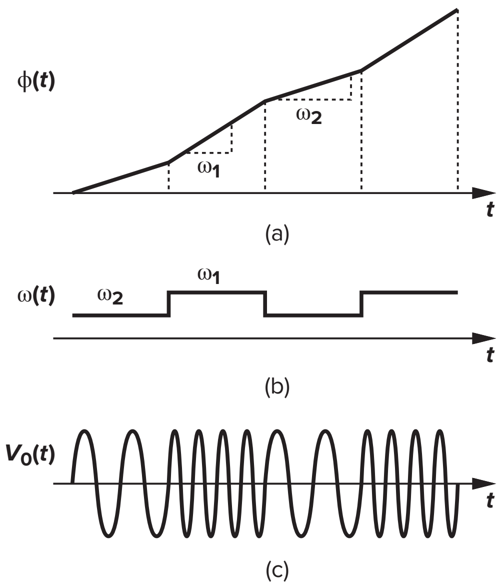


Figure 15.62

Solution

Taking the time derivative of $\phi(t)$, we obtain the behavior illustrated in Fig. 15.62(b). The frequency therefore periodically toggles between ω_1 and ω_2, yielding the waveform shown in Fig. 15.62(c). (This is a simple example of binary frequency modulation, called "frequency shift keying" and utilized in wireless pagers and many other communication systems.)

Equation (15.73) indicates that, if the frequency of a waveform is known as a function of time, then the phase can be computed as

$$\phi = \int \omega dt + \phi_0 \tag{15.74}$$

In particular, since for a VCO, $\omega_{out} = \omega_0 + K_{VCO} V_{cont}$, we have

$$V_{out}(t) = V_m \cos\left(\int \omega_{out} dt + \phi_0\right) \tag{15.75}$$

$$= V_m \cos\left(\omega_0 t + K_{VCO} \int V_{cont} dt + \phi_0\right) \tag{15.76}$$

Equation (15.76) proves essential in the analysis of VCOs and PLLs.[15] The initial phase ϕ_0 is usually unimportant and is assumed zero hereafter.

▶ Example 15.16

The control line of a VCO senses a rectangular signal toggling between V_1 and V_2 at a period T_m. Plot the frequency, phase, and output waveform as a function of time.

[15] Note that K_{VCO} cannot be brought out of the integral if the characteristic is nonlinear.

Solution

Since $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$, the output frequency toggles between $\omega_1 = \omega_0 + K_{VCO}V_1$ and $\omega_2 = \omega_0 + K_{VCO}V_2$ (Fig. 15.63). The phase is equal to the time integral of this result, rising linearly with time at a slope of ω_1 for half the input period and ω_2 for the other half. The output waveform of the VCO is similar to that shown in Fig. 15.62. Thus, a VCO can operate as a frequency modulator.

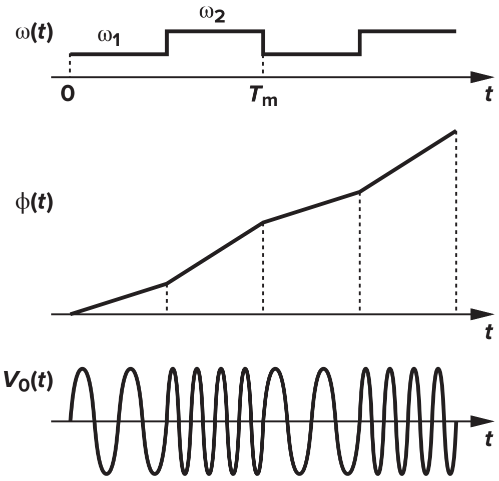


Figure 15.63

As explained in Chapter 16, if a VCO is placed in a phase-locked loop, then only the second term of the total phase in Eq. (15.76) is of interest. This term, $K_{VCO}\int V_{cont}dt$, is called the "excess phase," ϕ_{ex}. In fact, in the analysis of PLLs, we view the VCO as a system whose input and output are the control voltage and the excess phase, respectively:

$$\phi_{ex} = K_{VCO}\int V_{cont}dt \tag{15.77}$$

That is, the VCO operates as an *ideal* integrator, providing a transfer function:

$$\frac{\Phi_{ex}}{V_{cont}}(s) = \frac{K_{VCO}}{s} \tag{15.78}$$

Example 15.17

A VCO senses a small sinusoidal control voltage $V_{cont} = V_m \cos\omega_m t$. Determine the output waveform and its spectrum.

Solution

The output is expressed as

$$V_{out}(t) = V_0 \cos\left(\omega_0 t + K_{VCO}\int V_{cont}dt\right) \tag{15.79}$$

$$= V_0 \cos\left(\omega_0 t + K_{VCO}\frac{V_m}{\omega_m}\sin\omega_m t\right) \tag{15.80}$$

$$= V_0 \cos\omega_0 t \cos\left(K_{VCO}\frac{V_m}{\omega_m}\sin\omega_m t\right) \tag{15.81}$$

$$-V_0 \sin\omega_0 t \sin\left(K_{VCO}\frac{V_m}{\omega_m}\sin\omega_m t\right)$$

If V_m is small enough that $K_{VCO}V_m/\omega_m \ll 1$ rad, then

$$V_{out}(t) \approx V_0 \cos\omega_0 t - V_0(\sin\omega_0 t)(K_{VCO}\frac{V_m}{\omega_m}\sin\omega_m t) \tag{15.82}$$

$$= V_0 \cos\omega_0 t - \frac{K_{VCO}V_m V_0}{2\omega_m}[\cos(\omega_0 - \omega_m)t - \cos(\omega_0 + \omega_m)t] \tag{15.83}$$

The output therefore consists of three sinusoids having frequencies of ω_0, $\omega_0 - \omega_m$, and $\omega_0 + \omega_m$. The spectrum is shown in Fig. 15.64. The components at $\omega_0 \pm \omega_m$ are called "sidebands."

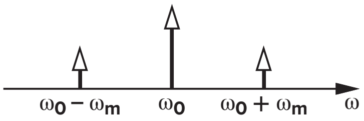


Figure 15.64

The above example reveals that variation of the control voltage with time may create unwanted components at the output. Indeed, when a VCO operates in the steady state, the control voltage must experience very little variation.[16] This issue is studied in Chapter 16.

A common mistake in expressing the phase of signals arises from the familiar form $V_m \cos\omega_0 t$. Here, the phase is equal to the product of frequency and time, creating the impression that such equality holds in all conditions. We may even deduce that, since the output frequency of a VCO is given by $\omega_0 + K_{VCO}V_{cont}$, the output waveform can be written as $V_m \cos[(\omega_0 + K_{VCO}V_{cont})t]$. To understand why this is incorrect, let us compute the frequency as the derivative of the phase:

$$\omega = \frac{d}{dt}[(\omega_0 + K_{VCO}V_{cont})t] \tag{15.84}$$

$$= K_{VCO}\frac{dV_{cont}}{dt}t + \omega_0 + K_{VCO}V_{cont} \tag{15.85}$$

The first term in this expression is redundant, vanishing only if $dV_{cont}/dt = 0$. Thus, in the general case, the phase cannot be written as the product of time and frequency.

Our study of VCOs in this section has assumed sinusoidal output waveforms. In practice, depending on the type and speed of the oscillator, the output may contain significant harmonics, even approaching a rectangular waveform. How should Eq. (15.76) be modified in this case? We expect that $V_{out}(t)$ can be expressed as a Fourier series:

$$V_{out}(t) = V_1\cos(\omega_0 t + \phi_1) + V_2\cos(2\omega_0 t + \phi_2) + \cdots \tag{15.86}$$

We also note that if the (fundamental) frequency of a rectagular waveform is changed by Δf, the frequency of its second harmonic must change by $2\Delta f$, etc. Thus, if V_{cont} varies by ΔV, then the frequency of the first harmonic varies by $K_{VCO}\Delta V$, the frequency of the second harmonic by $2K_{VCO}\Delta V$, etc. That is

$$V_{out}(t) = V_1\cos(\omega_0 t + K_{VCO}\int V_{cont}dt + \theta_1) + V_2\cos(2\omega_0 t + 2K_{VCO}\int V_{cont}dt + \theta_2) + \cdots \tag{15.87}$$

[16] Except when the VCO senses a signal to perform frequency modulation.

where $\theta_1, \theta_2, \cdots$ are constant phases necessary for the representation of each harmonic in the Fourier series expansion.

Equation (15.87) suggests that the harmonics of an oscillator output can be readily taken into account. For this reason, we often limit our calculations to the first harmonic, even though we may draw the waveforms in rectangular shape rather than sinusoidal shape.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

15.1. For the circuit of Fig. 15.6, determine the open-loop tranfer function and calculate the phase margin. Assume that $g_{m1} = g_{m2} = g_m$ and neglect other capacitances.

15.2. In the circuit of Fig. 15.8, assume that $g_{m1} = g_{m2} = g_{m3} = (200\ \Omega)^{-1}$.
(a) What is the minimum value of R_D that ensures oscillation?
(b) Determine the value of C_L for an oscillation frequency of 1 GHz and a total low-frequency loop gain of 16.

15.3. For the circuit of Fig. 15.12, determine the minimum value of I_{SS} that guarantees oscillation. (Hint: if the circuit is at the edge of oscillation, the swings are quite small.)

15.4. Prove that the small-signal resistance of the composite load in Fig. 15.18(c) is roughly equal to $1/g_{m3}$.

15.5. Including only the gate-source capacitance of M_3 in Fig. 15.18(c), explain under what condition the impedance of the composite load (seen at the drain of M_3) becomes inductive.

15.6. If each inductor in Fig. 15.25 exhibits a series resistance of R_S, how low must R_S be to ensure that the low-frequency loop gain is less than unity? (This condition is necessary to avoid latch-up.)

15.7. Explain why the V_X and V_Y waveforms in Fig. 15.28 are closer to sinusoids (i.e., they contain smaller harmonics) than are the I_{D1} and I_{D2} waveforms do.

15.8. Determine the minimum value of I_{SS} in Fig. 15.47(c) that guarantees oscillation. Estimate the maximum value of I_{SS} that guarantees that M_1 and M_2 do not enter the triode region.

15.9. Repeat Example 15.7 by applying a current stimulus to the drain of M_1.

15.10. Prove that if a capacitor C_P is placed in parallel with L_P in Fig. 15.31(a), then Eq. (15.47) results.

15.11. The Colpitts oscillator of Fig. 15.31(a) was analyzed and its oscillation conditions were derived by applying a current stimulus to the source. Repeat the analysis by applying a voltage stimulus to the gate of M_1.

15.12. Repeat the analysis of the Colpitts oscillator for the topologies in Figs. 15.38(a) and (c). Determine the oscillation condition and the frequency of oscillation.

15.13. The stage of Fig. 15.45 is designed with $I_T = 1$ mA and $(W/L)_{1,2} = 50/0.5$. Assume that $I_H \ll I_1$.
(a) Determine the minimum value of $R_1 = R_2 = R$ to ensure oscillation in a three-stage ring.
(b) Determine $(W/L)_{3,4}$ such that $g_{m3,4}R = 0.5$ when each of M_3 and M_4 carries $I_T/2$.
(c) Calculate the minimum value of I_H to guarantee oscillation.
(d) If the common-mode level of V_{cont1} and V_{cont2} is 1.5 V, calculate $(W/L)_{5,6}$ such that I_T sustains 0.5 V when $V_{cont1} = V_{cont2}$.

15.14. Repeat Example 15.14 if each inductor in the circuit contributes a constant capacitance equal to C_1.

15.15. The VCO of Fig. 15.53 is designed for operation at 1 GHz.
(a) If $L_P = 5$ nH and the total (fixed) parasitic capacitance seen at X (and Y) to ground is 500 fF, determine the maximum capacitance that D_1 and D_2 can add to the circuit.
(b) If the tail current is equal to 1 mA and the Q of each inductor at 1 GHz is equal to 4, estimate the output voltage swing.

References

[1] N. M. Nguyen and R. G. Meyer, "Start-up and Frequency Stability in High-Frequency Oscillators," *IEEE J. of Solid-State Circuits,* vol. 27, pp. 810–820, May 1992.

[2] I. A. Young, J. K. Greason, and K. L. Wong, "A PLL Clock Generator with 5 to 110 MHz of Lock Range for Microprocessors," *IEEE J. of Solid-State Circuits,* vol. SC-27, pp. 1599–1607, November 1992.

[3] B. Lai and R. C. Walker, "A Monolithic 622 Mb/sec Clock Extraction and Data Retiming Circuit," *ISSCC Dig. of Tech. Papers,* pp. 144–145, February 1991.

[4] S. K. Enam and A. A. Abidi, "NMOS ICs for Clock and Data Regeneration in Gigabit-per-Second Optical-Fiber Receivers," *IEEE J. of Solid-State Circuits,* vol. SC-27, pp. 1763–1774, December 1992.

[5] B. Razavi, *RF Microelectronics*, 2nd ed. (Upper Saddle River, NJ: Prentice-Hall, 2012).

CHAPTER

16

Phase-Locked Loops

The concept of phase locking was invented in the 1930s and swiftly found wide usage in electronics and communication. While the basic phase-locked loop has remained nearly the same since then, its implementation in different technologies and for different applications continues to challenge designers. A PLL serving the task of clock generation in a microprocessor appears quite similar to a frequency synthesizer used in a cellphone, but the actual circuits are designed quite differently.

This chapter deals with the analysis and design of PLLs, with particular attention to implementations in VLSI technologies. A thorough study of PLLs would require an entire book by itself, but our objective here is to lay the foundation for more advanced work. Beginning with a simple PLL architecture, we study the phenomenon of phase locking and analyze the behavior of PLLs in the time and frequency domains. We then address the problem of lock acquisition and describe charge-pump PLLs (CPPLLs) and their nonidealities. Finally, we examine jitter in PLLs, study delay-locked loops (DLLs), and present a number of PLL applications.

16.1 ■ Simple PLL

A PLL is a feedback system that compares the output phase with the input phase. The comparison is performed by a "phase comparator" or "phase detector" (PD). It is therefore beneficial to define the PD rigorously.

16.1.1 Phase Detector

A phase detector is a circuit whose average output, $\overline{V_{out}}$, is linearly proportional to the phase difference, $\Delta\phi$, between its two inputs (Fig. 16.1). In the ideal case, the relationship between $\overline{V_{out}}$ and $\Delta\phi$ is linear, crossing the origin for $\Delta\phi = 0$. Called the "gain" of the PD, the slope of the line, K_{PD}, is expressed in V/rad.

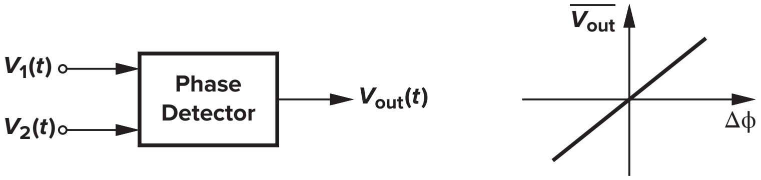


Figure 16.1 Definition of phase detector.

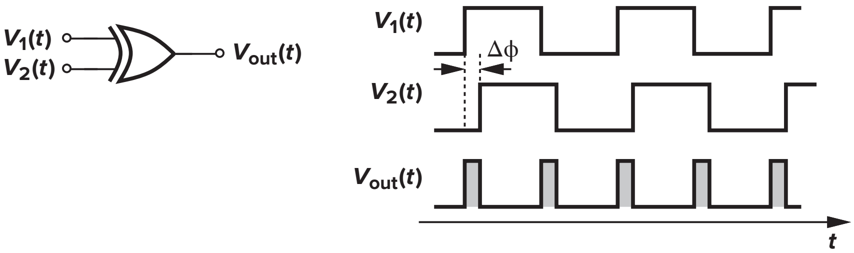


Figure 16.2 Exclusive OR gate as phase detector.

A familiar example of a phase detector is the exclusive OR (XOR) gate. As shown in Fig. 16.2, as the phase difference between the inputs varies, so does the width of the output pulses, thereby providing a dc level proportional to $\Delta\phi$. While the XOR circuit produces error pulses on both rising and falling edges, other types of PD may respond only to positive or negative transitions.

▶ Example 16.1

If the output swing of the XOR in Fig. 16.2 is V_0 volts, what is the gain of the circuit as a phase detector? Plot the input-output characteristic of the PD.

Solution

If the phase difference increases from zero to $\Delta\phi$ radians, the area under each pulse increases by $V_0 \cdot \Delta\phi$. Since each period contains *two* pulses, the average value rises by $2[V_0 \cdot \Delta\phi/(2\pi)]$, yielding a gain of V_0/π. Note that the gain is independent of the input frequency.

To construct the input-output characteristic, we examine the circuit's response to various input phase differences. As illustrated in Fig. 16.3, the average output voltage rises to $[V_0/\pi] \times \pi/2 = V_0/2$ for $\Delta\phi = \pi/2$ and V_0 for

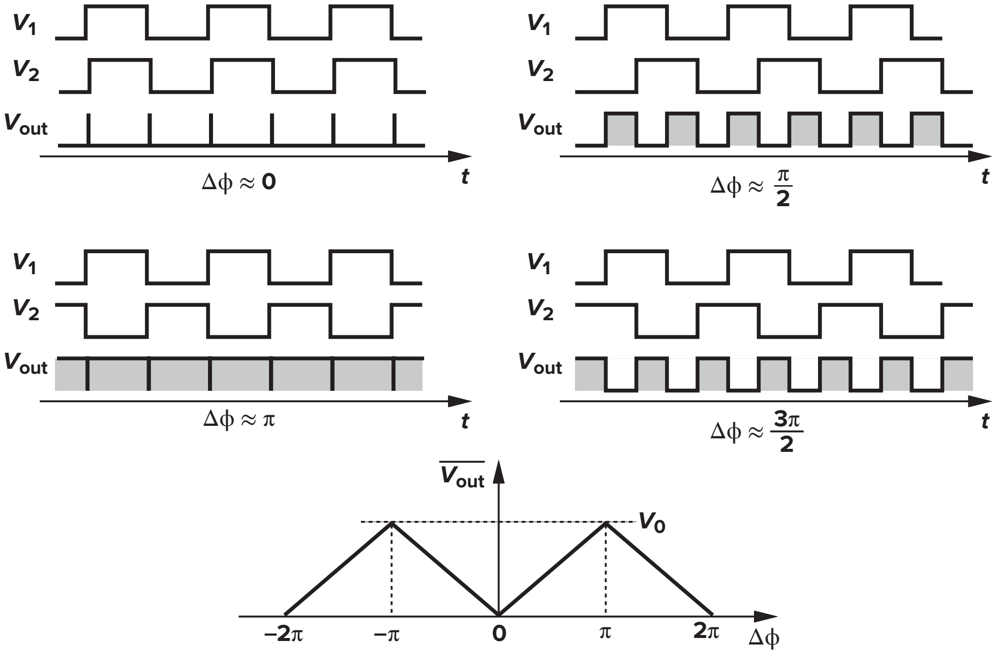


Figure 16.3

$\Delta\phi = \pi$. For $\Delta\phi > \pi$, the average begins to *drop*, falling to $V_0/2$ for $\Delta\phi = 3\pi/2$ and zero for $\Delta\phi = 2\pi$. The characteristic is therefore periodic, exhibiting both negative and positive gains.

16.1.2 Basic PLL Topology

To arrive at the concept of phase locking, let us consider the problem of aligning the output phase of a VCO with the phase of a reference clock. (The reader is encouraged to review the VCO mathematical model in the previous chapter.) As illustrated in Fig. 16.4(a), the rising edges of V_{out} are "skewed" by Δt seconds with respect to V_{CK}, and we wish to eliminate this error. Assuming that the VCO has a single control input, V_{cont}, we note that to vary the phase, we *must* vary the frequency and allow the integration $\phi = \int(\omega_0 + K_{VCO}V_{cont})dt$ to take place. For example, suppose that, as shown in Fig. 16.4(b), the VCO frequency is stepped to a higher value at $t = t_1$. The circuit then accumulates phase faster, gradually decreasing the phase error. At $t = t_2$, the phase error drops to zero and, if V_{cont} returns to its original value, V_{VCO} and V_{CK} remain aligned. Interestingly, the alignment can be accomplished by stepping the VCO frequency to a *lower* value for a certain time interval as well (Problem 16.2). Thus, phase alignment can be achieved only by a (temporary) frequency change.

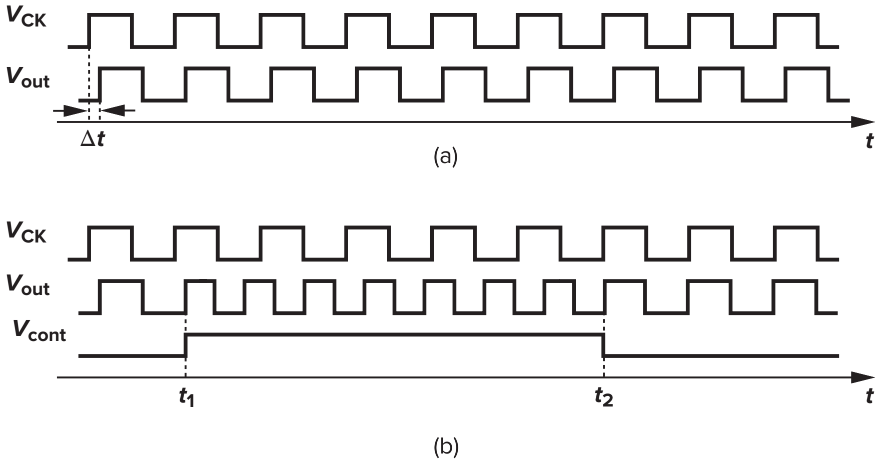


Figure 16.4 (a) Two waveforms with a skew; (b) change of VCO frequency to eliminate the skew.

The foregoing experiment suggests that the output phase of a VCO can be aligned with the phase of a reference if (1) the frequency of the VCO is changed momentarily, and (2) a means of comparing the two phases, i.e., a phase detector, is used to determine when the VCO and the reference signals are aligned. The task of aligning the output phase of the VCO with the phase of the reference is called "phase locking."

From the above observations, we surmise that a PLL simply consists of a PD and a VCO in a feedback loop [Fig. 16.5(a)]. The PD compares the phases of V_{out} and V_{in}, generating an error that varies the VCO frequency until the phases are aligned, i.e., the loop is locked. This topology, however, must be modified

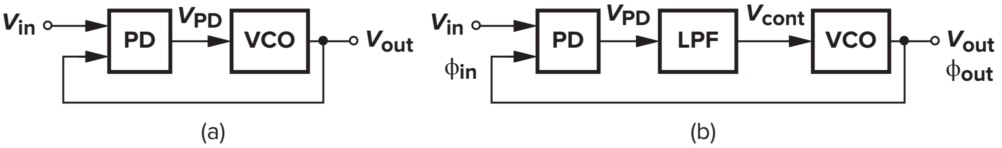


Figure 16.5 (a) Feedback loop comparing input and output phases; (b) simple PLL.

because (1) as exemplified by the waveforms of Fig. 16.2, the PD output, V_{PD}, consists of a dc component (desirable) and high-frequency components (undesirable), and (2) as mentioned in Chapter 15, the control voltage of the oscillator must remain quiet in the steady state, i.e., the PD output must be filtered. We therefore interpose a low-pass filter (LPF) between the PD and the VCO [Fig. 16.5(b)], suppressing the high-frequency components of the PD output and presenting the dc level to the oscillator. This forms the basic PLL topology. For now, we assume that the LPF has a gain of unity at low frequencies (e.g., as in a first-order RC section).

It is important to bear in mind that the feedback loop of Fig. 16.5(b) compares the *phases* of the input and output. Unlike the feedback topologies studied in the previous chapters, PLLs typically require no knowledge of voltages or currents in their feedback operation. If the loop gain is large enough, the difference between the input phase, ϕ_{in}, and the output phase, ϕ_{out}, falls to a small value in the steady state, providing phase alignment.

For subsequent analyses of PLLs, we must define the phase-lock condition carefully. If the loop of Fig. 16.5(b) is locked, we postulate that $\phi_{out} - \phi_{in}$ is constant and preferably small. We therefore define the loop to be locked if $\phi_{out} - \phi_{in}$ does not change with time. An important corollary of this definition is that

$$\frac{d\phi_{out}}{dt} - \frac{d\phi_{in}}{dt} = 0 \tag{16.1}$$

and hence

$$\omega_{out} = \omega_{in} \tag{16.2}$$

This is a unique property of PLLs and will be revisited more closely later.

In summary, when locked, a PLL produces an output that has a small phase error with respect to the input but exactly the same frequency. The reader may then wonder why a PLL is used at all. A short piece of wire would seem to perform the task even better! We answer this question in Sec. 16.5.

▶ **Example 16.2**

Implement a simple PLL in CMOS technology.

Solution

Figure 16.6 illustrates an implementation utilizing an XOR gate as the phase detector. The VCO is configured as a negative-G_m LC oscillator whose frequency is tuned by varactor diodes.

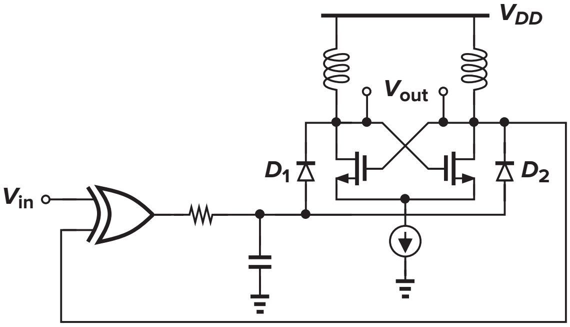


Figure 16.6

◀

PLL Waveforms in Locked Condition In order to familiarize ourselves with the behavior of PLLs, we begin with the simplest case: the circuit is locked and we wish to examine the waveforms at each point around the loop. As illustrated in Fig. 16.7(a), V_{in} and V_{out} exhibit a small phase difference but equal frequencies. The PD therefore generates pulses as wide as the skew between the input and the output,[1] and the low-pass filter extracts the dc component of V_{PD}, applying the result to the VCO. We assume that the LPF has a gain of unity at low frequencies. The small pulses in V_{LPF} are called "ripple."

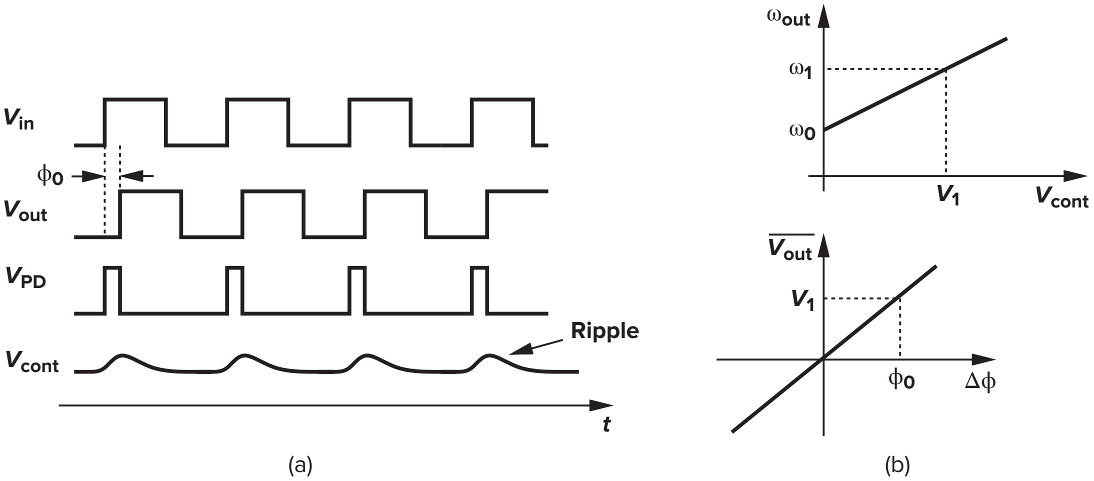


Figure 16.7 (a) Waveforms in a PLL in locked condition; (b) calculation of phase error.

In the waveforms of Fig. 16.7(a), two quantities are unknown: ϕ_0 and the dc level of V_{cont}. To determine these values, we construct the VCO and PD characteristics [Fig. 16.7(b)]. If the input and output frequencies are equal to ω_1, then the required oscillator control voltage is unique and equal to V_1. This voltage must be produced by the phase detector, demanding a phase error determined by the PD characteristic. More specifically, since $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$ and $\overline{V_{PD}} = K_{PD}\Delta\phi$, we can write

$$V_1 = \frac{\omega_1 - \omega_0}{K_{VCO}} \tag{16.3}$$

and

$$\phi_0 = \frac{V_1}{K_{PD}} \tag{16.4}$$

$$= \frac{\omega_1 - \omega_0}{K_{PD}K_{VCO}} \tag{16.5}$$

Equation (16.5) reveals two important points: (1) as the input frequency of the PLL varies, so does the phase error; and (2) to minimize the phase error, $K_{PD}K_{VCO}$ must be maximized.

▶ **Example 16.3**

A PLL incorporates a VCO and a PD having the characteristics shown in Fig. 16.8. Explain what happens as the input frequency varies in the locked condition.

[1] In this example, the PD produces pulses only on the rising transitions.

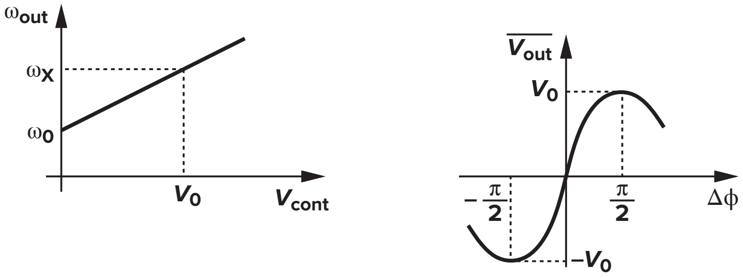


Figure 16.8

Solution

The PD characteristic is relatively linear near the origin but exhibits a small-signal gain of zero if the phase difference equals $\pm\pi/2$, at which point the average output is equal to $\pm V_0$. Now suppose the input frequency increases from ω_0, requiring a greater control voltage. If the frequency is high enough ($= \omega_x$) to dictate $V_{cont} = V_0$, then the PD must operate at the peak of its characteristic. However, the PD gain drops to zero here and the feedback loop fails. Thus, the circuit cannot lock if the input frequency reaches ω_X.

With the basic understanding of PLLs developed thus far, we now return to Eq. (16.2). The exact equality of the input and output frequencies of a PLL in the locked condition is a critical attribute. The significance of this property can be seen from two observations. First, in many applications, even a very small (deterministic) frequency error may prove unacceptable. For example, if a data stream is to be processed synchronously by a clocked system, even a slight difference between the data rate and the clock frequency results in a "drift," creating errors (Fig. 16.9). Second, the equality would *not* exist if the PLL compared the input and output *frequencies* rather than phases. As illustrated in Fig. 16.10(a), a loop employing a frequency detector (FD) would suffer from a finite difference between ω_{in} and ω_{out} due to various mismatches and other nonidealities. This can be understood by an analogy with the unity-gain feedback circuit of Fig. 16.10(b). Even if the op amp's open-loop gain is infinity, the input-referred offset voltage leads to a finite error between V_{in} and V_{out}.

Small Transients in Locked Condition Let us now analyze the response of a PLL in the locked condition to small phase or frequency transients at the input.

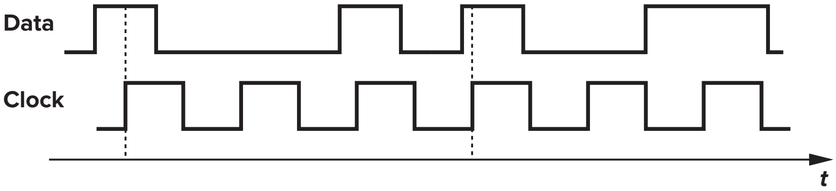


Figure 16.9 Drift of data with respect to clock in the presence of small frequency error.

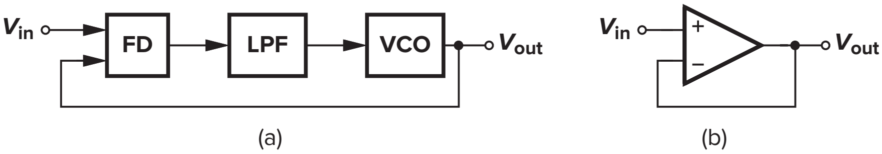


Figure 16.10 (a) Frequency-locked loop; (b) unity-gain feedback amplifier.

Consider a PLL in the locked condition and assume that the input and output waveforms can be expressed as

$$V_{in}(t) = V_A \cos \omega_1 t \tag{16.6}$$

$$V_{out}(t) = V_B \cos(\omega_1 t + \phi_0) \tag{16.7}$$

where higher harmonics are neglected and ϕ_0 is the static phase error. Suppose, as shown in Fig. 16.11, the input experiences a phase step of ϕ_1 at $t = t_1$, i.e., $\phi_{in} = \omega_1 t + \phi_1 u(t - t_1)$.[2] The phase step manifests itself as a rising edge in V_{in} that occurs earlier (or later) than the periodicity would dictate. Alternatively, we can say that the phase step results in a shorter (or longer) period just before t_1. Since the output of the LPF does not change instantaneously, the VCO initially continues to oscillate at ω_1. The growing phase difference between the input and the output then creates wide pulses at the output of the PD, forcing V_{LPF} to rise gradually. As a result, the VCO frequency begins to change, attempting to minimize the phase error. Note that the loop is not locked during the transient because the phase error varies with time.

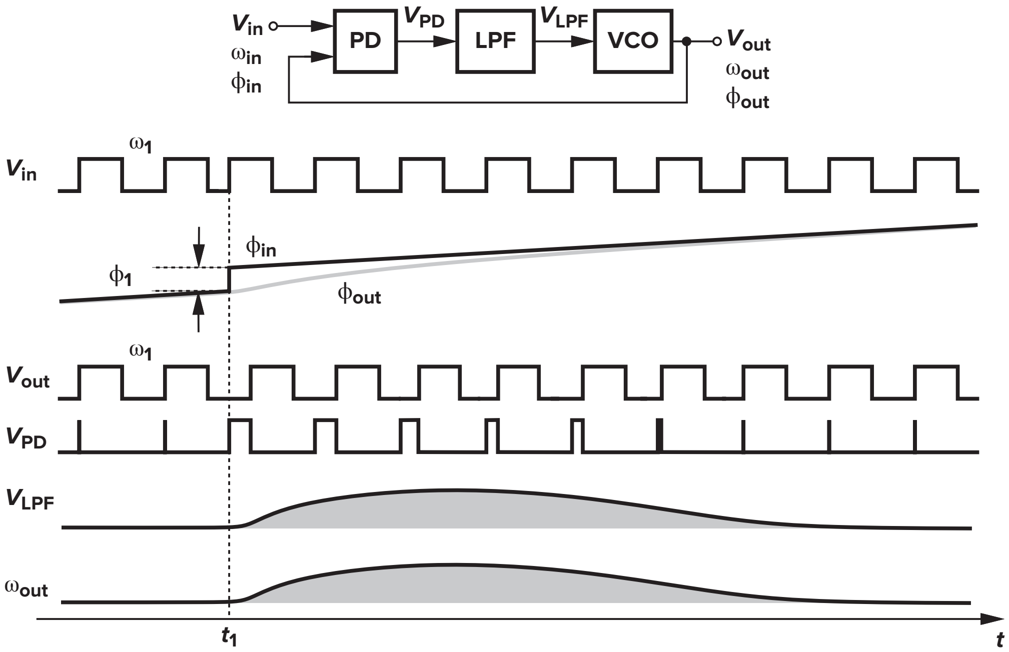


Figure 16.11 Response of a PLL to a phase step.

What happens after the VCO frequency begins to change? If the loop is to return to lock, ω_{out} must eventually go back to ω_1, requiring that V_{LPF} and hence $\phi_{out} - \phi_{in}$ also return to their original values. Since ϕ_{in} has changed by ϕ_1, the variation in the VCO frequency is such that the *area* under ω_{out} provides an additional phase of ϕ_1 in ϕ_{out}:

$$\int_{t1}^{\infty} \omega_{out} dt = \phi_1 \tag{16.8}$$

[2] In this example, ϕ_{in} and ϕ_{out} denote the *total* phases of the input and output, respectively.

Thus, when the loop settles, the output becomes equal to

$$V_{out}(t) = V_B \cos[\omega_1 t + \phi_0 + \phi_1 u(t - t_1)] \tag{16.9}$$

Consequently, as shown in Fig. 16.11, ϕ_{out} gradually "catches up" with ϕ_{in}.

It is important to make two observations. (1) After the loop returns to lock, *all* of the parameters (except for the total input and output phases) assume their original values. That is, $\phi_{in} - \phi_{out}$, V_{LPF}, and the VCO frequency remain unchanged—an expected result because these three parameters bear a one-to-one relationship and the input frequency has stayed the same. (2) The control voltage of the oscillator can serve as a suitable test point in the analysis of PLLs. While it is difficult to measure the time variations of phase and frequency in Fig. 16.11, $V_{cont}(= V_{LPF})$ can be readily monitored in simulations and measurements.

The reader may wonder whether an input phase step always gives rise to the response shown in Fig. 16.11. For example, is it possible for V_{LPF} to ring before settling to its final value? Such behavior is indeed possible and will be quantified in Sec. 16.1.3.

Let us now examine the response of PLLs to a small input frequency step $\Delta\omega$ at $t = t_1$ (Fig. 16.12). As with the case of a phase step, the VCO continues to oscillate at ω_1 immediately after t_1. Thus, the PD generates increasingly wider pulses, and V_{LPF} rises with time. As ω_{out} approaches $\omega_1 + \Delta\omega$, the width of the pulses generated by the PD decreases, eventually settling to a value that produces a dc component equal to $(\omega_1 + \Delta\omega - \omega_0)/K_{VCO}$. In contrast to the case of a phase step, the response of a PLL to a frequency step entails a permanent change in both the control voltage and the phase error. If the input frequency is varied slowly, ω_{out} simply "tracks" ω_{in}.

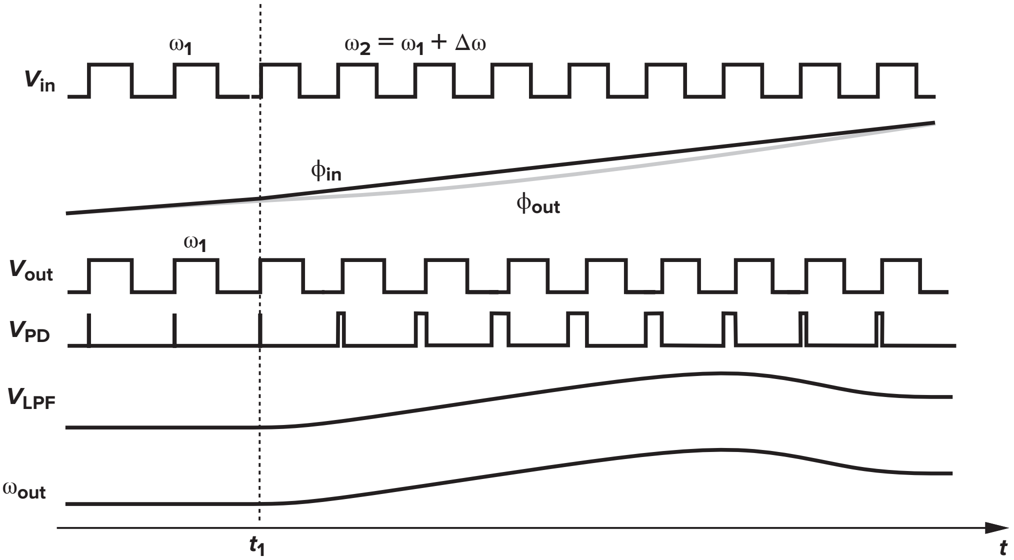


Figure 16.12 Response of a PLL to a small frequency step.

The exact settling behavior of PLLs depends on the various loop parameters and will be studied in Sec. 16.1.3. But, to arrive at an important observation, we consider the phase step response depicted in Fig. 16.13, where V_{cont} rings before settling to its final value. Consider the state of the loop at $t = t_2$. At this point, the output frequency is equal to its final value (because V_{cont} is equal to its final value), but the loop continues the transient because the phase error deviates from the required value. Similarly, at $t = t_3$, the phase error is equal to its final value, but the output frequency is not. In other words, for the loop to settle, both the phase and the frequency must settle to their proper values.

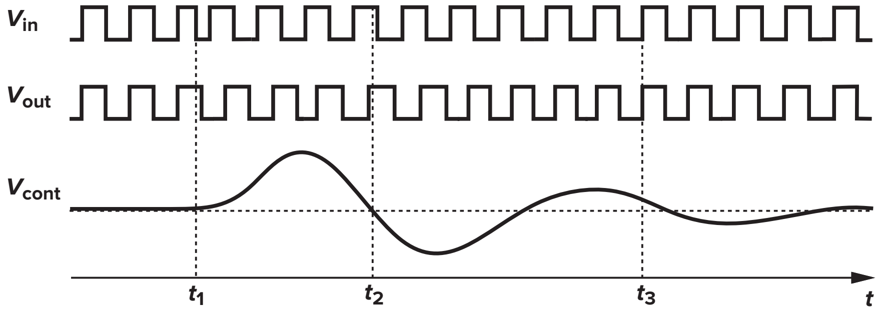


Figure 16.13 Example of phase step response.

▶ Example 16.4

In the PLL shown in Fig. 16.14, an external voltage V_{ex} is added to the output of the low-pass filter.[3] (a) Determine the phase error and V_{LPF} if the loop is locked and $V_{ex} = V_1$. (b) Suppose V_{ex} steps from V_1 to V_2 at $t = t_1$. How does the loop respond?

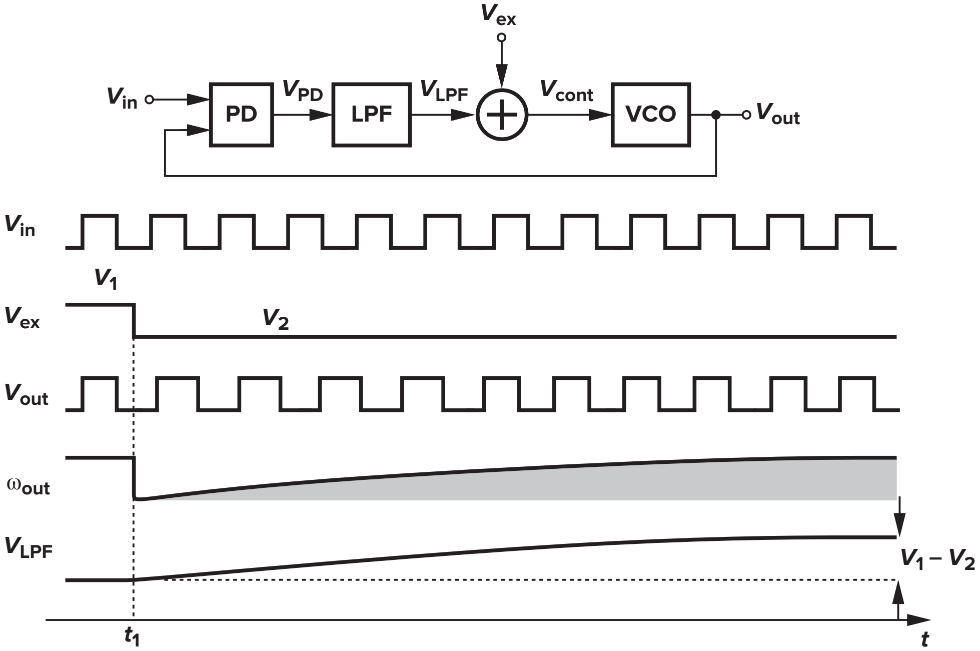


Figure 16.14

Solution

(a) If the loop is locked, $\omega_{out} = \omega_{in}$ and $V_{cont} = (\omega_{in} - \omega_0)/K_{VCO}$. Thus, $V_{LPF} = (\omega_{in} - \omega_0)/K_{VCO} - V_1$ and $\Delta\phi = V_{LPF}/K_{PD} = (\omega_{in} - \omega_0)/(K_{PD}K_{VCO}) - V_1/K_{PD}$.

(b) When V_{ex} steps from V_1 to V_2, V_{cont} immediately goes from $(\omega_{in} - \omega_0)/K_{VCO}$ to $(\omega_{in} - \omega_0)/K_{VCO} + (V_2 - V_1)$, changing the VCO frequency to $\omega_{in} - K_{VCO}(V_1 - V_2)$. Since V_{LPF} cannot change instantaneously, the PD begins to

[3] This topology is used for some types of frequency modulation in wireless communication.

generate increasingly wider pulses, raising V_{LPF} and increasing ω_{out}. When the loop returns to lock, ω_{out} becomes equal to ω_{in} and $V_{LPF} = (\omega_{in} - \omega_0)/K_{VCO} - V_2$. The phase error also changes to $(\omega_{in} - \omega_0)/(K_{PD}K_{VCO}) - V_2/K_{PD}$. Note that the area under ω_{out} during the transient is equal to the change in the output phase and hence the change in the phase error:

$$\int_{t1}^{\infty} \omega_{out} dt = \frac{V_1 - V_2}{K_{PD}} \tag{16.10}$$

From our study thus far, we conclude that phase-locked loops are "dynamic" systems, i.e., their response depends on the past values of the input and output. This is to be expected because the low-pass filter and the VCO introduce poles (and possibly zeros) in the loop transfer function. Moreover, we note that, so long as the input and the output remain perfectly periodic (i.e., $\phi_{in} = \omega_{in}t$ and $\phi_{out} = \omega_{in}t + \phi_0$), the loop operates in the steady state, exhibiting no transient. Thus, the PLL responds only to variations in the *excess* phase of the input or output. For example, in Fig. 16.11, $\phi_{in} = \omega_1 t + \phi_1 u(t - t_1)$, and in Fig. 16.12, $\phi_{in} = \omega_1 t + \Delta\omega \cdot t u(t - t_1)$.

16.1.3 Dynamics of Simple PLL

With the qualitative analysis of PLLs in the previous section, we can now study their transient behavior more rigorously. Assuming that the loop is initially locked, we treat the PLL as a feedback system but recognize that the output quantity in this analysis must be the (excess) phase of the VCO because the "error amplifier" can only compare phases. Our objective is to determine the transfer function $\Phi_{out}(s)/\Phi_{in}(s)$ for both open-loop and closed-loop systems and subsequently study the time-domain response. Note that the dimensions change from phase to voltage through the PD and from voltage to phase through the VCO.

What does $\Phi_{out}(s)/\Phi_{in}(s)$ signify? An analogy with more familiar transfer functions proves useful here. A circuit having a transfer function $V_{out}(s)/V_{in}(s) = 1/(1 + s/\omega_0)$ is considered a low-pass filter because if V_{in} varies rapidly, V_{out} cannot fully track the input variations. Similarly, $\Phi_{out}(s)/\Phi_{in}(s)$ reveals how the output phase tracks the input phase if the latter changes slowly or rapidly.

To visualize the variation of the excess phase with time, consider the waveforms in Fig. 16.15. The period varies slowly in Fig. 16.15(a) and rapidly in Fig. 16.15(b). Thus, $y_2(t)$ experiences faster phase variations than does $y_1(t)$.

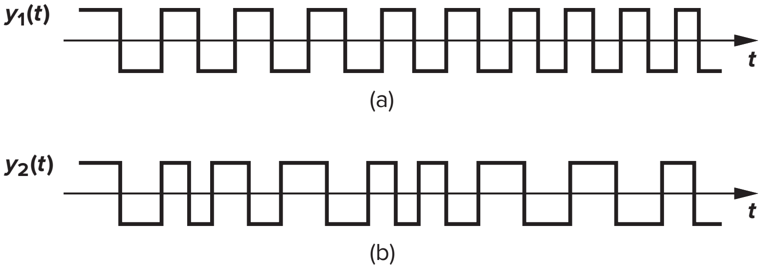


Figure 16.15 Slow and fast variation of the excess phase.

Let us construct a linear model of the PLL, assuming a first-order low-pass filter for simplicity. The PD output contains a dc component equal to $K_{PD}(\phi_{out} - \phi_{in})$ as well as high-frequency components. Since the latter are suppressed by the LPF, we simply model the PD by a subtractor whose output is "amplified" by K_{PD}. Illustrated in Fig. 16.16, the overall PLL model consists of the phase subtractor, the LPF transfer function $1/(1 + s/\omega_{LPF})$, where ω_{LPF} denotes the -3-dB bandwidth, and the VCO transfer function K_{VCO}/s (Chapter 15). Here, Φ_{in} and Φ_{out} denote the excess phases of the input and

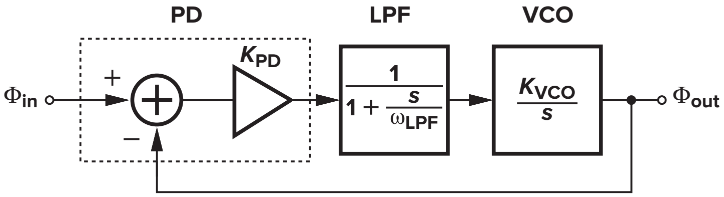


Figure 16.16 Linear model of type I PLL.

output waveforms, respectively. For example, if the total input phase experiences a step change, $\phi_1 u(t)$, then $\Phi_{in}(s) = \phi_1/s$.

The open-loop transfer function is given by

$$H(s)|_{\text{open}} = \frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} \tag{16.11}$$

$$= K_{PD} \cdot \frac{1}{1+\frac{s}{\omega_{LPF}}} \cdot \frac{K_{VCO}}{s} \tag{16.12}$$

revealing one pole at $s = -\omega_{LPF}$ and another at $s = 0$. Note that the loop gain is equal to $H(s)|_{\text{open}}$ because of the unity feedback factor. Since the loop gain contains a pole at the origin, the system is called "type I."

Before computing the closed-loop transfer function, let us make an important observation. What is the loop gain if s is very small, i.e., if the input excess phase varies very slowly? Owing to the pole at the origin, the loop gain goes to infinity as s approaches zero, a point of contrast to the feedback circuits studied in Chapters 8 and 10. Thus, the phase-locked loop (under closed-loop, locked condition) ensures that the change in ϕ_{out} is *exactly* equal to the change in ϕ_{in} as s goes to zero. This result predicts two interesting properties of PLLs. First, if the input excess phase varies very slowly, the output excess phase "tracks" it. (After all, ϕ_{out} is "locked" to ϕ_{in}.) Second, if the transients in ϕ_{in} have decayed (another case corresponding to $s \to 0$), then the change in ϕ_{out} is precisely equal to the change in ϕ_{in}. This is indeed true in the example depicted in Fig. 16.11.

From (16.12), we can write the closed-loop transfer function as

$$H(s)|_{\text{closed}} = \frac{K_{PD}K_{VCO}}{\frac{s^2}{\omega_{LPF}} + s + K_{PD}K_{VCO}} \tag{16.13}$$

For the sake of brevity, we hereafter denote $H(s)|_{\text{closed}}$ simply by $H(s)$ or Φ_{out}/Φ_{in}. As expected, if $s \to 0$, $H(s) \to 1$ because of the infinite loop gain.

In order to analyze $H(s)$ further, we derive a relationship that allows a more intuitive understanding of the system. Recall from Chapter 15 that the instantaneous frequency of a waveform is equal to the time derivative of the phase: $\omega = d\phi/dt$. Since the frequency and the phase are related by a linear operator, the transfer function of (16.13) applies to variations in the input and output frequencies as well:

$$\frac{\omega_{out}}{\omega_{in}}(s) = \frac{K_{PD}K_{VCO}}{\frac{s^2}{\omega_{LPF}} + s + K_{PD}K_{VCO}} \tag{16.14}$$

For example, this result predicts that if ω_{in} changes very slowly ($s \to 0$), then ω_{out} tracks ω_{in}, again an expected result because the loop is assumed locked. Equation (16.14) also indicates that if ω_{in} changes abruptly, but the system is given enough time to settle ($s \rightarrow 0$), then the change in ω_{out} equals that in ω_{in} (as illustrated in the example of Fig. 16.12).

The above observation aids the analysis in two directions. First, some transient responses of the closed-loop system may be simpler to visualize in terms of changes in the frequency quantities rather than the phase quantities. Second, since a change in ω_{out} must be accompanied by a change in V_{cont}, we have

$$H(s) = K_{VCO} \cdot \frac{V_{cont}}{\omega_{in}}(s) \tag{16.15}$$

That is, monitoring the response of V_{cont} to variations in ω_{in} indeed yields the response of the closed-loop system.

The second-order transfer function of (16.13) suggests that the step response of the type I system can be overdamped, critically damped, or underdamped. To derive the condition for each case, we rewrite the denominator in a familiar form used in control theory, $s^2 + 2\zeta\omega_n s + \omega_n^2$, where ζ is the "damping factor" and ω_n is the "natural frequency." That is

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{16.16}$$

where

$$\omega_n = \sqrt{\omega_{LPF} K_{PD} K_{VCO}} \tag{16.17}$$

$$\zeta = \frac{1}{2}\sqrt{\frac{\omega_{LPF}}{K_{PD} K_{VCO}}} \tag{16.18}$$

The two poles of the closed-loop system are given by

$$s_{1,2} = -\zeta\omega_n \pm \sqrt{(\zeta^2 - 1)\omega_n^2} \tag{16.19}$$

$$= (-\zeta \pm \sqrt{\zeta^2 - 1})\omega_n \tag{16.20}$$

Thus, if $\zeta > 1$, both poles are real, the system is overdamped, and the transient response contains two exponentials with time constants $1/s_1$ and $1/s_2$. On the other hand, if $\zeta < 1$, the poles are complex and the response to an input frequency step $\omega_{in} = \Delta\omega u(t)$ is equal to

$$\omega_{out}(t) = \left\{1 - e^{-\zeta\omega_n t}[\cos(\omega_n\sqrt{1-\zeta^2}t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\omega_n\sqrt{1-\zeta^2}t)]\right\}\Delta\omega u(t) \tag{16.21}$$

$$= [1 - \frac{1}{\sqrt{1-\zeta^2}}e^{-\zeta\omega_n t}\sin(\omega_n\sqrt{1-\zeta^2}t + \theta)]\Delta\omega u(t) \tag{16.22}$$

where ω_{out} denotes the change in the output frequency and $\theta = \sin^{-1}\sqrt{1-\zeta^2}$. Thus, as shown in Fig. 16.17, the step response contains a sinusoidal component with a frequency $\omega_n\sqrt{1-\zeta^2}$ that decays with a time constant $(\zeta\omega_n)^{-1}$. Note that the system exhibits the same response if a phase step is applied to the input and the output phase is observed.

The settling speed of PLLs is of great concern in most applications. Equation (16.22) indicates that the exponential decay determines how fast the output approaches its final value, implying that $\zeta\omega_n$ must be maximized. For the type I PLL under study here, (16.17) and (16.18) yield

$$\zeta\omega_n = \frac{1}{2}\omega_{LPF} \tag{16.23}$$

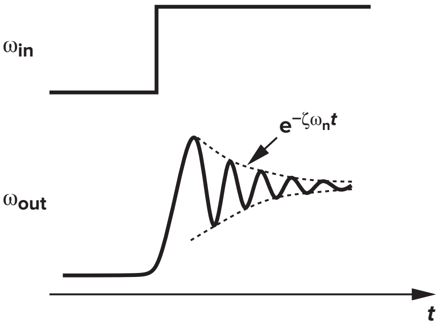


Figure 16.17 Underdamped response of PLL to a frequency step.

This result reveals a critical trade-off between the settling speed and the ripple on the VCO control line: the lower the ω_{LPF}, the greater the suppression of the high-frequency components produced by the PD, but the longer the settling time constant.

▶ **Example 16.5**

A cellular telephone incorporates a 900-MHz phase-locked loop to generate the carrier frequencies. If $\omega_{LPF} = 2\pi \times (20\text{ kHz})$ and the output frequency is to be changed from 901 MHz to 901.2 MHz, how long does the PLL output frequency take to settle within 100 Hz of its final value?

Solution

Since the step size is 200 kHz, we have

$$[1 - e^{-\zeta\omega_n t_s}\sin(\omega_n\sqrt{1-\zeta^2}t_s + \theta)] \times 200\text{ kHz} = 200\text{ kHz} - 100\text{ Hz} \tag{16.24}$$

Thus,

$$e^{-\zeta\omega_n t_s}\sin(\omega_n\sqrt{1-\zeta^2}t_s + \theta) = \frac{100\text{ Hz}}{200\text{ kHz}} \tag{16.25}$$

In the worst case, the sinusoid is equal to unity and

$$e^{-\zeta\omega_n t_s} = 0.0005 \tag{16.26}$$

That is

$$t_s = \frac{7.6}{\zeta\omega_n} \tag{16.27}$$

$$= \frac{15.2}{\omega_{LPF}} \tag{16.28}$$

$$= 0.12\text{ ms} \tag{16.29}$$

◀

In addition to the product $\zeta\omega_n$, the value of ζ itself is also important. Illustrated in Fig. 16.18 for several values of ζ and a constant ω_n, the step response exhibits severe ringing for $\zeta < 0.5$. In view of process and temperature variation of the loop parameters, ζ is usually chosen to be greater than $\sqrt{2}/2$ or even 1 to avoid excessive ringing.[4]

[4] A low ζ may also produce peaking in the transfer function. Thus, some applications require a ζ of 5 to 10 to avoid this effect.

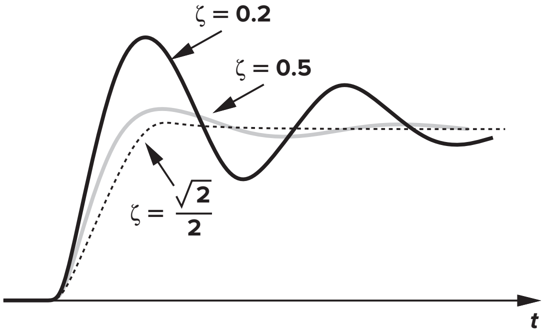


Figure 16.18 Underdamped response of a second-order system for various values of ζ.

The choice of ζ entails other trade-offs as well. First, (16.18) implies that as ω_{LPF} is reduced to minimize the ripple on the control voltage, the stability degrades. Second, (16.5) and (16.18) indicate that both the phase error and ζ are inversely proportional to $K_{PD}K_{VCO}$; lowering the phase error inevitably makes the system less stable. In summary, the type I PLL suffers from trade-offs among the settling speed, the ripple on the control voltage (i.e., the quality of the output signal), the phase error, and the stability.

The stability behavior of PLLs can also be analyzed graphically, providing more insight. Recall from Chapter 10 that the Bode plots of the magnitude and phase of the loop gain readily yield the phase margin. Let us utilize (16.12) to construct such plots. As shown in Fig. 16.19, the loop gain begins from infinity at $\omega = 0$ and falls at a rate of 20 dB/dec for $\omega < \omega_{LPF}$ and at a rate of 40 dB/dec thereafter. The phase begins at $-90°$ and asymptotically reaches $-180°$.

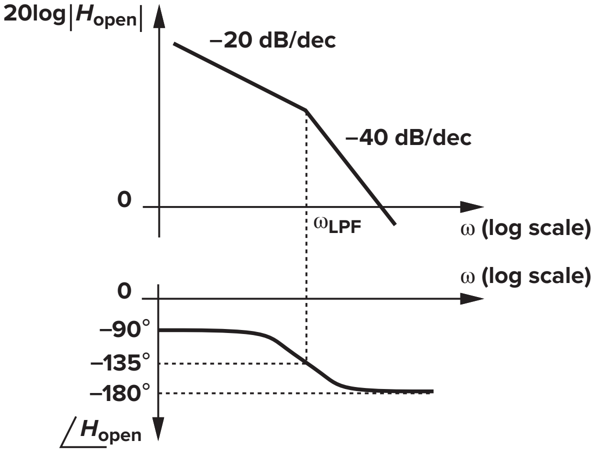


Figure 16.19 Bode plots of type I PLL.

What happens if a higher $K_{PD}K_{VCO}$ is chosen so as to minimize $\phi_{out} - \phi_{in}$? Since the entire gain plot in Fig. 16.19 is shifted up, the gain crossover moves to the right, thus degrading the phase margin. This is consistent with the dependence of ζ upon $K_{PD}K_{VCO}$.

As observed thus far, $K_{PD}K_{VCO}$ affects many important parameters of PLLs. This quantity is sometimes called the loop gain (even though it is not dimensionless) because of the resemblance of $\Delta\phi = (\omega_{out} - \omega_0)/(K_{PD}K_{VCO})$ to the error equation in a feedback system.

The stability behavior of type I PLLs can also be analyzed by the locus of their poles in the complex plane as the parameter $K_{PD}K_{VCO}$ varies (Fig. 16.20). With $K_{PD}K_{VCO} = 0$, the loop is open, $\zeta = \infty$, and the two poles are given by $s_1 = -\omega_{LPF}$ and $s_2 = 0$. As $K_{PD}K_{VCO}$ increases (i.e., the feedback becomes

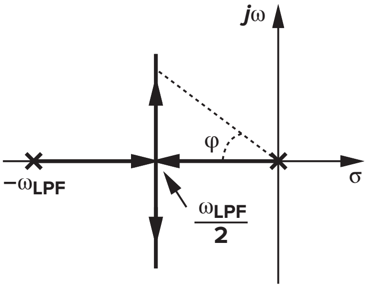


Figure 16.20 Root locus of type I PLL.

stronger), ζ drops and the two poles, given by $s_{1,2} = (-\zeta \pm \sqrt{\zeta^2 - 1})\omega_n$, move toward each other on the real axis. For $\zeta = 1$ (i.e., $K_{PD}K_{VCO} = \omega_{LPF}/4$), $s_1 = s_2 = -\zeta\omega_n = -\omega_{LPF}/2$. As $K_{PD}K_{VCO}$ increases further, the two poles become complex, with a real part equal to $-\zeta\omega_n = -\omega_{LPF}/2$, moving in parallel with the $j\omega$ axis.

We recognize from Fig. 16.20 that, as s_1 and s_2 move away from the real axis, the system becomes less stable. In fact, the reader can prove that $\cos\psi = \zeta$ (Problem 16.8), concluding that as ψ approaches 90°, ζ drops to zero.

Another transfer function that reveals the settling behavior of PLLs is that of the error at the output of the phase subtractor in Fig. 16.16. Defined as $H_e(s) = (\phi_{in} - \phi_{out})/\phi_{in}$, this transfer function can be obtained by noting that $\phi_{out}/\phi_{in} = H(s)$ and, from (16.13),

$$H_e(s) = 1 - H(s) \tag{16.30}$$

$$= \frac{s^2 + 2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n} \tag{16.31}$$

As expected, $H_e(s) \to 0$ if $s \to 0$ because the output tracks the input when the input varies very slowly or the transient has settled.

Example 16.6

Suppose a type I PLL experiences a frequency step $\Delta\omega$ at $t = 0$. Calculate the change in the phase error.

Solution

The Laplace transform of the frequency step equals $\Delta\omega/s$. Since $H_e(s)$ relates the phase error to the input phase, we write $\Phi_{in}(s) = (\Delta\omega/s)/s = \Delta\omega/s^2$. Thus, the Laplace transform of the phase error is

$$\Phi_e(s) = H_e(s) \cdot \frac{\Delta\omega}{s^2} \tag{16.32}$$

$$= \frac{s^2 + 2\zeta\omega_n s}{s^2 + 2\zeta\omega_n s + \omega_n^2} \cdot \frac{\Delta\omega}{s^2} \tag{16.33}$$

From the final value theorem,

$$\phi_e(t = \infty) = \lim_{s \to 0} s\Phi_e(s) \tag{16.34}$$

$$= \frac{2\zeta}{\omega_n}\Delta\omega \tag{16.35}$$

$$= \frac{\Delta\omega}{K_{PD}K_{VCO}} \tag{16.36}$$

which agrees with (16.5).

16.2 ■ Charge-Pump PLLs

While type I PLLs have been realized widely in discrete form, their shortcomings often prohibit usage in high-performance integrated circuits. In addition to the trade-offs among ζ, ω_{LPF}, and the phase error, type I PLLs suffer from another critical drawback: limited acquisition range.

16.2.1 Problem of Lock Acquisition

Suppose that when a PLL circuit is turned on, its oscillator operates at a frequency far from the input frequency, i.e., the loop is not locked. Under what conditions does the loop "acquire" lock? The transition of the loop from an unlocked to a locked condition is a very nonlinear phenomenon because the phase detector senses unequal frequencies. The problem of lock acquisition in type I PLLs has been studied extensively [1, 2], but we state without proof that the "acquisition range"[5] is on the order of ω_{LPF}; that is, the loop locks only if the difference between ω_{in} and ω_{out} is less than roughly ω_{LPF}.[6]

The problem of lock acquisition further tightens the trade-offs in type I PLLs. If ω_{LPF} is reduced to suppress the ripple on the control voltage, the acquisition range decreases. Note that even if the input frequency has a precisely-controlled value, a wide acquisition range is often necessary because the VCO center frequency may vary considerably with process and temperature. In most of today's applications, the acquisition range of the simple PLL studied thus far proves inadequate.

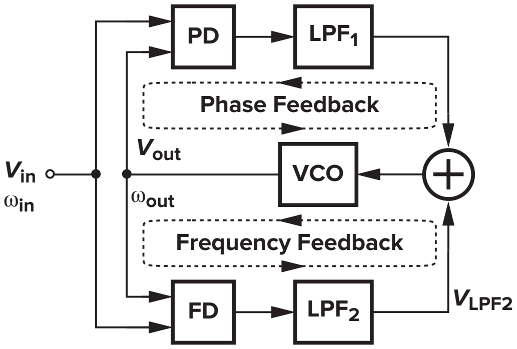


Figure 16.21 Addition of frequency detection to increase the acquisition range.

In order to remedy the acquisition problem, modern PLLs incorporate frequency detection in addition to phase detection. Called "aided acquisition" and illustrated in Fig. 16.21, the idea is to compare ω_{in} and ω_{out} by means of a frequency detector, generate a dc component V_{LPF2} proportional to $\omega_{in} - \omega_{out}$, and apply the result to the VCO in a negative-feedback loop. At the beginning, the FD drives ω_{out} toward ω_{in} while the PD output remains "quiet." When $|\omega_{out} - \omega_{in}|$ is sufficiently small, the phase-locked loop takes over, acquiring lock. Such a scheme increases the acquisition range to the tuning range of the VCO.[7]

[5] Acquisition range, tracking range, lock range, capture range, and pull-in range are often used to describe the behavior of PLLs in the presence of input or VCO frequency variation. For our purposes, the acquisition range, the capture range, and the pull-in range are the same. The tracking range refers to the input frequency range across which a locked PLL can track the input. With the addition of frequency detection, the acquisition range becomes equal to the tracking range (for periodic signals).

[6] This is a very rough estimate. In practice, the acquisition range may be several times narrower or wider. It is also assumed that the tuning range of the VCO is large enough not to limit the acquisition range.

[7] This may not be true if the input is not periodic.

16.2.2 Phase/Frequency Detector

For periodic signals, it is possible to merge the two loops of Fig. 16.21 by devising a circuit that can detect both phase and frequency differences. Called a phase/frequency detector (PFD) and illustrated conceptually in Fig. 16.22, the circuit employs sequential logic to create three states and respond to the rising (or falling) edges of the two inputs. If initially $Q_A = Q_B = 0$, then a rising transition on A leads to $Q_A = 1$, $Q_B = 0$. The circuit remains in this state until B goes high, at which point Q_A returns to zero. In other words, if a rising edge on A is followed by a rising edge on B, then Q_A goes high and returns to low. The behavior is similar for the B input.

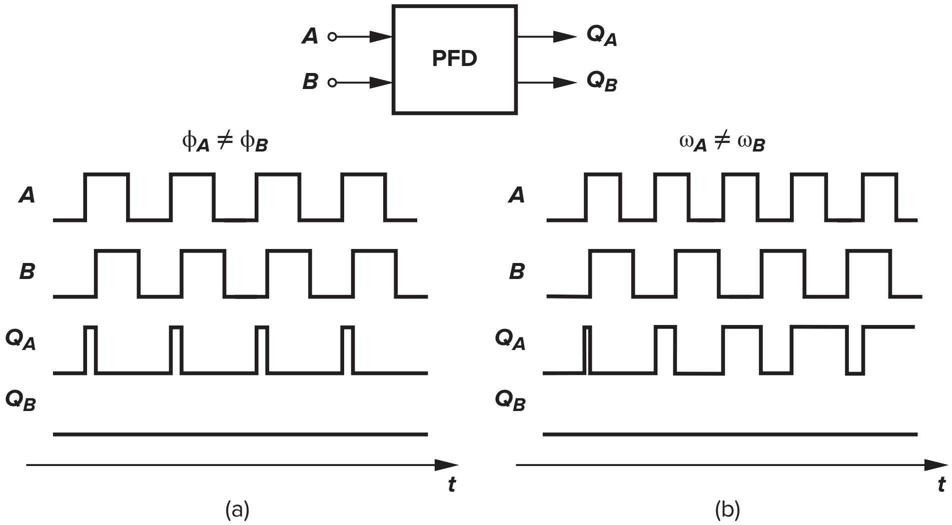


Figure 16.22 Conceptual operation of a PFD.

In Fig. 16.22(a), the two inputs have equal frequencies, but A leads B. The output Q_A continues to produce pulses whose width is proportional to $\phi_A - \phi_B$ while Q_B remains at zero. In Fig. 16.22(b), A has a higher frequency than B, and Q_A generates pulses while Q_B does not. By symmetry, if A lags B or has a lower frequency than B, then Q_B produces pulses and Q_A remains quiet. Thus, the dc contents of Q_A and Q_B provide information about $\phi_A - \phi_B$ or $\omega_A - \omega_B$. The outputs Q_A and Q_B are called the "UP" and "DOWN" pulses, respectively.

▶ **Example 16.7**

Explain whether a master-slave D flipflop can operate as a phase detector or a frequency detector. Assume that the flipflop provides differential outputs.

Solution

As shown in Fig. 16.23(a), we first apply inputs having equal frequencies and a finite phase difference, assuming that the output changes on the rising edge of the clock input. If A leads B, then V_{out} remains at a logical ONE indefinitely because the flipflop continues to sample the high levels of A. Conversely, if A lags B, then V_{out} remains low. Plotted in Fig. 16.23(b), the input-output characteristic of the circuit displays a very high gain at $\Delta\phi = 0, \pm\pi, \cdots$ and a zero gain at other values of $\Delta\phi$. The D flipflop is sometimes called a "bang-bang" phase detector to emphasize that the average value of V_{out} jumps from $-V_1$ to $+V_1$ as $\Delta\phi$ varies from slightly below zero to slightly above zero.

Now let us assume unequal frequencies for A and B. If the flipflop is to behave as a frequency detector, then the average value of V_{out} must exhibit different polarities for $\omega_A > \omega_B$ and $\omega_A < \omega_B$. However, as illustrated in Fig. 16.23(c), the average value is zero in both cases.

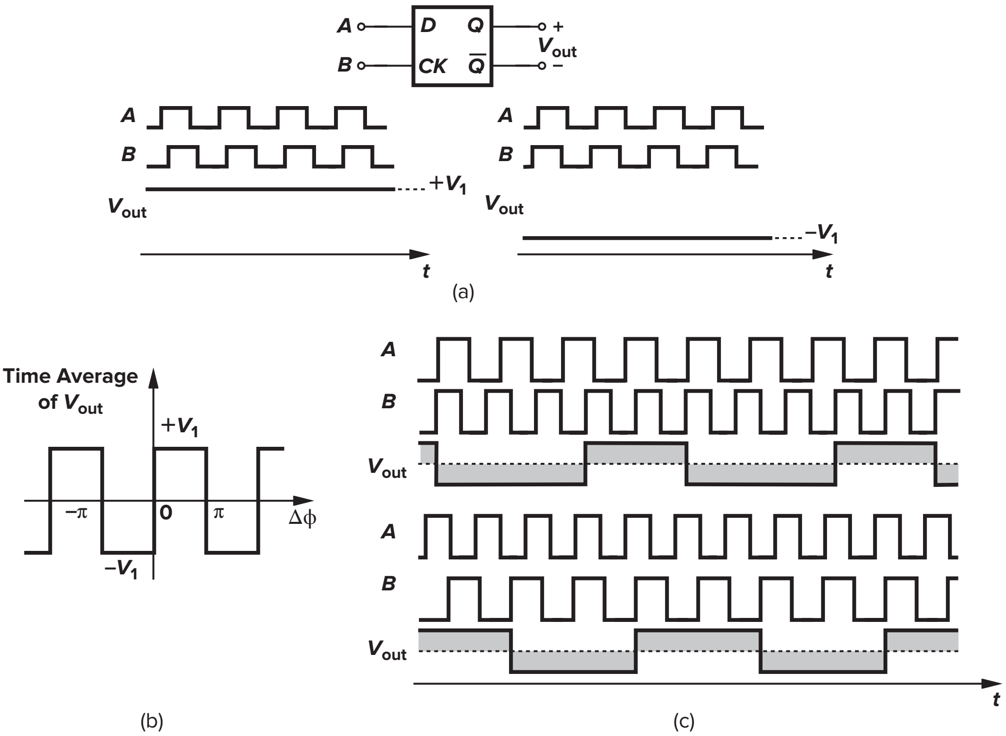


Figure 16.23 (a) D flipflop as a phase detector; (b) input-output characteristic; (c) response of D flipflop to unequal input frequencies.

The circuit of Fig. 16.22 can be realized in various forms. Figure 16.24(a) shows a simple implementation consisting of two edge-triggered, resettable D flipflops with their D inputs tied to a logical ONE.

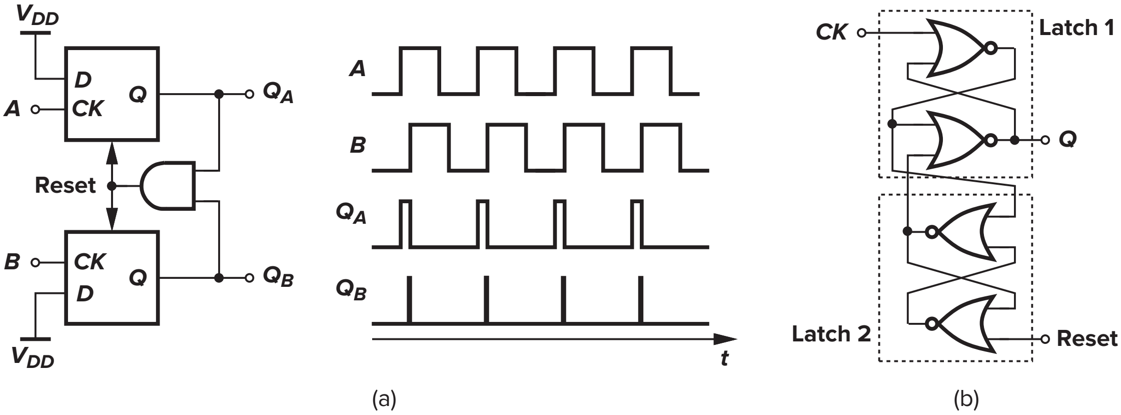


Figure 16.24 (a) Implementation of PFD; (b) implementation of D flipflop.

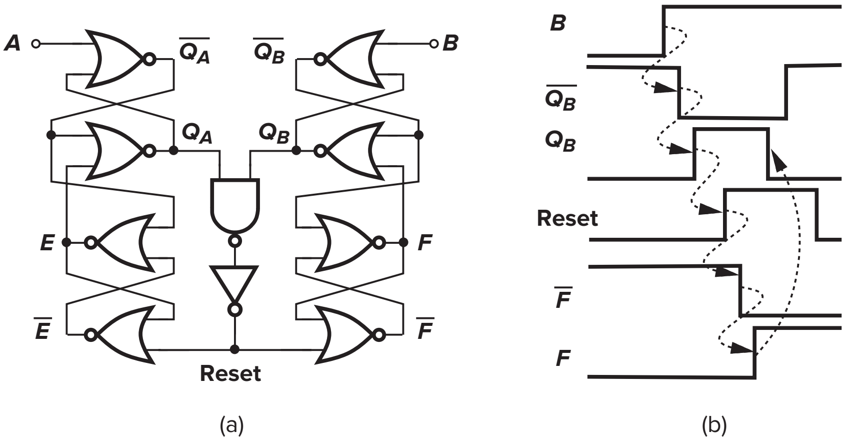
The inputs of interest, A and B, serve as the clocks of the flipflops. If $Q_A = Q_B = 0$ and A goes high, Q_A rises. If this event is followed by a rising transition on B, Q_B goes high and the AND gate resets both flipflops. In other words, Q_A and Q_B are simultaneously high for a short time, but the difference between their average values still represents the input phase or frequency difference correctly. Each flipflop can be implemented as shown in Fig. 16.24(b), where two RS latches are cross-coupled. Latch 1 and Latch 2 respond to the rising edges of CK and Reset, respectively.

▶ **Example 16.8**

Determine the width of the narrow reset pulses that appear in the Q_B waveform in Fig. 16.24(a).

Solution

Figure 16.25(a) illustrates the overall PFD at the gate level. If the circuit begins with $A = 1$, $Q_A = 1$, and $Q_B = 0$, a rising edge on B forces $\overline{Q_B}$ to go low and, one gate delay later, Q_B to go high. As shown in Fig. 16.25(b), this transition propagates to Reset, $\overline{E}$ and $\overline{F}$, E and F, and finally to Q_A and Q_B. Thus, the width of the pulse on Q_B is approximately equal to 5 gate delays.[8]



(a) (b)

Figure 16.25

◀

It is instructive to plot the input-output characteristic of the above PFD. Defining the output as the difference between the average values of Q_A and Q_B when $\omega_A = \omega_B$ and neglecting the effect of the narrow reset pulses, we note that the output varies symmetrically as $|\Delta\phi|$ begins from zero (Fig. 16.26). For $\Delta\phi = \pm 360^\circ$, V_{out} reaches its extrema and subsequently changes sign. The slope of the characteristic can be viewed as the gain.

How is the PFD of Fig. 16.24(a) utilized in a phase-locked loop? Since the difference between the average values of Q_A and Q_B is of interest, the two outputs can be low-pass filtered and sensed differentially (Fig. 16.27). A PLL employing such a topology always locks, but, due to the finite "loop gain," $K_{PFD}K_{VCO}$, it suffers from a finite phase error.

16.2.3 Charge Pump

In order to avoid the finite phase error present in type I PLLs, we wish to raise the loop gain to infinity, perhaps by means of an integrator. As our first step, we interpose a "charge pump" (CP) between the PFD

[8] This is a rough approximation because the NAND gate, the inverter, and the NOR gates have different delays and fanouts.

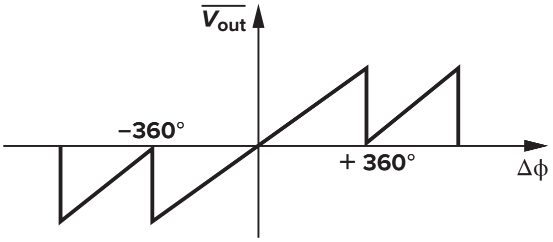


Figure 16.26 Input-output characteristic of the three-state PFD.

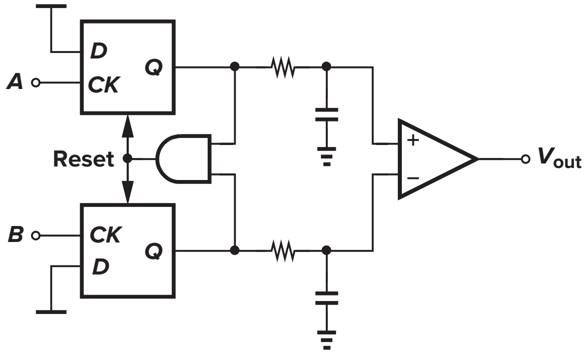


Figure 16.27 PFD followed by low-pass filters.

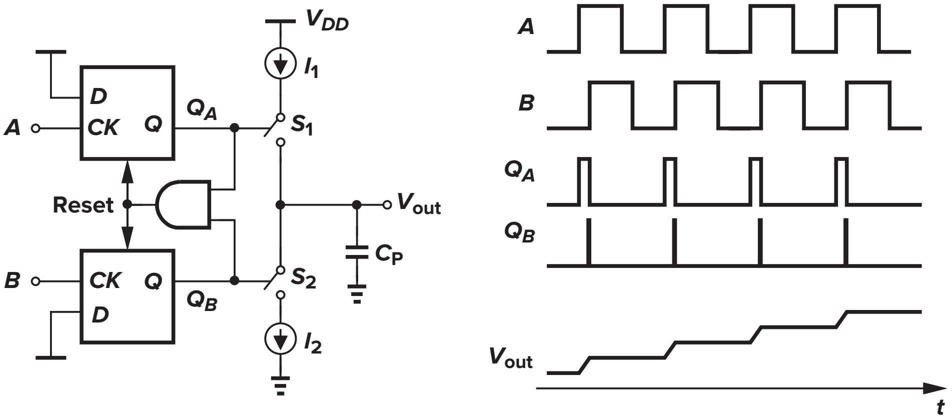


Figure 16.28 PFD with charge pump.

and the loop filter. A charge pump consists of two switched current sources that pump charge into or out of the loop filter according to two logical inputs. Figure 16.28 illustrates a charge pump driven by a PFD and driving a capacitor. The circuit has three states. If $Q_A = Q_B = 0$, then S_1 and S_2 are off and V_{out} remains constant. If Q_A is high and Q_B is low, then I_1 charges C_P. Conversely, if Q_A is low and Q_B is high, then I_2 discharges C_P. Thus, if, for example, A leads B, then Q_A continues to produce pulses and V_{out} rises steadily. Called UP and DOWN currents, respectively, I_1 and I_2 are nominally equal.

Example 16.9

What is the effect of the narrow pulses that appear in the Q_B waveform in Fig. 16.28?

Solution

Since Q_A and Q_B are simultaneously high for a finite period (approximately 5 gate delays from Example 16.8), the current supplied by the charge pump to C_P is affected. In fact, if $I_1 = I_2$, the current through S_1 simply flows through S_2 during the narrow reset pulse, leaving no current to charge C_P. As shown in Fig. 16.29, V_{out} remains constant after Q_B goes high.

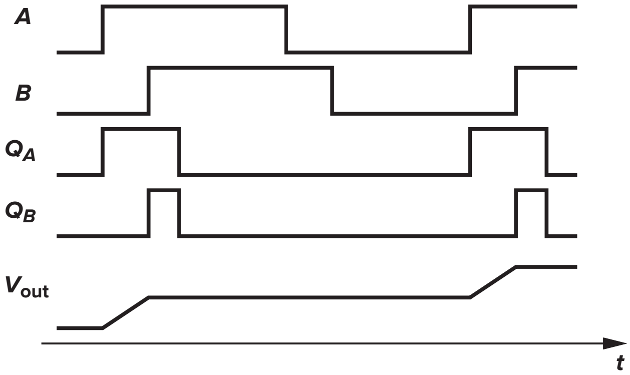


Figure 16.29

The PFD/CP/LPF cascade shown in Fig. 16.28 has an interesting property. If A, say, leads B by a finite amount, Q_A produces pulses indefinitely, allowing the charge pump to inject I_1 into C_P and forcing V_{out} to rise steadily. In other words, for a finite input error, the output eventually goes to $+\infty$ or $-\infty$, i.e., the "gain" of the circuit is infinity. In this cascade, the PFD converts the input phase error to a pulse width on Q_A or Q_B, the charge pump translates this pulse width to charge, and the capacitor accumulates this charge.

16.2.4 Basic Charge-Pump PLL

Let us now construct a PLL using the circuit of Fig. 16.28. Shown in Fig. 16.30 and called a charge-pump PLL, such an implementation senses the transitions at the input and output, detects phase or frequency differences, and activates the charge pump accordingly. When the loop is turned on, ω_{out} may be far from ω_{in}, and the PFD and the charge pump adjust the control voltage such that ω_{out} approaches ω_{in}. When the input and output frequencies are sufficiently close, the PFD operates as a phase detector, performing phase lock. The loop locks when the phase difference drops to zero and the charge pump remains relatively idle.

As observed above, the gain of the PFD/CP/LPF combination is infinite, i.e., a nonzero (deterministic) difference between ϕ_{in} and ϕ_{out} leads to indefinite charge buildup on C_P. What is the consequence of this attribute in a charge-pump PLL? When the loop of Fig. 16.30 is locked, V_{cont} is finite. Therefore, the input phase error must be exactly *zero*.[9] This is in contrast to the behavior of the type I PLL, in which the phase error is finite and a function of the output frequency.

[9] As explained in Sec. 16.3.1, mismatches still yield a finite phase error.

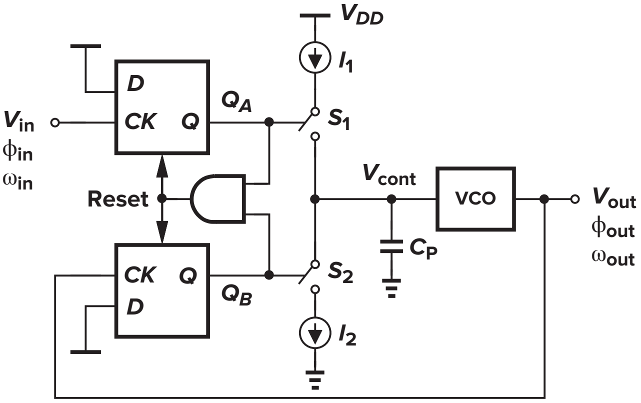


Figure 16.30 Simple charge-pump PLL.

To gain more insight into the operation of the PLL shown in Fig. 16.30, let us ignore the narrow reset pulses on Q_A and Q_B and assume that after $\phi_{out} - \phi_{in}$ drops to zero, the PFD simply produces $Q_A = Q_B = 0$. The charge pump thus remains idle, and C_P sustains a constant control voltage. Does this mean that the PFD and the CP are no longer needed?! If V_{cont} remains constant for a long time, the VCO frequency and phase begin to drift. In particular, the noise sources in the VCO create random variations in the oscillation frequency that can result in a large accumulation of phase error. The PFD then detects the phase difference, producing a corrective pulse on Q_A or Q_B that adjusts the VCO frequency through the charge pump and the filter. This is why we stated earlier that the PLL responds only to the *excess* phase of waveforms. We also note that, since in Fig. 16.30 phase comparison is performed in every cycle, the VCO phase and frequency cannot drift substantially.

Dynamics of CPPLL In order to quantify the behavior of charge-pump PLLs, we develop a linear model for the combination of the PFD, the charge pump, and the low-pass filter, thereby obtaining the transfer function. We raise two questions: (1) Is the PFD/CP/LPF combination in Fig. 16.28 a linear system? (2) If so, how can its transfer function be computed?

To answer the first question, we test the system for linearity. For example, as illustrated in Fig. 16.31(a), we double the input phase difference and see if V_{out} exactly doubles. Interestingly, the flat sections of V_{out} double, but not the ramp sections. After all, the current charging or discharging C_P is constant, yielding a constant slope for the ramp—an effect similar to slewing in op amps. Thus, the system is not linear in the strict sense. To overcome this quandary, we approximate the output waveform by a ramp [Fig. 16.31(b)], arriving at a linear relationship between V_{out} and $\Delta\phi$. In a sense, we approximate a discrete-time system by a continuous-time model.

To answer the second question, we recall that the transfer function is the Laplace transform of the impulse response, requiring that we apply a phase difference impulse and compute V_{out} in the time domain. Since a phase difference impulse is difficult to visualize, we apply a phase difference step, obtain V_{out}, and differentiate the result with respect to time.

Let us assume that the input period is T_{in} and the charge pump provides a current of $\pm I_P$ to the capacitor. As shown in Fig. 16.32, we begin with a zero phase difference and, at $t = 0$, step the phase of B by ϕ_0, i.e., $\Delta\phi = \phi_0 u(t)$. As a result, Q_A or Q_B continues to produce pulses that are $\phi_0 T_{in}/(2\pi)$ seconds wide, raising the output voltage by $(I_P/C_P)\phi_0 T_{in}/(2\pi)$ in every period.[10] Approximated by a ramp, V_{out} thus exhibits a slope of $(I_P/C_P)\phi_0/(2\pi)$ and can be expressed as

$$V_{out}(t) = \frac{I_P}{2\pi C_P} t \cdot \phi_0 u(t) \tag{16.37}$$

[10] We neglect the effect of the narrow reset pulses that appear in the other output.

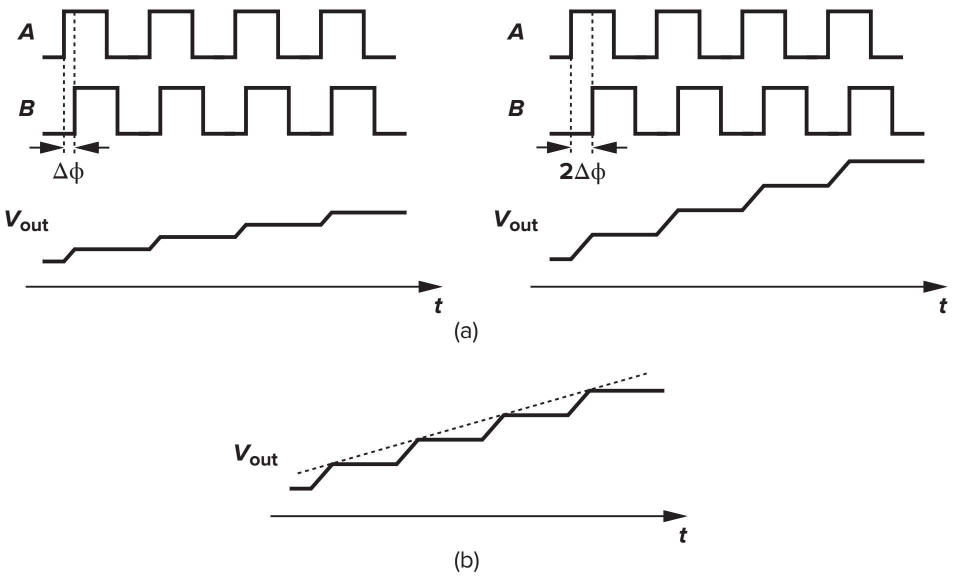


Figure 16.31 (a) Test of linearity of PFD/CP/LPF combination; (b) ramp approximation of the response.

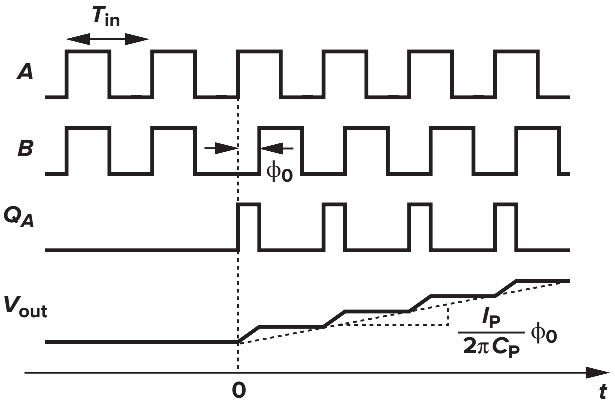


Figure 16.32 Step response of PFD/CP/LPF combination.

The impulse response is therefore given by

$$h(t) = \frac{I_P}{2\pi C_P} u(t) \tag{16.38}$$

yielding the transfer function

$$\frac{V_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi C_P} \cdot \frac{1}{s} \tag{16.39}$$

Consequently, the PFD/CP/LPF combination contains a pole at the origin, a point of contrast to the PD/LPF circuit used in the type I PLL. In analogy with the expression K_{VCO}/s, we call $I_P/(2\pi C_P)$ the "gain" of the PFD and denote it by K_{PFD}.

▶ **Example 16.10**

Suppose the output quantity of interest in the circuit of Fig. 16.28 is the current injected by the charge pump into the capacitor. Determine the transfer function from $\Delta\phi$ to this current, I_{out}.

Solution

Since $V_{out}(s) = I_{out}/(C_P s)$, we have

$$\frac{I_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi} \tag{16.40}$$

◀

Let us now construct a linear model of charge-pump PLLs. Shown in Fig. 16.33, the model gives an open-loop transfer function

$$\frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} = \frac{I_P}{2\pi C_P}\frac{K_{VCO}}{s^2} \tag{16.41}$$

Since the loop gain has two poles at the origin, this topology is called a "type II" PLL. The closed-loop transfer function, denoted by $H(s)$ for the sake of brevity, is thus equal to

$$H(s) = \frac{\dfrac{I_P K_{VCO}}{2\pi C_P}}{s^2 + \dfrac{I_P K_{VCO}}{2\pi C_P}} \tag{16.42}$$

This result is alarming because the closed-loop system contains two imaginary poles at $s_{1,2} = \pm j\sqrt{I_P K_{VCO}/(2\pi C_P)}$ and is therefore unstable. The instability arises because the loop gain has only two poles at the origin (i.e., two ideal integrators). As shown in Fig. 16.34(a), each integrator contributes a constant phase shift of 90°, allowing the system to oscillate at the gain crossover frequency.

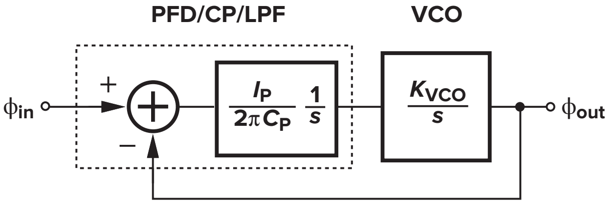


Figure 16.33 Linear model of simple charge-pump PLL.

In order to stabilize the system, we must modify the phase characteristic such that the phase shift is less than 180° at the gain crossover. As shown in Fig. 16.34(b), this is accomplished by introducing a zero in the loop gain, i.e., by adding a resistor in series with the loop filter capacitor (Fig. 16.35). Using the result of Example 16.10, the reader can prove (Problem 16.11) that the PFD/CP/LPF now has a transfer function

$$\frac{V_{out}}{\Delta\phi}(s) = \frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right) \tag{16.43}$$

It follows that the PLL open-loop transfer function is equal to

$$\frac{\Phi_{out}}{\Phi_{in}}(s)|_{\text{open}} = \frac{I_P}{2\pi}\left(R_P + \frac{1}{C_P s}\right)\frac{K_{VCO}}{s} \tag{16.44}$$

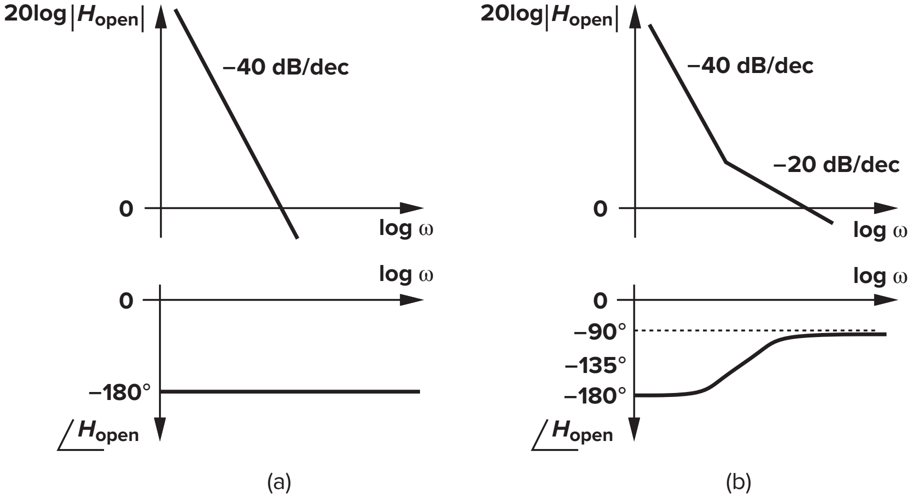


Figure 16.34 (a) Loop gain characteristics of simple charge-pump PLL; (b) addition of zero.

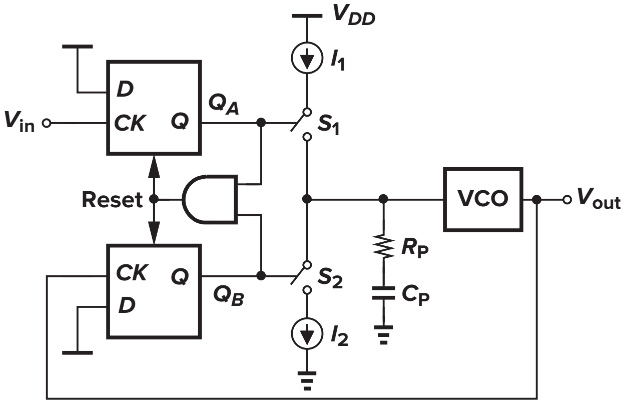


Figure 16.35 Addition of zero to charge-pump PLL.

and hence

$$H(s) = \frac{\dfrac{I_P K_{VCO}}{2\pi C_P}(R_P C_P s + 1)}{s^2 + \dfrac{I_P}{2\pi} K_{VCO} R_P s + \dfrac{I_P}{2\pi C_P} K_{VCO}} \tag{16.45}$$

The closed-loop system contains a zero at $s_z = -1/(R_P C_P)$. Using the same notation as that for the type I PLL, we have

$$\omega_n = \sqrt{\frac{I_P K_{VCO}}{2\pi C_P}} \tag{16.46}$$

$$\zeta = \frac{R_P}{2} \sqrt{\frac{I_P C_P K_{VCO}}{2\pi}} \tag{16.47}$$

As expected, if $R_P = 0$, then $\zeta = 0$. With complex poles, the decay time constant is given by $1/(\zeta \omega_n) = 4\pi/(R_P I_P K_{VCO})$.

Stability Issues The stability behavior of type II PLLs is quite different from that of type I PLLs. We begin the analysis with the Bode plots of the loop gain (the loop transmission) [Eq. (16.44)]. Shown in Fig. 16.36, these plots suggest that if $I_P K_{VCO}$ decreases, the gain crossover frequency moves toward the origin, *degrading* the phase margin. Predicted by (16.47), this trend is in sharp contrast to that expressed by (16.18) and illustrated in Fig. 16.19.

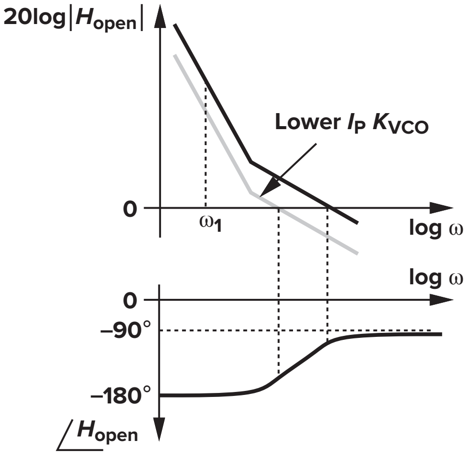


Figure 16.36 Stability degradation of charge-pump PLL as $I_P K_{VCO}$ decreases.

It is also possible to construct the root locus of the closed-loop system in the complex plane. For $I_P K_{VCO} = 0$ (e.g., $I_P = 0$), the loop is open and both poles lie at the origin. For $I_P K_{VCO} > 0$, we have $s_{1,2} = -\zeta\omega_n \pm \omega_n\sqrt{\zeta^2 - 1}$, and, since $\zeta \propto \sqrt{I_P K_{VCO}}$, the poles are complex if $I_P K_{VCO}$ is small. The reader can prove (Problem 16.14) that as $I_P K_{VCO}$ increases, s_1 and s_2 move on a circle centered at $\sigma = -1/(R_P C_P)$ with a radius $1/(R_P C_P)$ (Fig. 16.37). The poles return to the real axis at $\zeta = 1$, assuming a value of $-2/(R_P C_P)$. For $\zeta > 1$, the poles remain real, one approaching $-1/(R_P C_P)$ and the other going to $-\infty$ as $I_P K_{VCO} \to +\infty$. Since for complex s_1 and s_2, $\zeta = \cos\psi$, we observe that as $I_P K_{VCO}$ exceeds zero, the system becomes more stable.

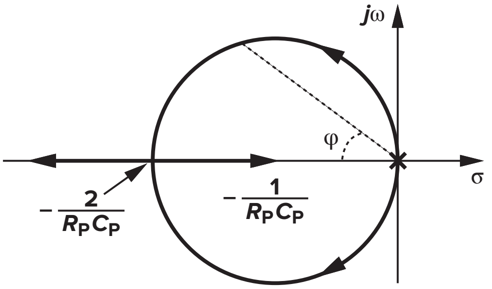


Figure 16.37 Root locus of type II PLL.

▶ **Example 16.11**

A student considers the Bode plots in Fig. 16.36 and observes that at ω_1, the loop gain exceeds unity and the phase shift is $-180°$. The student then reasons that the PLL must *oscillate* at this frequency! Explain the flaw in this reasoning.

Solution

The phase shift is in fact slightly less than zero unless $\omega_1 = 0$. As explained using Nyquist's approach in Chapter 10, a system containing two integrators and one zero does not oscillate.

◀

The compensated type II PLL of Fig. 16.35 suffers from a critical drawback. Since the charge pump drives the series combination of R_P and C_P, each time a current is injected into the loop filter, the control voltage experiences a large jump. Even in the locked condition, the mismatches between I_1 and I_2 and the charge injection and clock feedthrough of S_1 and S_2 introduce voltage jumps in V_{cont}. The resulting ripple severely disturbs the VCO, corrupting the output phase. To relax this issue, a second capacitor is usually added in parallel with R_P and C_P (Fig. 16.38), suppressing the initial step. The loop filter now is of *second* order, yielding a third-order PLL and creating stability difficulties [4]. Nonetheless, if C_2 is about one-fifth to one-tenth of C_P, the closed-loop time and frequency responses remain relatively unchanged.

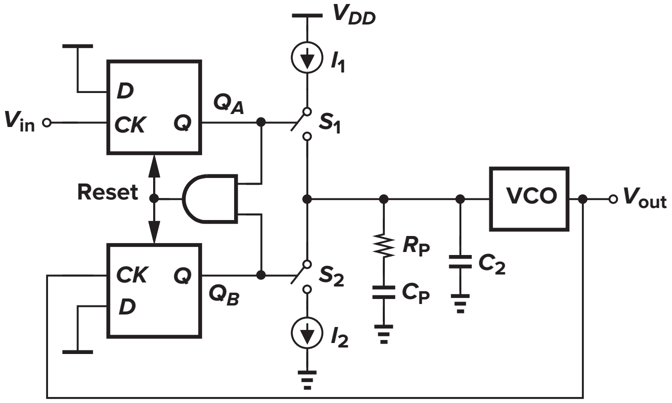


Figure 16.38 Addition of C_2 to reduce ripple on the control line.

Equation (16.47) implies that the loop becomes more stable as R_P increases. In reality, as R_P becomes very large, the stability degrades again. This effect is not predicted by the foregoing derivations because we have approximated the discrete-time system by a continuous-time loop. A more accurate analysis is given in [2], but simulations are often necessary to determine the stability bounds of CPPLLs.

16.3 ■ Nonideal Effects in PLLs

16.3.1 PFD/CP Nonidealities

Several imperfections in the PFD/CP circuit lead to high ripple on the control voltage even when the loop is locked. As mentioned earlier, the ripple modulates the VCO frequency, producing a waveform that is no longer periodic. In this section, we study these nonidealities.

The PFD implementation of Fig. 16.24(a) generates narrow, coincident pulses on both Q_A and Q_B even when the input phase difference is zero. As illustrated in Fig. 16.39, if A and B rise simultaneously, so do Q_A and Q_B, thereby activating the reset. That is, even when the PLL is locked, Q_A and Q_B simultaneously turn on the charge pump for a finite period $T_P \approx 5T_D$, where T_D denotes the gate delay (Example 16.8).

What are the consequences of the reset pulses on Q_A and Q_B? To understand why these pulses are *desirable*, we consider a hypothetical PFD that produces no pulses for a zero input phase difference [Fig. 16.40(a)]. How does such a PFD respond to a small phase error? As shown in Fig. 16.40(b), the circuit generates very narrow pulses on Q_A or Q_B. However, owing to the finite rise time and fall time resulting from the capacitance seen at these nodes, the pulse may not find enough time to reach a logical high level fail to turn on the charge pump switches. In other words, if the input phase difference, $\Delta\phi$, falls below a certain value ϕ_0, then the output voltage of the PFD/CP/LPF combination is no longer a

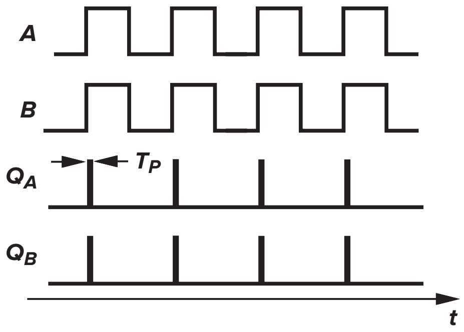


Figure 16.39 Coincident pulses generated by PFD with zero phase difference.

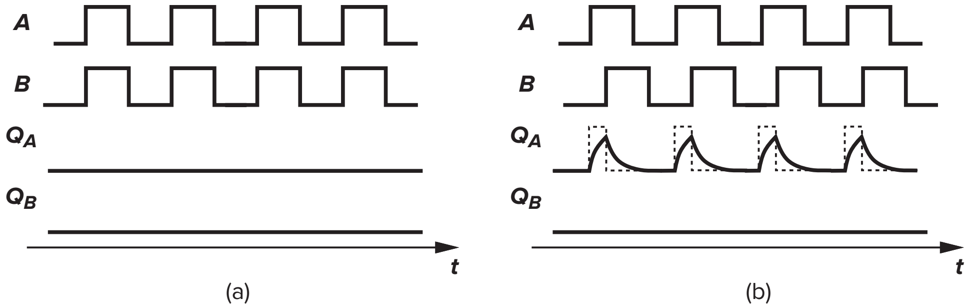


Figure 16.40 Output waveforms of a hypothetical PD with (a) zero input phase difference, and (b) a small input phase difference.

function of $\Delta\phi$. Since, as depicted in Fig. 16.41, for $|\Delta\phi| < \phi_0$ the charge pump injects no current, Eq. (16.41) implies that the loop gain drops to zero and the output phase is not locked. We say that the PFD/CP circuit suffers from a dead zone equal to $\pm\phi_0$ around $\Delta\phi = 0$.

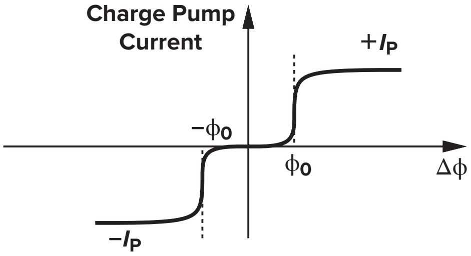


Figure 16.41 Dead zone in the charge-pump current.

The dead zone is highly undesirable because it allows the VCO to accumulate as much random phase error as ϕ_0 with respect to the input while receiving no corrective feedback. Thus, as illustrated in Fig. 16.42, the zero crossing points of the VCO output experience substantial random variations, an effect called "jitter."

Interestingly, the coincident pulses on Q_A and Q_B can eliminate the dead zone. This is because, for $\Delta\phi = 0$, the pulses always turn on the charge pump if they are sufficiently wide. Consequently, as shown in Fig. 16.43, an infinitesimal increment in the phase difference results in a proportional increase in the net current produced by the charge pump. In other words, the dead zone vanishes if T_P is long enough to allow Q_A and Q_B to reach a valid logical level and turn on the switches in the charge pump.

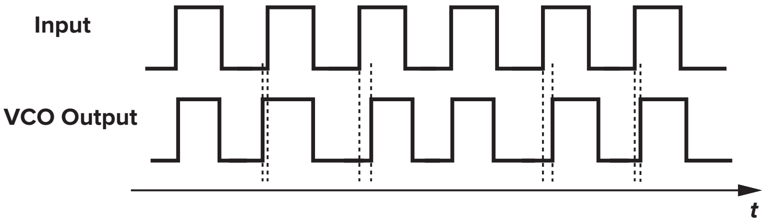


Figure 16.42 Jitter resulting from the dead zone.

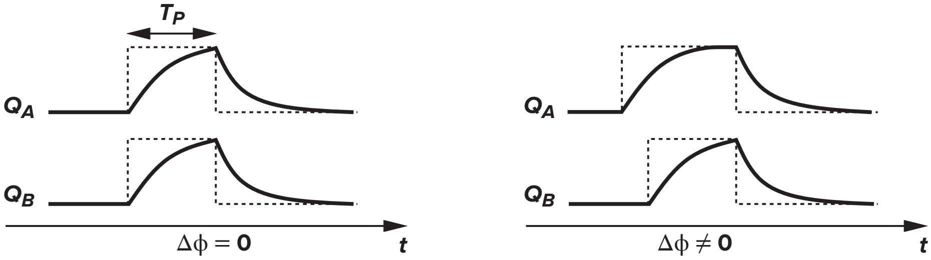


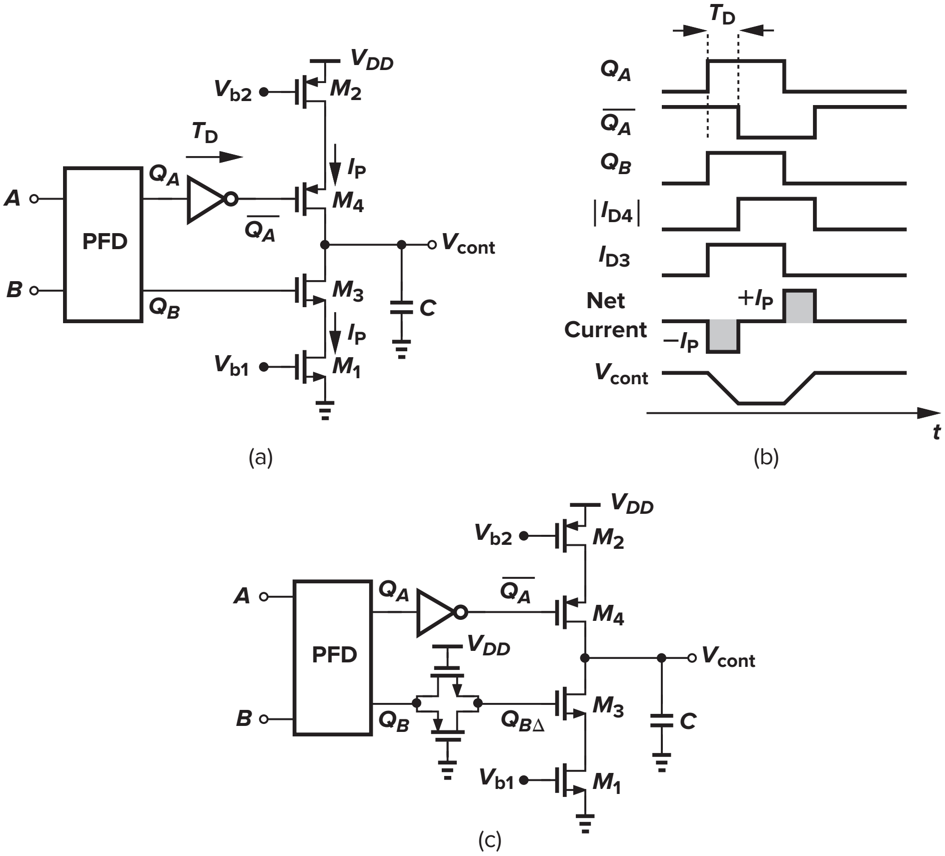
Figure 16.43 Response of actual PD to a small input phase difference.

While eliminating the dead zone, the reset pulses on Q_A and Q_B introduce other difficulties. Let us first implement the charge pump using MOS transistors [Fig. 16.44(a)]. Here, M_1 and M_2 operate as current sources and M_3 and M_4 as switches. The output Q_A is inverted so that when it goes high, M_4 turns on.

The first issue in the circuit of Fig. 16.44(a) stems from the delay difference between $\overline{Q_A}$ and Q_B in turning on their respective switches. As shown in Fig. 16.44(b), the net current injected by the charge pump into the loop filter jumps to $+I_P$ and $-I_P$, disturbing the oscillator control voltage periodically even if the loop is locked. To suppress this effect, a complementary pass gate can be interposed between Q_B and the gate of M_3, equalizing the delays [Fig. 16.44(c)].

The second issue in the CP of Fig. 16.44(c) relates to the mismatch between the drain currents of M_1 and M_2. As depicted in Fig. 16.45(a), even with perfect alignment of the UP and DOWN pulses, the net current produced by the charge pump is nonzero, changing V_{cont} by a constant increment at each phase comparison instant. How does the PLL respond to this error? For the loop to remain locked, the average value of the control voltage must remain constant. The PLL therefore creates a phase error between the input and the output such that the net current injected by the CP in every cycle is zero [Fig. 16.45(b)]. The relationship between the current mismatch and the phase error is determined in Problem 16.12. It is important to note that (1) the control voltage still experiences a periodic ripple; (2) owing to the low output impedance of short-channel MOSFETs, the current mismatch *varies* with the output voltage (i.e., with the VCO frequency); and (3) the clock feedthrough and charge injection mismatch between M_3 and M_4 further increase both the phase error and the ripple.

The third issue in the circuit of Fig. 16.44(c) originates from the finite capacitance seen at the drains of the current sources. Suppose, as illustrated in Fig. 16.46(a), S_1 and S_2 are off, allowing M_1 to discharge X to ground and M_2 to charge Y to V_{DD}. At the next phase comparison instant, both S_1 and S_2 turn on, V_X rises, V_Y falls, and $V_X \approx V_Y \approx V_{cont}$ if the voltage drop across S_1 and S_2 is neglected [Fig. 16.46(b)]. If the phase error is zero and $I_{D1} = |I_{D2}|$, does V_{cont} remain constant after the switches turn on? Even if $C_X = C_Y$, the change in V_X is not equal to that in V_Y. For example, if V_{cont} is relatively high, V_X changes by a large amount and V_Y by a small amount. The difference between the two changes must therefore be supplied by C_P, leading to a jump in V_{cont}.



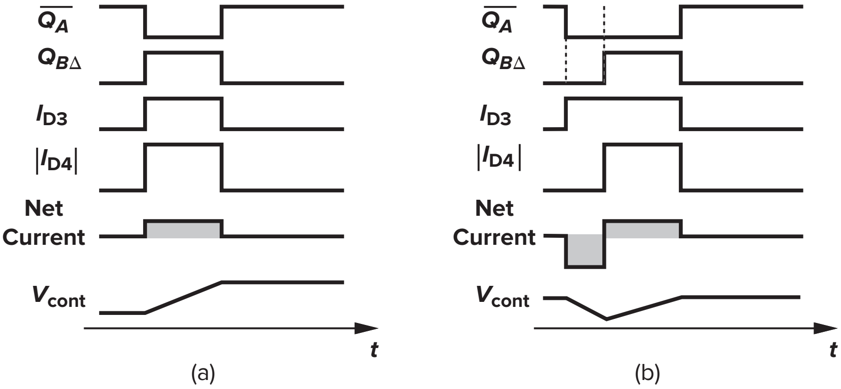
(a)

(b)

(c)

Figure 16.44 (a) Implementation of charge pump; (b) effect of skew between $\overline{Q_A}$ and Q_B; (c) suppression of skew by a pass gate.

The above charge-sharing phenomenon can be suppressed by "bootstrapping." Illustrated in Fig. 16.47 [3], the idea is to "pin" V_X and V_Y to V_{cont} after phase comparison is finished. When S_1 and S_2 turn off, S_3 and S_4 turn on, allowing the unity-gain amplifier to hold nodes X and Y at a potential equal to V_{cont}. Note that the amplifier need not provide much current because $I_1 \approx I_2$. At the next phase comparison instant, S_1 and S_2 turn on, S_3 and S_4 turn off, and V_X and V_Y begin with a value equal to V_{cont}. Thus, no charge sharing occurs between C_P and the capacitances at X and Y.



(a)

(b)

Figure 16.45 Effect of UP and DOWN current mismatch.

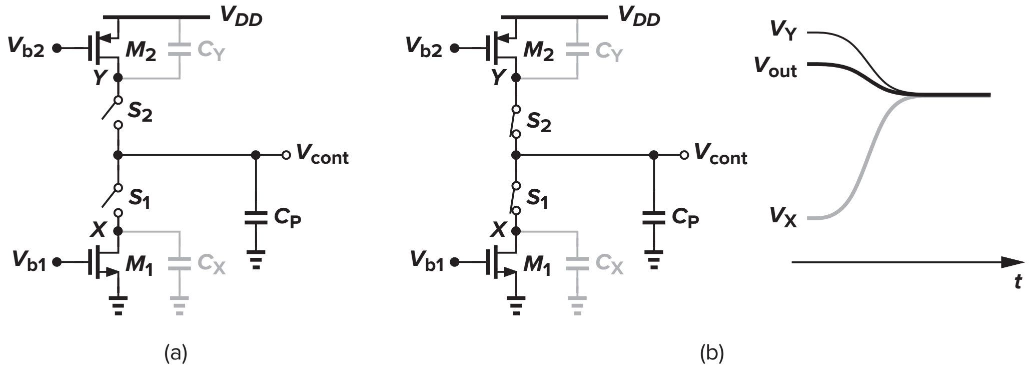


Figure 16.46 Charge sharing between C_P and capacitances at X and Y.

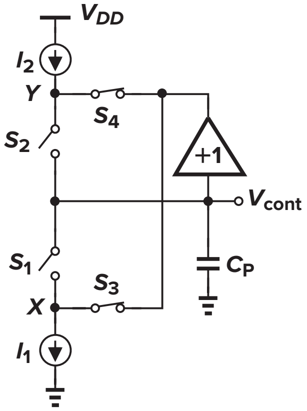


Figure 16.47 Bootstrapping X and Y to minimize charge sharing.

16.3.2 Jitter in PLLs

The response of phase-locked loops to jitter is of extreme importance in most applications. We first describe the concepts of jitter and the rate of change of jitter.

As shown in Fig. 16.48, a strictly periodic waveform, $x_1(t)$, contains zero crossings that are evenly spaced in time. Now consider the nearly periodic signal $x_2(t)$, whose period experiences small changes, displacing the zero crossings from their ideal points. We say that the latter waveform suffers from jitter.[11] Plotting the total phase, ϕ_{tot}, and the excess phase, ϕ_{ex}, of the two waveforms, we observe that jitter manifests itself as variation of the excess phase with time. In fact, ignoring the harmonics above the fundamental, we can write $x_1(t) = A\cos\omega t$ and $x_2(t) = A\cos[\omega t + \phi_n(t)]$, where $\phi_n(t)$ models the variation of the period.[12]

The rate at which the jitter varies is also important. Consider the two jittery waveforms depicted in Fig. 16.49. The first signal, $y_1(t)$, experiences "slow jitter" because its instantaneous frequency varies slowly from one period to the next. The second signal, $y_2(t)$, experiences "fast jitter." The rate of change is also evident from the excess phase plots of the two waveforms.

[11] Jitter is quantified by several different mathematical definitions, e.g., as in [5].

[12] The quantity $\phi_n(t)$ (or more commonly its spectrum) is called the "phase noise." In this book, we assume that the jitter is uniquely represented by $\phi_n(t)$.

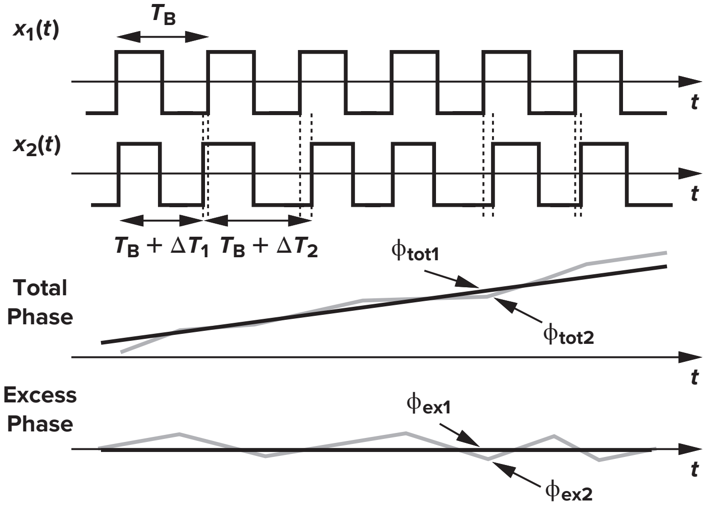


Figure 16.48 Ideal and jittery waveforms.

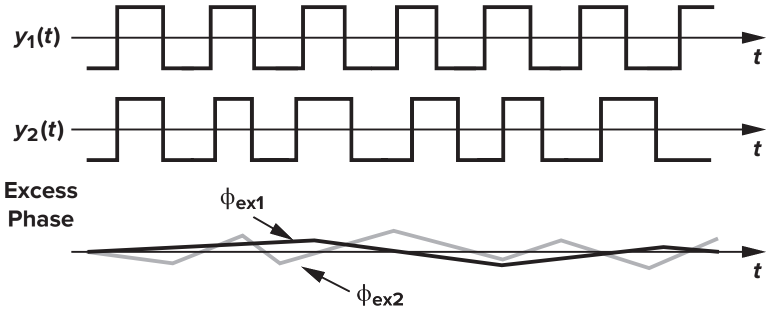


Figure 16.49 Illustration of slow and fast jitter.

Two jitter phenomena in phase-locked loops are of great interest: (1) the input exhibits jitter, and (2) the VCO produces jitter. Let us study each case, assuming that the input and output waveforms are expressed as $x_{in}(t) = A\cos[\omega t + \phi_{in}(t)]$ and $x_{out}(t) = A\cos[\omega t + \phi_{out}(t)]$.

The transfer functions derived for type I and type II PLLs have a low-pass characteristic, suggesting that if $\phi_{in}(t)$ varies rapidly, then $\phi_{out}(t)$ does not fully track the variations. In other words, slow jitter at the input propagates to the output unattenuated, but fast jitter does not. We say the PLL low-pass filters $\phi_{in}(t)$.

Now suppose the input is strictly periodic, but the VCO suffers from jitter. Viewing jitter as random phase variations, we construct the model depicted in Fig. 16.50, where the input excess phase is set to zero

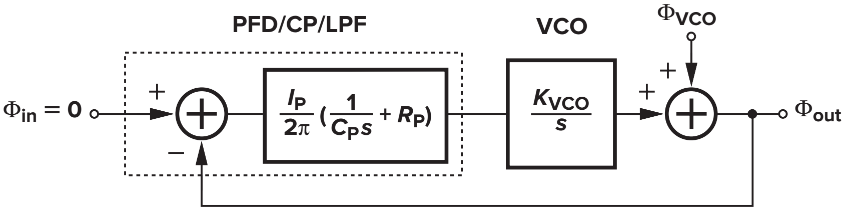


Figure 16.50 Effect of VCO jitter.

[i.e., $x_{in}(t) = A\cos\omega t$] and a random component Φ_{VCO} is added to the output of the VCO to represent its jitter. The reader can show that the transfer function from Φ_{VCO} to Φ_{out} for a type II PLL is equal to

$$\frac{\Phi_{out}}{\Phi_{VCO}}(s) = \frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{16.48}$$

Interestingly, the characteristic has a high-pass nature, indicating that slow jitter components generated by the VCO are suppressed, but fast jitter components are not. This can be understood with the aid of Fig. 16.50: if $\phi_{VCO}(t)$ changes slowly (e.g., the oscillation period drifts with temperature), then the comparison with $\phi_{in} = 0$ (i.e., a perfectly periodic signal) generates a slowly-varying error that propagates through the LPF and adjusts the VCO frequency, thereby counteracting the change in ϕ_{VCO}. On the other hand, if ϕ_{VCO} varies rapidly (e.g., high-frequency noise modulates the oscillation period), then the error produced by the phase detector is heavily attenuated by the poles in the loop, failing to correct for the change.

Figure 16.51 conceptually summarizes the response of PLLs to input jitter and VCO jitter. Depending on the application and the environment, one or both sources may be significant, requiring an optimum choice of the loop bandwidth.

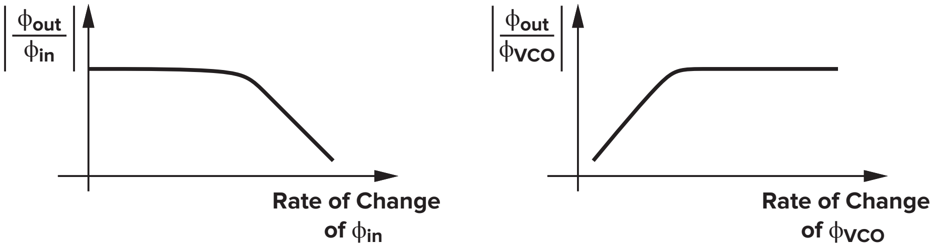


Figure 16.51 Transfer functions of jitter from input and VCO to the output.

16.4 ■ Delay-Locked Loops

A variant of PLLs that finds usage in many applications is the "delay-locked loop." To arrive at the concept, let us begin with an example. Suppose an application requires four clock phases with a precise spacing of $\Delta T = 1$ ns between consecutive edges [Fig. 16.52(a)]. How should these phases be generated? We can use a two-stage differential ring oscillator[13] to produce the four phases, but how do we guarantee that $\Delta T = 1$ ns despite process and temperature variations? This requires that the oscillator be locked to a 250-MHz reference so that the output period is exactly equal to 4 ns [Fig. 16.52(b)].

An alternative approach to generating the clock phases of Fig. 16.52(a) is to apply the input clock to four delay stages in a cascade. Illustrated in Fig. 16.53(a), this technique nonetheless does not produce a well-defined edge spacing because the delay of each stage varies with process and temperature. Now consider the circuit shown in Fig. 16.53(b), where the phase difference between CK_{in} and CK_4 is sensed by a phase detector, a proportional average voltage, V_{cont}, is generated, and the delay of the stages is adjusted with negative feedback. For a large loop gain, the phase difference between CK_{in} and CK_4 is small; that is, the four stages delay the clock by almost exactly one period, thereby establishing precise edge spacing.[14] This topology is called a delay-locked loop to emphasize that it incorporates a voltage-controlled delay line (VCDL) rather than a VCO. In practice, a charge pump is interposed between the PD

[13] As explained in Chapter 15, a simple two-stage CMOS ring oscillator may not oscillate. This example is merely for illustration purposes.

[14] The total delay through the four stages may be equal to two or more periods. We return to this issue later.

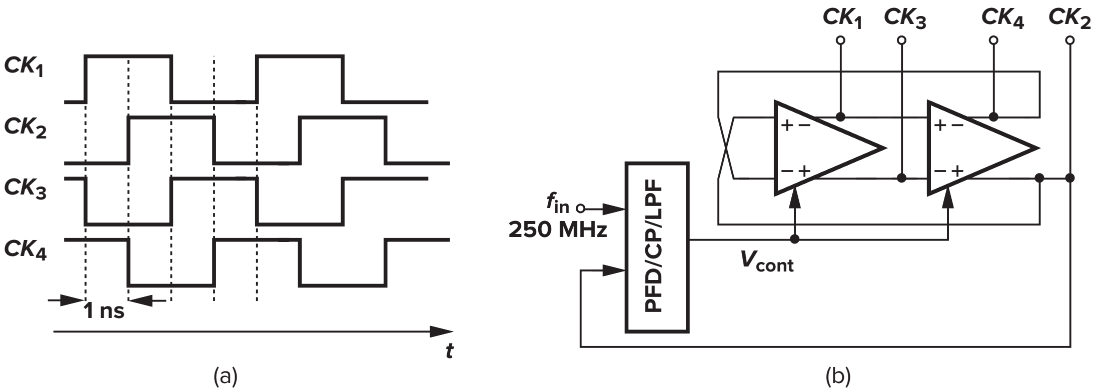


Figure 16.52 (a) Clock phases with edge-to-edge delay of 1 ns; (b) use of a phase-locked ring oscillator to generate the clock phases.

and the LPF to achieve an infinite loop gain. Each delay stage may be based on one of the ring oscillator stages described in Chapter 15.

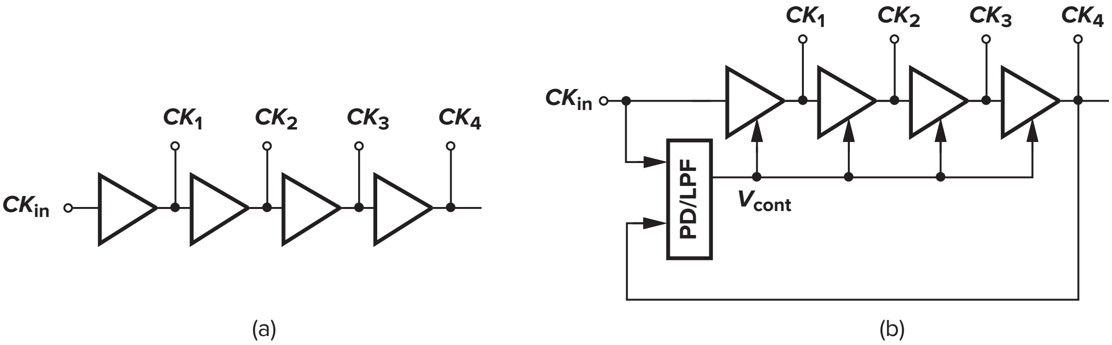


Figure 16.53 (a) Generation of clock edges by delay stages; (b) simple delay-locked loop.

The reader may wonder about the advantages of DLLs over PLLs. First, delay lines are generally less susceptible to noise than are oscillators because corrupted zero crossings of a waveform disappear at the end of a delay line, whereas they are recirculated in an oscillator, thereby experiencing more corruption. Second, in the VCDL of Fig. 16.53(b), a change in the control voltage immediately changes the delay; that is, the transfer function $\Phi_{out}(s)/V_{cont}(s)$ is simply equal to the gain of the VCDL, K_{VCDL}. Thus, the feedback system of Fig. 16.53(b) has the same order as the LPF, and its stability and settling issues are more relaxed than those of a PLL.

▶ Example 16.12

Explain qualitatively what type of transfer function the DLL of Fig. 16.54 has.

Solution

Suppose the input exhibits slow phase fluctuations. Then, the phase error sees a high gain through the PD/CP/LPF combination, and the delay of the line is adjusted so as to minimize this error. That is, ϕ_{out} tracks ϕ_{in}, and the gain is about unity. Now, suppose the input exhibits very fast phase changes. The feedback loop thus has little gain, providing little correction at the control of the delay line; i.e., V_{cont} remains relatively constant. As a result, the input phase variations directly propagate to the output, yielding a gain of about unity. We conclude that the DLL exhibits an all-pass response, but that for moderately fast phase fluctuations, the response may have a dip or a peak.

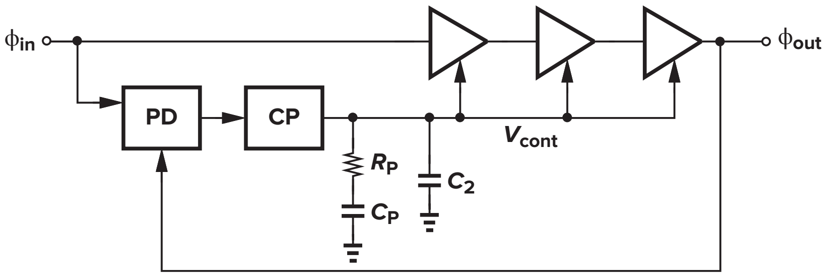


Figure 16.54

The principal drawback of DLLs is that they cannot generate a variable output frequency. This issue becomes clearer when we study the frequency synthesis capabilities of PLLs in Sec. 16.5.1. DLLs may also suffer from locked delay ambiguity. That is, if the total delay of the four stages in Fig. 16.53(b) can vary from below T_{in} to above $2T_{in}$, then the loop may lock with a CK_{in}-to-CK_4 delay equal to either T_{in} or $2T_{in}$. This ambiguity proves detrimental if the DLL must provide precisely-spaced clock edges because the edge-to-edge delay may settle to $2T_{in}/4$ rather than $T_{in}/4$. In such cases, additional circuitry is necessary to avoid the ambiguity. Also, mismatches between the delay stages and their load capacitances introduce error in the edge spacing, requiring large devices and careful layout.

16.5 ■ Applications

After nearly 90 years since its invention, phase locking continues to find new applications in electronics, communication, and instrumentation. Examples include memories, microprocessors, hard disk drive electronics, RF and wireless transceivers, and optical fiber receivers.

The reader may recall from Sec. 16.1.2 that a PLL appears no more useful than a short piece of wire because both guarantee a small phase difference between the input and the output. In this section, we present a number of applications that demonstrate the versatility of phase locking. The concepts described below have been the topic of numerous books and papers, e.g., [6, 7].

16.5.1 Frequency Multiplication and Synthesis

Frequency Multiplication A PLL can be modified such that it multiplies its input frequency by a factor of M. To arrive at the implementation, we exploit an analogy with voltage multiplication. As depicted in Fig. 16.55(a), a feedback system amplifies the input voltage by a factor of M if the output voltage is divided by M [i.e., if $R_2/(R_1 + R_2) = 1/M$] and the result is compared with the input. Thus, as shown in Fig. 16.55(b), if the output *frequency* of a PLL is divided by M and applied to the phase detector, we have $f_{out} = Mf_{in}$. From another point of view, since $f_D = f_{out}/M$ and f_D and f_{in} must be equal in the

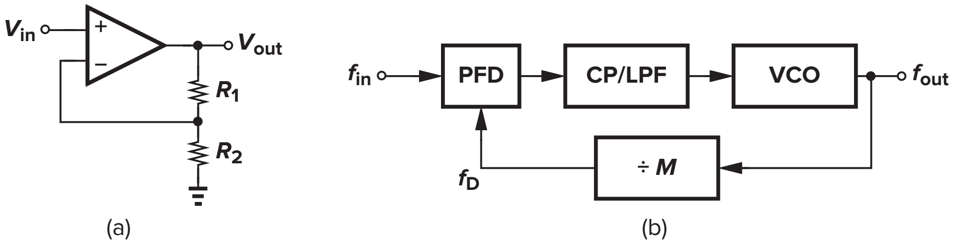


Figure 16.55 (a) Voltage amplification and (b) frequency multiplication.

locked condition, the PLL multiplies f_{in} by M. The $\div M$ circuit is realized as a counter that produces one output pulse for every M input pulses.

As with voltage division in Fig. 16.55(a), the feedback divider in the loop of Fig. 16.55(b) alters the system characteristics. Using (16.44), we rewrite (16.45) as

$$H(s) = \frac{\dfrac{I_P}{2\pi}\left(R_P + \dfrac{1}{C_P s}\right)\dfrac{K_{VCO}}{s}}{1 + \dfrac{1}{M}\dfrac{I_P}{2\pi}\left(R_P + \dfrac{1}{C_P s}\right)\dfrac{K_{VCO}}{s}} \tag{16.49}$$

$$= \frac{\dfrac{I_P K_{VCO}}{2\pi C_P}(R_P C_P s + 1)}{s^2 + \dfrac{I_P}{2\pi}\dfrac{K_{VCO}}{M}R_P s + \dfrac{I_P}{2\pi C_P}\dfrac{K_{VCO}}{M}} \tag{16.50}$$

Note that $H(s) \to M$ as $s \to 0$, i.e., phase or frequency changes at the input result in an M-fold change in the corresponding output quantity. Comparing the denominators of (16.45) and (16.50), we observe that frequency division in the loop manifests itself as division of K_{VCO} by M. In other words, as far as the poles of the closed-loop system are concerned, we can assume that the oscillator and the divider form a VCO with an equivalent gain of K_{VCO}/M. This is, of course, to be expected because, for the VCO/divider cascade shown in Fig. 16.56, we have

$$\omega_{out} = \frac{\omega_0 + K_{VCO}V_{cont}}{M} \tag{16.51}$$

$$= \frac{\omega_0}{M} + \frac{K_{VCO}}{M}V_{cont} \tag{16.52}$$

Thus, the combination cannot be distinguished from a VCO having an intercept frequency of ω_0/M and a gain of K_{VCO}/M.

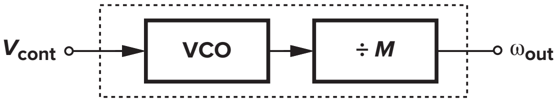


Figure 16.56 Equivalency of VCO/divider combination to a single VCO.

The foregoing discussion suggests that (16.46) and (16.47) can be respectively rewritten as

$$\omega_n = \sqrt{\frac{I_P}{2\pi C_P}\frac{K_{VCO}}{M}} \tag{16.53}$$

$$\zeta = \frac{R_P}{2}\sqrt{\frac{I_P C_P}{2\pi}\frac{K_{VCO}}{M}} \tag{16.54}$$

Also, the decay time constant is modified to $(\zeta\omega_n)^{-1} = 4\pi M/(R_P I_P K_{VCO})$. It follows that inserting a divider in a type II loop degrades both the stability and the settling speed, requiring a proportional increase in the charge-pump current.

The frequency-multiplying loop of Fig. 16.55(b) exhibits two interesting properties. First, unlike the voltage amplifier of Fig. 16.55(a), the PLL provides a multiplication factor *exactly* equal to M, a unique attribute resulting from phase locking. Second, the output frequency can be varied by changing the divide ratio M, an extremely useful property in synthesizing frequencies. Note that DLLs cannot perform such synthesis.

Frequency Synthesis Some systems require a periodic waveform whose frequency (1) must be very accurate (e.g., exhibit an error less than 10 ppm), and (2) can be varied in very fine steps (e.g., in steps of 30 kHz from 900 MHz to 925 MHz). Commonly encountered in wireless transceivers, such requirements can be met through frequency multiplication by PLLs.

Figure 16.57 shows the architecture of a phase-locked frequency synthesizer. The channel control input is a digital word that defines the value of M. Since $f_{out} = Mf_{REF}$, the relative accuracy of f_{out} is equal to that of f_{REF}. For this reason, f_{REF} is derived from a stable, low-noise crystal oscillator. Note that f_{out} varies in steps equal to f_{REF} if M changes by one each time.

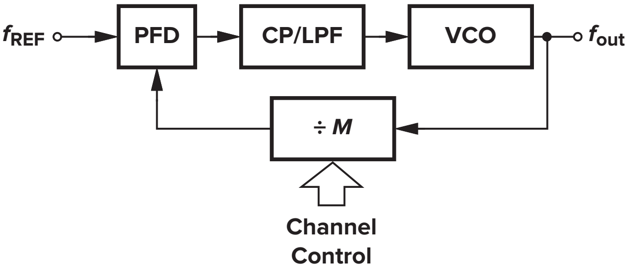


Figure 16.57 Frequency synthesizer.

CMOS frequency synthesizers achieving gigahertz output frequencies have been reported. Issues such as noise, sidebands, settling speed, frequency range, and power dissipation continue to challenge synthesizer designers.

16.5.2 Skew Reduction

The earliest usage of phase locking in digital systems was for skew reduction. Suppose a synchronous pair of data and clock lines enter a large digital chip, as shown in Fig. 16.58. Since the clock typically drives a large number of transistors and long interconnects, it is first applied to a large buffer. Thus, the clock distributed on the chip may suffer from substantial skew, ΔT, with respect to the data, an undesirable effect because it reduces the timing budget for on-chip operations.

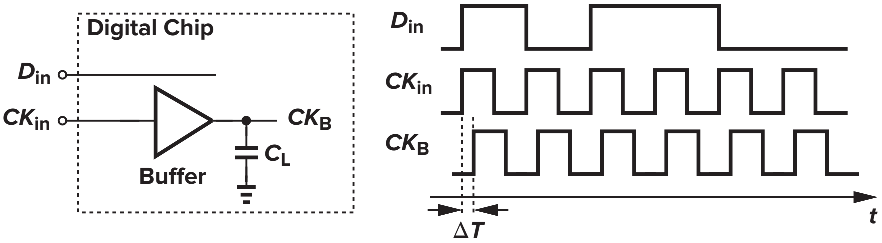


Figure 16.58 Skew between data and buffered clock.

Now consider the circuit shown in Fig. 16.59, where CK_{in} is applied to an on-chip PLL and the buffer is placed *inside* the loop. Since the PLL guarantees a nominally-zero phase difference between CK_{in} and CK_B, the skew is eliminated. From another point of view, the constant phase shift introduced by the buffer is divided by the infinite loop gain of the feedback system. Note that the VCO output, V_{VCO}, may not be aligned with CK_{in}, a nonetheless unimportant issue because V_{VCO} is not used.

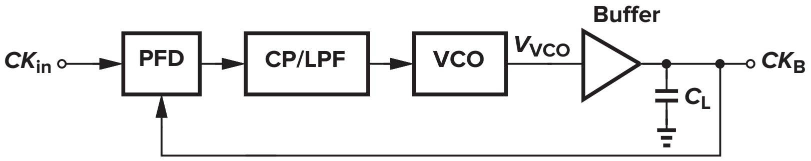


Figure 16.59 Use of a PLL to eliminate skew.

▶ **Example 16.13**

Construct the voltage-domain counterpart of the loop shown in Fig. 16.59.

Solution

The buffer creates a constant phase shift in the signal generated by the VCO. The voltage-domain counterpart therefore assumes the topology shown in Fig. 16.60. We have

$$(V_{in} - V_{out})A + V_M = V_{out} \tag{16.55}$$

and hence

$$V_{out} = \frac{AV_{in} + V_M}{1 + A} \tag{16.56}$$

As $A \to \infty$, $V_{out} \to V_{in}$.

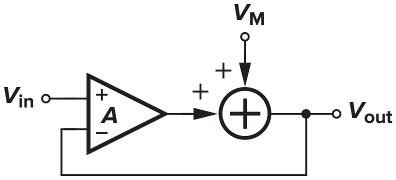


Figure 16.60

◀

We should note that the skew can be suppressed by a delay-locked loop as well. In fact, if frequency multiplication is not required, DLLs are preferred because they are less susceptible to noise.

16.5.3 Jitter Reduction

Recall from Sec. 16.3.2 that PLLs suppress fast jitter components at the input. For example, if a 1-GHz jittery signal is applied to a PLL having a bandwidth of 10 MHz, then input jitter components that vary faster than 10 MHz are attenuated. In a sense, the phase-locked loop operates as a narrowband filter centered around 1 GHz with a total bandwidth of 20 MHz. This is another important and useful property of PLLs.

Many applications must deal with jittery waveforms. Random binary signals experience jitter because of (1) crosstalk on the chip and in the package (Chapter 19), (2) package parasitics (Chapter 19), (3) additive electronic noise of devices, etc. Such waveforms are typically "retimed" by a low-noise clock so as to reduce the jitter. Illustrated in Fig. 16.61(a), the idea is to resample the midpoint of each bit by a D flipflop that is driven by the clock. However, in many applications, the clock may not be available independently. For example, an optical fiber carries only the random data stream, providing no separate clock waveform at the receive end. The circuit of Fig. 16.61(a) is therefore modified as shown in Fig. 16.61(b), where a "clock recovery circuit" (CRC) produces the clock from the data. Employing phase locking with a relatively narrow loop bandwidth, the circuit minimizes the effect of the input jitter on the recovered clock.

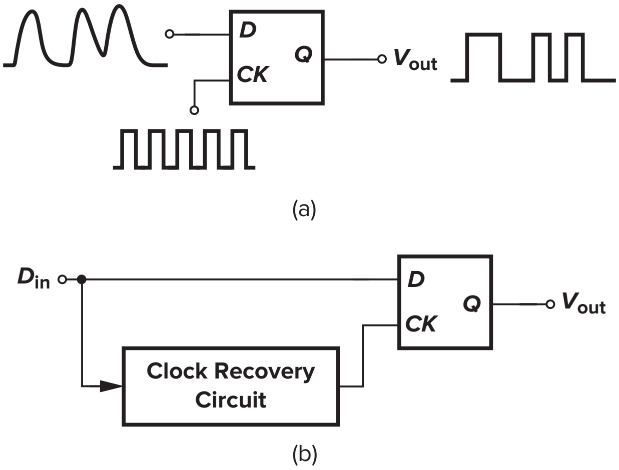


Figure 16.61 (a) Retiming data with D flipflop driven by a low-noise clock; (b) use of a phase-locked clock recovery circuit to generate the clock.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

16.1. The Gilbert cell (Chapter 4) operates as an XOR gate with large input swings and as an analog multiplier with small input swings. Prove that an analog multiplier can be used to detect the phase difference between two sinusoids. Is the input-output characteristic of such a phase detector linear?

16.2. Redraw the waveforms of Fig. 16.4(b) if the VCO frequency is lowered at $t = t_1$. If the phase error between V_{CK} and V_{VCO} before $t = t_1$ is equal to ϕ_0 and f_{VCO} is lowered from f_H to f_L, determine the minimum $t_2 - t_1$ that is sufficient for phase alignment.

16.3. Explain why the low-pass filter in Fig. 16.5(b) cannot be replaced by a high-pass filter.

16.4. A PLL using an XOR gate as a phase detector locks with $\phi_{in} - \phi_{out} \approx 90°$ if $K_{PD}K_{VCO}$ is large. Explain why.

16.5. Using the characteristic of Fig. 16.3 as an example, explain why the polarity of feedback in a PLL (without frequency detection) is unimportant. (Hint: prove that the loop locks regardless of whether the initial phase difference falls in the positive-slope region or the negative-slope region.)

16.6. Assuming a first-order LPF in Fig. 16.14, determine the transfer function Φ_{out}/Φ_{ex}, where Φ_{out} denotes the excess phase of V_{out}.

16.7. A VCO used in a type I PLL exhibits nonlinearity in its input-output characteristic, i.e., K_{VCO} varies across the tuning range. If the damping ratio must remain between 1 and 1.5, how much variation can be tolerated in K_{VCO}?

16.8. Prove that in the root locus of Fig. 16.20, $\cos\theta = \zeta$.

16.9. A type I PLL incorporates a VCO with $K_{VCO} = 100$ MHz/V, a PD with $K_{PD} = 1$ V/rad, and an LPF with $\omega_{LPF} = 2\pi(1\text{ MHz})$. Determine the step response of the PLL.

16.10. Explain why in the charge-pump PLL of Fig. 16.35, the control voltage of the VCO cannot be connected to the top plate of C_P.

16.11. Prove that the transfer function of the PFD/CP/LPF circuit in Fig. 16.35 is given by Eq. (16.43).

16.12. As illustrated in Fig. 16.45, mismatches between the UP and DOWN currents translate to phase offset at the input of a CPPLL. With the aid of the waveforms in Fig. 16.45, calculate the phase offset in terms of current mismatch.

16.13. For a VCO, we have $\omega_{out} = \omega_0 + K_{VCO}V_{cont}$. The control line experiences a small sinusoidal ripple, $V_{cont} = V_m \cos \omega_m t$. If the VCO is followed by a $\div M$ circuit, determine the output spectrum of the divider. Consider two cases: $\omega_0/M > \omega_m$ and $\omega_0/M < \omega_m$.

16.14. Prove that the root locus of a type II PLL is as shown in Fig. 16.37.

16.15. Determine the transfer function Φ_{out}/Φ_{ex} for the circuit of Fig. 16.14 if the PLL is modified to the architecture of Fig. 16.35.

16.16. When a charge-pump PLL incorporating a PFD is turned on, the VCO frequency may be far from the input frequency. Explain why the order of the PLL transfer function is lower by one while the PFD operates as a frequency detector.

References

[1] R. E. Best, *Phase-Locked Loops*, 2nd ed. (New York: McGraw-Hill, 1993).

[2] F. M. Gardner, *Phaselock Techniques*, 2nd ed. (New York: John Wiley & Sons, 1979).

[3] M. G. Johnson and E. L. Hudson, "A Variable Delay Line PLL for CPU-Coprocessor Synchronization," *IEEE J. of Solid-State Circuits,* vol. 23, pp. 1218–1223, October 1988.

[4] F. M. Gardner, "Charge-Pump Phase-Locked Loops," *IEEE Trans. Comm.*, vol. COM-28, pp. 1849–1858, November 1980.

[5] F. Herzel and B. Razavi, "A Study of Oscillator Jitter Due to Supply and Substrate Noise," *IEEE Transactions on Circuits and Systems, Part II*, vol. 46, pp. 56–62, January 1999.

[6] W. F. Egan, *Frequency Synthesis by Phase Lock* (New York: John Wiley & Sons, 1981).

[7] J. A. Crawford, *Frequency Synthesizer Design Handbook* (Boston: Artech House, 1994).

CHAPTER

17

Short-Channel Effects and Device Models

The square-law characteristics derived for MOSFETs in Chapter 2 provide moderate accuracies for devices with minimum channel lengths of greater than several microns, a value corresponding to technologies in production in the early 1980s. As device dimensions continue to scale down, reaching below 12 nm, higher-order effects necessitate more complex models so as to attain enough accuracy in simulations.

The problem of device models in CMOS technology has constantly haunted analog designers, manifesting itself as substantial discrepancies between simulated and measured results. A number of comprehensive books [1, 2, 3] and hundreds of papers deal with the subject in great detail, but our objective here is to provide a basic understanding of short-channel effects and review some of the SPICE models developed to reflect such phenomena. Knowledge of these issues also proves useful in interpreting the anomalies that the designer may encounter in SPICE simulations.

We first describe the ideal scaling theory of MOS transistors. Next, we study short-channel effects such as threshold voltage variation, velocity saturation, and the dependence of the output impedance on the drain-source voltage. We then review MOS device models, including Levels 1–3 and the BSIM series. Finally, we discuss charge and capacitance modeling, temperature dependence, and process corners.

17.1 ■ Scaling Theory

The two principal reasons for the dominance of CMOS technology in today's semiconductor industry are the zero static power dissipation of CMOS logic and the scalability of MOSFETs. In a paper published in 1974 [4], Dennard et al. recognized the tremendous potential of scaling MOS transistors, making predictions about speed and power dissipation of digital CMOS circuits as devices are shrunk.

The ideal scaling theory follows three rules: (1) reduce all lateral and vertical dimensions by $\alpha(> 1)$; (2) reduce the threshold voltage and the supply voltage by α; (3) increase all of the doping levels by α (Fig. 17.1). Since the dimensions and voltages scale together, all electric fields in the transistor remain

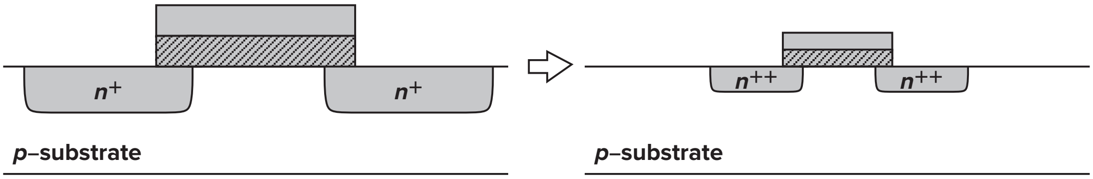


Figure 17.1 Ideal scaling of MOS transistor.

constant, hence the name "constant-field scaling." Note that $W, L, t_{ox}, V_{DD}, V_{TH}$, and the depth and perimeter of the source and drain junctions scale down by α.

Let us examine the saturation drain current of a square-law device after scaling. Writing

$$I_{D,scaled} = \frac{1}{2}\mu_n(\alpha C_{ox})\left(\frac{W/\alpha}{L/\alpha}\right)\left(\frac{V_{GS}}{\alpha} - \frac{V_{TH}}{\alpha}\right)^2 \tag{17.1}$$

$$= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2\frac{1}{\alpha} \tag{17.2}$$

we observe that the current capability of the transistor *drops* by a factor of α. Note that the same result applies for the drain current in the triode region. The advantage of scaling, however, lies in the reduction of capacitances and power dissipation. The total channel capacitance is

$$C_{ch,scaled} = \frac{W}{\alpha}\frac{L}{\alpha}(\alpha C_{ox}) \tag{17.3}$$

$$= \frac{1}{\alpha}WLC_{ox} \tag{17.4}$$

To calculate the source/drain junction capacitance, we first analyze the effect of ideal scaling on the total width of the depletion region. Recall that this width is given by

$$W_d = \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(\phi_B + V_R)} \tag{17.5}$$

where N_A and N_D denote the doping levels of the two sides of the junction, $\phi_B = V_T \ln(N_A N_D/n_i^2)$, and V_R is the reverse-bias voltage. The built-in potential, ϕ_B, is a weak function of $N_A N_D$, and in fact it *increases* if $N_A N_D$ is scaled up by α^2. For now, we assume $V_R \gg \phi_B$ so that

$$W_{d,scaled} \approx \sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{\alpha N_A} + \frac{1}{\alpha N_D}\right)\frac{V_R}{\alpha}} \tag{17.6}$$

$$\approx \frac{1}{\alpha}\sqrt{\frac{2\epsilon_{si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)V_R} \tag{17.7}$$

Thus, as with other dimensions, the width of each depletion region scales down by α, increasing the depletion-region capacitance per unit area by the same factor.

As illustrated in Fig. 17.2, the bottom-plate capacitance of the S/D junction (per unit area), C_j, increases by a factor of α. The sidewall capacitance (per unit width), C_{jsw}, on the other hand, remains constant because the depth of the junction is reduced by α. It follows that

$$C_{S/D,scaled} = \frac{W}{\alpha}\frac{E}{\alpha}(\alpha C_j) + 2\left(\frac{W}{\alpha} + \frac{E}{\alpha}\right)(C_{jsw}) \tag{17.8}$$

$$= [WEC_j + 2(W + E)C_{jsw}]\frac{1}{\alpha} \tag{17.9}$$

All of the capacitances therefore decrease by the scaling factor.

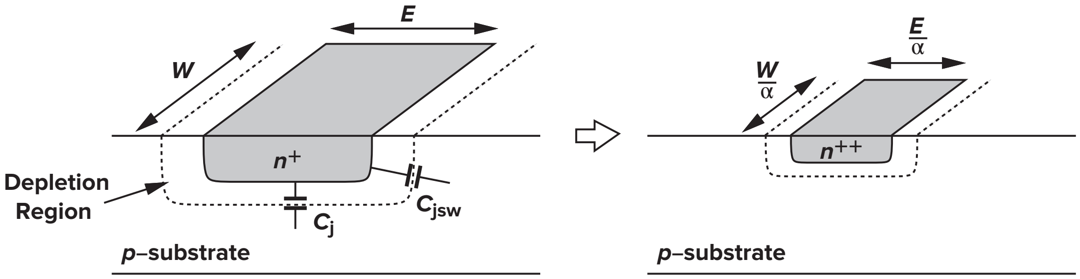


Figure 17.2 Scaling of S/D junction capacitances.

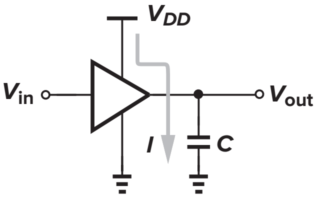


Figure 17.3 CMOS inverter.

In digital applications, the scaling of the gate delay and power dissipation is of interest. Approximating the delay of a CMOS inverter by $T_d = (C/I)V_{DD}$ (Fig. 17.3), we have

$$T_{d,scaled} = \frac{C/\alpha}{I/\alpha}\frac{V_{DD}}{\alpha} \tag{17.10}$$

$$= \left(\frac{C}{I}V_{DD}\right)\frac{1}{\alpha} \tag{17.11}$$

We conclude that the speed of digital circuits can potentially increase by the scaling factor. For power dissipation, we write $P = fCV_{DD}^2$, where f is the operating frequency. Thus, $P_{scaled} = f(C/\alpha)(V_{DD}/\alpha)^2 = fCV_{DD}^2/\alpha^3$ if f and the number of gates in the circuit remain constant. Note that the layout density, i.e., the number of transistors per unit area, also scales by α^2.

The reduction of power and delay and the increase in circuit density make scaling extremely attractive for digital systems. Based on these observations, Gordon Moore predicted in 1975 [5] that MOS device dimensions would continue to scale down by a factor of two every three years and the number of transistors per chip would double every one to two years. Such trends have indeed persisted over the past 40 years.

Let us now consider the effect of ideal scaling in analog circuits. Writing the transconductance as

$$g_{m,scaled} = \mu(\alpha C_{ox})\frac{W/\alpha}{L/\alpha}\frac{V_{GS}-V_{TH}}{\alpha} \tag{17.12}$$

$$= \mu C_{ox}\frac{W}{L}(V_{GS}-V_{TH}) \tag{17.13}$$

we note that the transconductance remains constant if all of the dimensions and voltages (and currents) scale down. To calculate the output impedance in saturation, we first observe from Fig. 17.4 and Eq. (17.7) that the width of the depletion region around the drain decreases by α, and hence $\Delta L/L$ remains

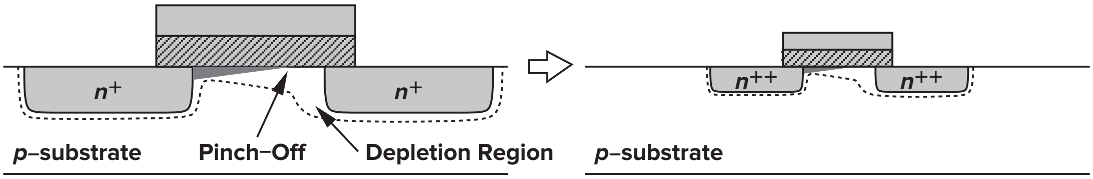


Figure 17.4 Effect of scaling on pinch-off.

constant. Since $\lambda = (\Delta L/L)/V_{DS}$ (Chapter 2), λ increases by α and

$$r_{O,scaled} = \frac{1}{\alpha\lambda\dfrac{I_D}{\alpha}} \tag{17.14}$$

$$= \frac{1}{\lambda I_D} \tag{17.15}$$

Thus, the intrinsic gain, $g_m r_O$, remains constant. Unfortunately, in practice, $g_m r_O$ has dropped considerably.

The greatest impact of scaling on analog circuits is the reduction of the supply voltage. With ideal scaling, the maximum allowable voltage swings decrease by a factor of α, lowering the dynamic range[1] of the circuit. For example, if the lower end of the dynamic range is limited by thermal noise, then scaling V_{DD} by α decreases the dynamic range by the same factor because g_m and hence thermal noise remain constant. Of course, since for analog circuits $(V_{DD}/\alpha)(I_{DD}/\alpha) = (V_{DD}I_{DD}/\alpha)^2$, the power dissipation drops by α^2.

In order to restore the dynamic range, the transconductance of the transistors must be increased by a factor of α^2 because thermal noise voltages and currents scale with $\sqrt{g_m}$. Thus, since voltage scaling requires that $V_{GS} - V_{TH}$ decrease by a factor of α, we note from $g_m = 2I_D/(V_{GS} - V_{TH})$ that I_D must increase by the same factor, leading to a power dissipation of $(V_{DD}/\alpha)(\alpha I_D) = V_{DD}I_D$. Also, from $g_m = \mu C_{ox}(W/L)(V_{GS} - V_{TH})$, we conclude that if C_{ox} is scaled up by α and L and $V_{GS} - V_{TH}$ are scaled down by α, then W must *increase* by α (whereas in ideal scaling, it would decrease by this factor). That is, for a constant (thermal-noise limited) dynamic range, ideal scaling of linear analog circuits requires a *constant* power dissipation and a *higher* device capacitance, e.g., $(\alpha W)(L/\alpha)(\alpha C_{ox}) = \alpha WLC_{ox}$. Interestingly, if the lower end of the dynamic range is determined by kT/C noise, then to maintain a constant slew rate in switched-capacitor circuits, the bias current must scale up by a factor of α^2, resulting in an increase in the power dissipation. (Problem 17.17.3).

In practice, technology scaling has deviated from the ideal, constant-field scenario considerably. The supply voltage and MOS threshold voltage have not scaled as rapidly as device dimensions. For example, V_{DD} has decreased from 5 V to 2.5 V and V_{TH} from 0.8 V to 0.4 V as minimum channel length has dropped from 1 μm to 0.25 μm. Furthermore, many "short-channel" effects have plagued the transistors, making it difficult to obtain all of the benefits that would accrue with ideal scaling.

The reluctance of circuit designers to use a lower supply voltage and the fundamental limitations in decreasing the MOS threshold voltage have led to another scaling scenario: constant-voltage scaling. In this case, the device dimensions shrink by α, the doping levels increase by α, and the voltages remain constant, thereby increasing the electric fields by α. Such high electric fields both raise the possibility of device breakdown and exacerbate short-channel effects. In reality, technology scaling has followed a

[1] Dynamic range is loosely defined as the maximum allowable voltage swing divided by the total noise voltage in the band of interest.

mixture of constant-field and constant-voltage trends, thus demanding innovative device design so as to achieve reliability and performance.

17.2 ■ Short-Channel Effects

In order to appreciate the need for sophisticated device models, we briefly study some of the phenomena that manifest themselves for short channels. As we will see, a basic understanding of these effects also proves essential to the design of analog (and digital) circuits.

Small-geometry effects arise because five factors deviate the scaling from the ideal scenario: (1) the electric fields tend to increase because the supply voltage has not scaled proportionally; (2) the built-in potential term in Eq. (17.5) is neither scalable nor negligible; (3) the depth of S/D junctions cannot be reduced easily; (4) the mobility decreases as the substrate doping increases; and (5) the subthreshold slope (described below) is not scalable.

17.2.1 Threshold Voltage Variation

The choice of the threshold voltage is based on the device performance in typical circuit applications. The upper bound is roughly equal to $V_{DD}/4$ to avoid degrading the speed of digital CMOS gates. The lower bound is determined by several factors: the subthreshold behavior, variation with temperature and process, and dependence upon the channel length [6].

Let us first consider the subthreshold behavior. For long-channel devices, the subthreshold drain current can be expressed as

$$I_D = \mu C_d \frac{W}{L} V_T^2 \left(\exp \frac{V_{GS} - V_{TH}}{\zeta V_T}\right)\left(1 - \exp \frac{-V_{DS}}{V_T}\right) \tag{17.16}$$

where $C_d = \sqrt{\epsilon_{si} q N_{sub}/(4\phi_B)}$ denotes the capacitance of the depletion region under the gate area, $V_T = kT/q$, and $\zeta = 1 + C_d/C_{ox}$ [6]. Equation (17.16) reveals two interesting properties. First, as V_{DS} exceeds a few V_T, I_D becomes independent of the drain-source voltage and the relationship reduces to Eq. (2.33). Second, under this condition, the slope of I_D on a logarithmic scale equals

$$\frac{\partial(\log_{10} I_D)}{\partial V_{GS}} = (\log_{10} e)\frac{1}{\zeta V_T} \tag{17.17}$$

The inverse of this quantity is usually called the "subthreshold slope," S:

$$S = 2.3 V_T \left(1 + \frac{C_d}{C_{ox}}\right) \ \text{V/dec} \tag{17.18}$$

For example, if $C_d = 0.67 C_{ox}$, then $S = 100$ mV/dec, suggesting that a change of 100 mV in V_{GS} leads to a tenfold reduction in the drain current. In order to turn off the transistor by lowering V_{GS} below V_{TH}, S must be as *small* as possible, i.e., C_d/C_{ox} must be minimized.

The relatively constant magnitude of S severely limits the scaling of the threshold voltage. For example, a subthreshold slope of 80 mV/dec imposes a lower bound of 400 mV for V_{TH} if the "off current" must be roughly five orders of magnitude lower than the "on current."

The difficulty in scaling V_{TH} becomes even more serious if we take into account the variation of V_{TH} with temperature and process. The threshold voltage exhibits a temperature coefficient of approximately

−1 mV/K, yielding a 50-mV change across the commercial temperature range (0 to 50°C).[2] Process-induced variation is also in the vicinity of 50 mV, raising the margin to approximately 100 mV. Thus, it is difficult to reduce V_{TH} below several hundred millivolts.

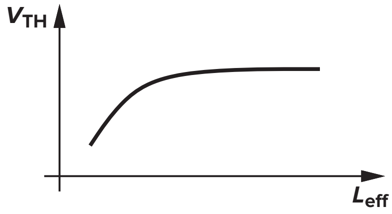


Figure 17.5 Variation of threshold with channel length.

An interesting phenomenon observed in scaled transistors is the dependence of the threshold voltage on the channel length. As shown in Fig. 17.5, transistors fabricated on the same wafer but with different lengths yield lower V_{TH} as L decreases. This is because the depletion regions associated with the source and drain junctions protrude into the channel area considerably, thereby reducing the immobile charge that must be imaged by the charge on the gate (Fig. 17.6). In other words, part of the immobile charge in the substrate is now imaged by the charge inside the source and drain areas rather than by the charge on the gate. As a result, the gate voltage required to create an inversion layer decreases. Since the channel length cannot be controlled accurately during fabrication, this effect introduces additional variations in V_{TH}. The implication of this phenomenon in analog design is that if the length of a device is increased so as to achieve a higher output impedance, then the threshold voltage also increases by as much as 100 to 200 mV.

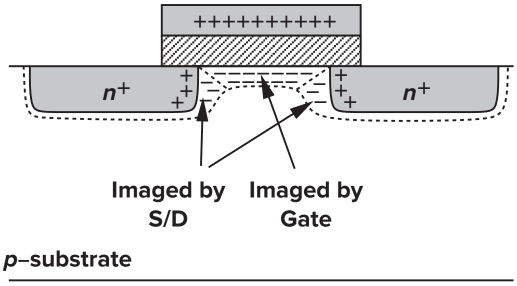


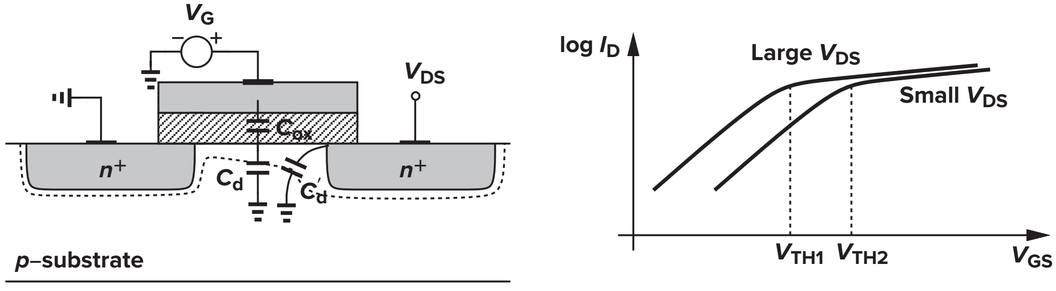
Figure 17.6 Charge sharing between source/drain depletion regions and the channel depletion region.

Another short-channel phenomenon related to the threshold voltage is "drain-induced barrier lowering" (DIBL). Recall from Chapter 2 that in weak inversion, as the gate voltage rises, the surface potential becomes more positive [Fig. 17.7(a)], attracting carriers from the source region. In short-channel devices, the *drain* voltage also makes the surface more positive by creating a two-dimensional field in the depletion region [6]. In essence, the drain introduces a capacitance C'_d that raises the surface potential in a manner similar to C_d. As a result, the barrier to the flow of charge and hence the threshold voltage are decreased. This effect manifests itself if the plot of Fig. 2.28 is drawn in both deep triode and saturation regions [Fig. 17.7(b)].

The principal impact of DIBL on circuit design is the degraded output impedance. This point is explained in Sec. 17.2.5.

Reverse Short-Channel Effect In nanometer CMOS technologies, the threshold voltage *decreases* as the channel length increases from its minimum value. To analyze this effect, let us consider the cross

[2] Interestingly, as the temperature rises, so does S, further exacerbating the situation.



(a) (b)

Figure 17.7 (a) DIBL in a short-channel device; (b) effect of DIBL on current characteristic.

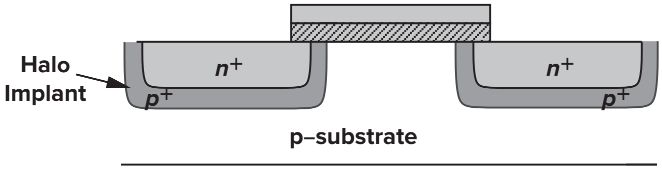


Figure 17.8 MOS structure with halo implant.

section of a modern device, shown in Fig. 17.8, wherein a “halo” implant of heavy doping surrounds the source and drain junctions. This implant reduces the penetration of the drain depletion region into the channel area, thereby improving the device characteristics.

Now recall from Chapter 2 that the threshold voltage is a function of the substrate doping level, N_{sub}. We have

$$V_{TH} = \phi_{MS} + 2\phi_F + \frac{Q_{dep}}{C_{ox}} \tag{17.19}$$

where both $\phi_F = (kT/q)\ln(N_{sub}/n_i)$ and $Q_{dep} = \sqrt{4q\epsilon_{si}|\phi_F|N_{sub}}$ increase as N_{sub} increases. Due to the nonuniform substrate doping along the channel in Fig. 17.8, the “local” threshold voltage also varies from the source to the drain. We can take the average along the channel to obtain an overall threshold for a given device structure. We then observe that, as the channel length increases, the average substrate doping decreases, and so does the threshold voltage.

17.2.2 Mobility Degradation with Vertical Field

At large gate-source voltages, the high electric field developed between the gate and the channel confines the charge carriers to a narrower region below the oxide-silicon interface, leading to more carrier scattering and hence lower mobility. Since scaling has substantially deviated from the constant-field scenario, small-geometry devices experience significant mobility degradation. An empirical equation modeling this effect is

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})} \tag{17.20}$$

where μ_0 denotes the “low-field” mobility and θ is a fitting parameter roughly equal to $(10^{-7}/t_{ox})$ V^{-1} [7]. For example, if $t_{ox} = 100$ Å, then $\theta \approx 1$ V^{-1} and the mobility begins to fall considerably as the overdrive exceeds 100 mV. Note that θ rises as t_{ox} drops because the electric field in the oxide becomes stronger.

In addition to lowering the current capability and transconductance of MOSFETs, mobility degradation causes the I/V characteristic to deviate from the simple square-law behavior. Specifically, whereas a square-law device generates only even harmonics in its drain current in response to a sinusoidal gate-source voltage, Eq. (17.20) predicts odd harmonics as well. In fact, writing

$$I_D = \frac{1}{2}\frac{\mu_0 C_{ox}}{1+\theta(V_{GS}-V_{TH})}\frac{W}{L}(V_{GS}-V_{TH})^2 \tag{17.21}$$

and assuming that $\theta(V_{GS}-V_{TH}) \ll 1$, we obtain

$$I_D \approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}[1-\theta(V_{GS}-V_{TH})](V_{GS}-V_{TH})^2 \tag{17.22}$$

$$\approx \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\left[(V_{GS}-V_{TH})^2-\theta(V_{GS}-V_{TH})^3\right] \tag{17.23}$$

This is a rough approximation, but it reveals the existence of higher harmonics in the drain current.

The mobility degradation with the vertical field affects the device transconductance as well. This is studied in Problem 17.9.

17.2.3 Velocity Saturation

The mobility of carriers also depends on the *lateral* electric field in the channel, which is beginning to drop as the field reaches levels of 1 V/μm. Since the carrier velocity $v = \mu E$, we note that v approaches a saturated value, about 10^7 cm/s, for sufficiently high fields. Thus, as carriers enter the channel from the source and accelerate toward the drain, they may eventually reach a saturated velocity at some point along the channel.[3] In the extreme case, where carriers experience velocity saturation along the entire channel, we can rewrite Eq. (2.2) as

$$I_D = v_{sat} Q_d \tag{17.24}$$

$$= v_{sat} W C_{ox}(V_{GS}-V_{TH}) \tag{17.25}$$

Interestingly, the current is *linearly* proportional to the overdrive voltage and does not depend on the length. In fact, as shown in Fig. 17.9, I_D-V_{DS} characteristics of devices with $L < 1\ \mu$m reveal velocity saturation because equal increments in $V_{GS}-V_{TH}$ result in roughly equal increments in I_D. We also note that $g_m = v_{sat} W C_{ox}$, concluding that the transconductance is a weak function of the drain current and channel length in the velocity-saturation regime.

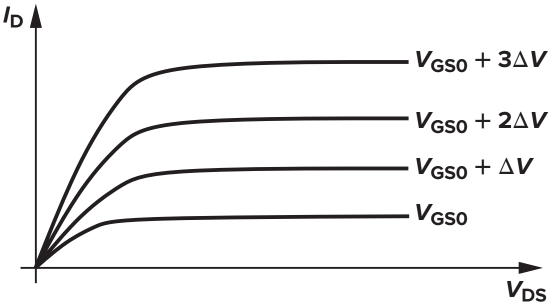


Figure 17.9 Effect of velocity saturation on drain-current characteristics.

[3] Even in long-channel devices, carriers experience velocity saturation if the drain-source voltage is high enough to pinch off the channel. At the pinch-off point, the mobile charge density is near zero, the electric field is very large, and hence the velocity of carriers is saturated.

Under typical bias conditions, MOSFETs experience some velocity saturation, displaying a characteristic between linear and square-law behavior. An important consequence is that, as V_{GS} increases, the drain current saturates well before pinch-off occurs. As shown in Fig. 17.10(a), carriers reach velocity saturation if V_{DS} exceeds $V_{D0} < V_{GS} - V_{TH}$, yielding a constant current quite a lot lower than that obtained if the device saturated for $V_{DS} > V_{GS} - V_{TH}$. Furthermore, as illustrated in Fig. 17.10(b), since an increment in V_{GS} gives a smaller increment for I_D when velocity saturation occurs, the transconductance is also lower than that predicted by the square law.

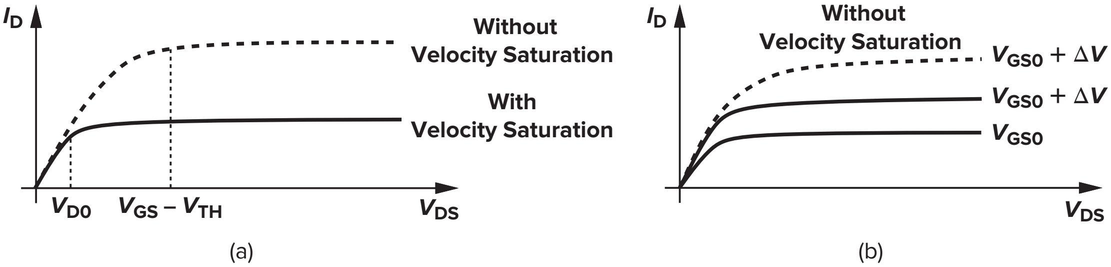


Figure 17.10 Effect of velocity saturation: (a) premature drain current saturation; (b) reduction of transconductance.

A compact and versatile equation developed to represent velocity saturation (in the saturation region) is

$$I_D = WC_{ox}v_{sat}\frac{(V_{GS}-V_{TH})^2}{V_{GS}-V_{TH}+2\dfrac{v_{sat}L}{\mu_{eff}}} \tag{17.26}$$

where μ_{eff} is given by Eq. (17.20) [7, 8]. The same work provides the following equation for the drain-source voltage at the onset of premature saturation [V_{D0} in Fig. 17.10(a)]:

$$V_{DS,sat} = \frac{2\mu_{eff}L(V_{GS}-V_{TH})}{2\mu_{eff}L+V_{GS}-V_{TH}} \tag{17.27}$$

Equation (17.26) provides two interesting results. First, if L or v_{sat} is large, the expression reduces to the square-law relationship. Second, if the *overdrive* voltage is so small that the denominator of (17.26) is approximated as $2v_{sat}L/\mu_{eff}$ and $\mu_{eff} \approx \mu_0$, then the device still follows the square-law behavior even if L is relatively small. For example, if $v_{sat} \approx 10^7$ cm/s, $L = 0.25$ μm, and $\mu_0 \approx 350$ cm^2/V/s, we have $2v_{sat}L/\mu_0 \approx 1.43$ V, recognizing that for overdrive voltages of a few hundred millivolts, the transistor operation is somewhat close to the square law. Thus, the simplified treatment of Chapter 2 can still provide insight for many analog applications.

Equation (17.26) can be further simplified to yield additional results. Substituting for μ_{eff} from Eq. (17.20), we have

$$I_D = WC_{ox}v_{sat}\frac{(V_{GS}-V_{TH})^2}{V_{GS}-V_{TH}+\dfrac{2v_{sat}L}{\mu_0}[1+\theta(V_{GS}-V_{TH})]} \tag{17.28}$$

$$= WC_{ox}v_{sat}\frac{(V_{GS}-V_{TH})^2}{\dfrac{2v_{sat}L}{\mu_0}+\left(1+\dfrac{2v_{sat}L\theta}{\mu_0}\right)(V_{GS}-V_{TH})} \tag{17.29}$$

$$= \frac{1}{2}\mu_0 C_{ox}\frac{W}{L}\frac{(V_{GS}-V_{TH})^2}{1+\left(\dfrac{\mu_0}{2v_{sat}L}+\theta\right)(V_{GS}-V_{TH})} \tag{17.30}$$

This equation is similar to (17.21), implying that the degradation of the mobility with both lateral and vertical fields can be represented by adding the terms $\mu_0/(2v_{sat}L)$ and θ. Thus, the results obtained from (17.21) apply here as well. For example, the drain current contains high-order nonlinear terms. Equation (17.30) can also predict the transconductance (Problem 17.10).

17.2.4 Hot Carrier Effects

Short-channel MOSFETs may experience high lateral electric fields if the drain-source voltage is large. While the *average* velocity of carriers saturates at high fields, the instantaneous velocity and hence the kinetic energy of the carriers continue to increase, especially as they accelerate toward the drain. These are called “hot” carriers [2].

In the vicinity of the drain region, hot carriers may “hit” the silicon atoms at high speeds, thereby creating impact ionization. As a result, new electrons and holes are generated, with the electrons absorbed by the drain and the holes by the substrate. Thus, a finite drain-substrate current appears. Also, if the carriers acquire a very high energy, they may be injected into the gate oxide and even flow out the gate terminal, introducing a gate current. The substrate and gate currents are often measured to study hot carrier effects.

The scaling of technologies proceeds so as to minimize hot carrier effects. This limitation and other breakdown phenomena make the supply voltage scaling inevitable.

In nanometer technologies, hot carrier effects have subsided. This is because the energy required to create an electron-hole pair, $E_g \approx 1.12$ eV, is simply not available if the supply voltage is around 1 V. That is, for an arbitrarily short channel and even in the absence of any lattice, an electron cannot attain 1.2 eV by traveling from 0 V at the source to 1 V at the drain. (Statistically, a small fraction of electrons may reach E_g at a finite temperature, but the effect is negligible.)

17.2.5 Output Impedance Variation with Drain-Source Voltage

In modeling channel-length modulation by a single constant λ, we have assumed that the output impedance of the transistor, r_O, is constant in the saturation region. In reality, however, r_O varies with V_{DS}. As V_{DS} increases and the pinch-off point moves toward the source, the rate at which the depletion region around the source becomes wider decreases, resulting in a higher incremental output impedance. Illustrated in Fig. 17.11, this effect is somewhat similar to the variation of the capacitance of a reversed-biased *pn* junction: with a small reverse bias, the width of the depletion region is a strong function of the voltage applied to the junction and with a large reverse bias, a weak function.

In this regime, the output impedance can be approximated as

$$r_O = \frac{2L}{1-\dfrac{\Delta L}{L}}\frac{1}{I_D}\sqrt{\frac{qN_B}{2\epsilon_{si}}(V_{DS}-V_{DS,sat})} \tag{17.31}$$

where $V_{D,sat}$ is the drain-source voltage at the onset of pinch-off [9]. Another approximation developed in conjunction with (17.26) and (17.27) is described in [8].

In short-channel devices, as V_{DS} increases further, drain-induced barrier lowering becomes significant, reducing the threshold voltage and increasing the drain current. This effect roughly cancels that expressed by (17.31), giving a relatively constant output impedance. At sufficiently high drain voltages, impact ionization near the drain produces a large current (flowing from the drain into the substrate), in essence lowering the output impedance. The overall behavior of r_O is plotted in Fig. 17.12.

The variation of r_O gives rise to nonlinearity in many circuits. In an op amp, for example, as the output voltage varies, so does the output impedance and hence the voltage gain of the circuit. Furthermore, impact

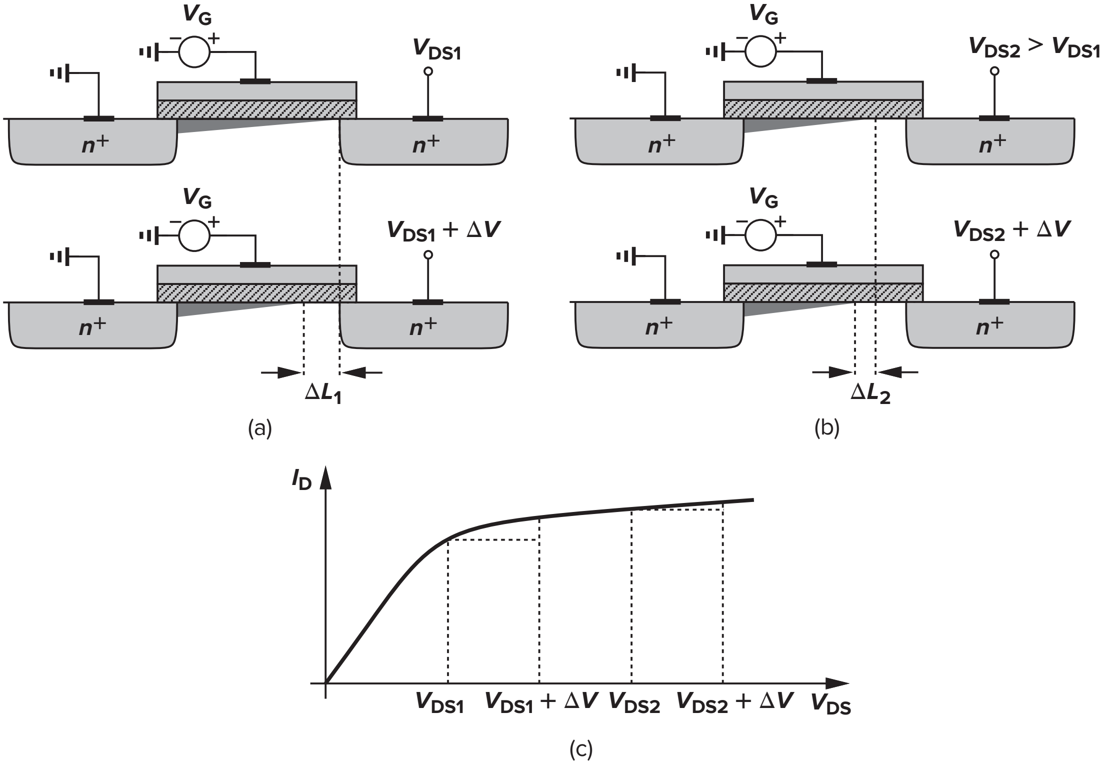


Figure 17.11 Decrement in channel length for (a) small V_{DS} and (b) large V_{DS} and (c) the resulting slope change.

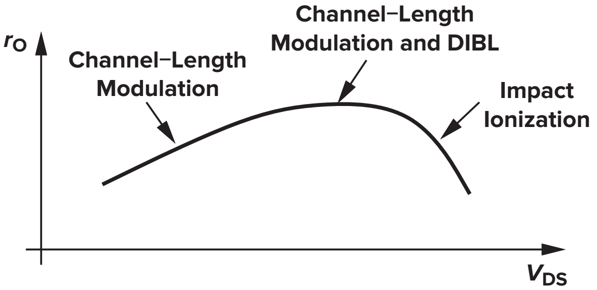


Figure 17.12 Overall variation of output resistance as a function of V_{DS}.

ionization limits the maximum gain that can be obtained from cascode structures because it introduces a small-signal resistance from the drain to the *substrate* rather than to the source.

17.3 ■ MOS Device Models

Since the introduction of the first MOS model in the mid-1960s [10], a tremendous amount of research has been expended on improving the accuracy of models as device dimensions scale down. Developed between the mid-1960s and the late 1970s, the Level 1, 2, and 3 models consecutively included higher-order effects so as to provide reasonable accuracy with respect to measured transistor characteristics for channel lengths as small as 1 μm. Following this set were the Compact Short-Channel IGFET Model (CSIM) from AT&T Bell Laboratories and the Berkeley Short-Channel IGFET Model (BSIM) from University of California, Berkeley, in the mid-1980s. These models proved inadequate for analog design and were followed by BSIM2, HSPICE level 28, BSIM3, BSIM4, and a number of others in the 1980s and 1990s.

MOS device modeling continues to pose a challenge—especially for high-frequency operation. Our objective is to develop a basic understanding of some of the models to the extent necessary for simulations. We should also mention that the utility of a model is given by the accuracy it provides in various regions of operation for different device dimensions, the ease with which its parameters can be measured, and the efficiency that it allows in simulations. The interested reader is referred to [1] for in-depth coverage.

17.3.1 Level 1 Model

Also known as the Shichman and Hodges Model [10], this representation uses the parameters listed in Table 2.1 and is based on the following equations:

$$I_D = \frac{1}{2} K_P \frac{W}{L - 2L_D} \left[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2\right](1 + \lambda V_{DS}) \quad \text{Triode Region} \tag{17.32}$$

$$I_D = \frac{1}{2} K_P \frac{W}{L - 2L_D} (V_{GS} - V_{TH})^2 (1 + \lambda V_{DS}) \quad \text{Saturation Region} \tag{17.33}$$

where $K_P = \mu C_{ox}$ and $V_{TH} = V_{TH0} + \gamma(\sqrt{2\phi_B - V_{BS}} - \sqrt{2\phi_B})$. Note that this model does not include subthreshold conduction or any short-channel effects.

The device capacitances are represented according to the simple model described in Chapter 2, but with one modification. Since in that model, C_{GS} abruptly changes from $(2/3)WLC_{ox} + WC_{ov}$ in saturation to $(1/2)WLC_{ox} + WC_{ov}$ in the triode region [and C_{GD} from WC_{ov} to $(1/2)WLC_{ox} + WC_{ov}$], most computation algorithms experience convergence difficulties here. For this reason, C_{GS} and C_{GD} in the triode region are formulated as

$$C_{GS} = \frac{2}{3} WLC_{ox} \left\{1 - \frac{(V_{GS} - V_{DS} - V_{TH})^2}{[2(V_{GS} - V_{TH}) - V_{DS}]^2}\right\} + WC_{ov} \tag{17.34}$$

$$C_{GD} = \frac{2}{3} WLC_{ox} \left\{1 - \frac{(V_{GS} - V_{TH})^2}{[2(V_{GS} - V_{TH}) - V_{DS}]^2}\right\} + WC_{ov} \tag{17.35}$$

$$C_{GB} = 0. \tag{17.36}$$

We note that if the device operates at the edge of saturation, $V_{GS} - V_{DS} = V_{TH}$, $C_{GS} = (2/3)WLC_{ox} + WC_{ov}$, and $C_{GD} = WC_{ov}$. Thus, the capacitance values change continuously from one region to another.

The Level 1 model maintains reasonable I/V accuracy for channel lengths as small as roughly 4 μm, but it still predicts the output impedance of transistors in saturation quite poorly.

17.3.2 Level 2 Model

The Level 1 model began to manifest its shortcomings as channel lengths fell below approximately 4 μm. The Level 2 model was then developed to represent many high-order effects.

An assumption that we made in Chapter 2 in deriving the square-law characteristics was a constant threshold voltage along the channel. This assumption is not correct even for long-channel devices because the charge in the depletion region under the channel varies according to the local voltage (Fig. 17.13). Since the inversion layer and the depletion region must image the charge on the gate, as the inversion layer

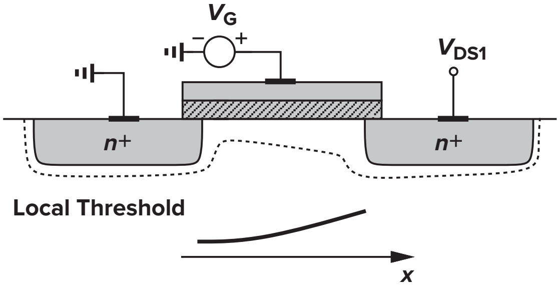


Figure 17.13 Variation of threshold along the channel.

vanishes in the direction toward the drain, the depletion region must enclose more charge. Performing the integration in Sec. 2.2.2 with a varying threshold voltage yields [1]

$$I_D = \mu C_{ox}\frac{W}{L}\{(V_{GS} - V_{TH0})V_{DS} - \frac{V_{DS}^2}{2} \\ -\frac{2}{3}\gamma[(V_{DS} - V_{BS} + 2\phi_F)^{3/2} - (-V_{BS} + 2\phi_F)^{3/2}]\} \tag{17.37}$$

Interestingly, even for $V_{BS} = 0$, I_D exhibits some dependence on γ. Moreover, for small V_{DS}, the equation reduces to that of the Level 1 model, but for large V_{DS}, the drain current is less than that predicted by the square law. It can also be shown that the edge of the saturation region is given by [1]

$$V_{D,sat} = V_{GS} - V_{TH0} - \phi_F + \gamma^2\left[1 - \sqrt{1 + \frac{2}{\gamma^2}(V_{GS} - V_{TH0} + \phi_F)}\right] \tag{17.38}$$

In the saturation region, the drain current is

$$I_{DS} = I_{D,sat}\frac{1}{1 - \lambda V_{DS}} \tag{17.39}$$

where $I_{D,sat}$ is calculated from (17.37) for $V_{DS} = V_{DS,sat}$.

Modeling channel-length modulation or, more generally, the finite output impedance has always remained a difficult problem. Representing such phenomena by only λ is far from accurate. In the Level 2 implementation, if λ is not specified, it is obtained by calculating the width of the depletion region between the pinch-off point and the edge of the drain. Using simple relationships for the depletion region of a *pn* junction, we can write

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}[\phi_B + (V_{DS} - V_{D,sat})]} \tag{17.40}$$

where $V_{D,sat}$ denotes the pinch-off voltage.[4]

The principal difficulty with the above approach is that both the drain current and its derivative are discontinuous at the edge of the triode region [1]! To resolve this issue, ΔL is actually obtained by a

[4] The junction is considered "one-sided" here; i.e., the drain doping level is much higher.

"fixed-up" equation:

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_{sub}}\left(V_1 + \sqrt{1 + V_1^2}\right)} \tag{17.41}$$

where $V_1 = (V_{DS} - V_{D,sat})/4$. The channel-length modulation coefficient is then expressed as $\lambda = \Delta L/(LV_{DS})$. An attribute of (17.41) is that the output conductance of the transistor varies as V_{DS} increases, an effect not represented by the first-order model using a constant λ.

The Level 2 model also includes the degradation of the mobility with the vertical field in the channel. The mobility is calculated from

$$\mu_s = \mu_0\left(\frac{\epsilon_{si}}{C_{ox}} \cdot \frac{U_c}{V_{GS} - V_{TH} - U_t V_{DS}}\right)^{U_e} \tag{17.42}$$

where U_c denotes the gate-channel critical electric field, U_t is a fitting parameter between 0 and 0.5, and U_e is an exponent in the vicinity of 0.15.

The subthreshold behavior implemented in the Level 2 model defines a voltage V_{on} as $V_{on} = V_{TH} + \zeta V_T$, where $\zeta = 1 + (qN_{FS}/C_{ox}) + C_d/C_{ox}$, and N_{FS} is an empirical constant. The drain current is then expressed as

$$I_{DS} = I_{on} \exp \frac{V_{GS} - V_{on}}{\zeta V_T} \tag{17.43}$$

where I_{on} is the drain current calculated in strong inversion [Eq. (17.37)] for $V_{GS} = V_{on}$. An important drawback of this representation is the discontinuity in the slope of I_D as the device goes from the subthreshold region to strong inversion (Fig. 17.14), leading to various difficulties and errors in simulation.

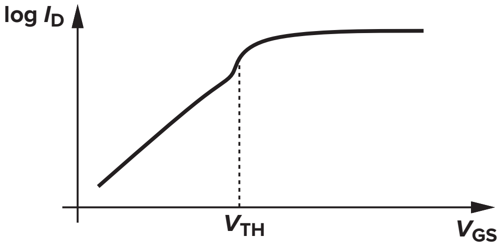


Figure 17.14 Kink in drain current characteristic in Level 2 model.

In addition to the above effects, the Level 2 model represents two other short-channel phenomena: the variation of V_{TH} with L, and velocity saturation. The implementation of these effects is quite involved and can be found in [1].

Measured data [1] indicate that the Level 2 model provides reasonable I/V accuracy for wide, short devices in the saturation region with $L \approx 0.7\ \mu$m, but it suffers from substantial error in representing the output impedance and the transition point between the saturation and triode regions. For narrow or long devices, the model is inaccurate.

17.3.3 Level 3 Model

The Level 3 model realization is somewhat similar to the Level 2 model, with some equations simplified and many empirical constants introduced to improve the accuracy for channel lengths as small as $1\ \mu$m.

This model expresses the threshold voltage as

$$V_{TH} = V_{TH0} + F_s\gamma\sqrt{2\phi_F - V_{BS}} + F_n(2\phi_F - V_{BS}) + \xi\frac{8.15\times 10^{-22}}{C_{ox}L_{eff}^3}V_{DS} \tag{17.44}$$

where F_s and F_n represent short-channel and narrow-channel effects,[5] respectively, and ξ models drain-induced barrier lowering.

The mobility equation involves both vertical and lateral field effects and is expressed as

$$\mu_1 = \frac{\mu_{eff}}{1 + \dfrac{\mu_{eff}V_{DS}}{v_{max}L_1}} \tag{17.45}$$

where

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_{TH})} \tag{17.46}$$

and v_{max} denotes the maximum velocity of the carriers in the channel. As can be seen from (17.45) and (17.46), μ_{eff} models the effect of the vertical field while μ_1 adds that of the lateral field as well.

The drain current is realized as

$$I_D = \mu_1 C_{ox}\frac{W_{eff}}{L_{eff}}\left[V_{GS} - V_{TH0} - \left(1 + \frac{F_s\gamma}{4\sqrt{2\phi_F - V_{BS}}} + F_n\right)\frac{V'_{DS}}{2}\right]V'_{DS} \tag{17.47}$$

where $V'_{DS} = V_{D,sat}$ if the device is in saturation. The quantity $V_{D,sat}$ represents both channel pinch-off and velocity saturation (Fig. 17.10) and is expressed by relatively complex equations [1].

The subthreshold current relations are similar to those of the Level 2 model, still suffering from derivative discontinuity near strong inversion.

The Level 3 model employs more sophisticated methods of computing channel-length modulation as well as charge and capacitance parameters. The details can be found in [1]. Comparison with measured data [1] suggests that the Level 3 model, as with the Level 2 model, exhibits moderate accuracy for wide, short transistors, but suffers from large errors for longer channels.

An important drawback of the Level 3 model is the discontinuity of the derivative of I_D with respect to V_{DS} at the edge of the triode region, leading to large errors in the calculation of the output impedance. Shown in Fig. 17.15 for a short-channel device, the variation of r_O with V_{DS} is quite poorly modeled.

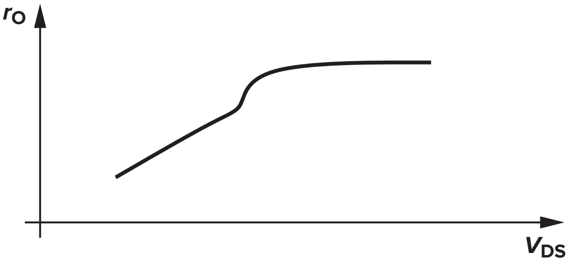


Figure 17.15 Kink in output resistance in Level 3 model.

[5] For narrow-channel devices, the threshold voltage *increases* if the *width* is reduced [6].

17.3.4 BSIM Series

The philosophy behind the Level 1–3 models was to express the device behavior by means of equations that originated from the physical operation. However, as transistors were scaled to submicron dimensions, it became increasingly more difficult to introduce physically meaningful equations that would be both accurate and computationally efficient. BSIM adopted a different approach: numerous empirical parameters were added so as to simplify the equations—but at the cost of losing touch with the actual device operation.

An interesting feature of BSIM is the addition of a simple equation to represent the geometry dependence of many of the device parameters. The general expression is of the form

$$P = P_0 + \frac{\alpha_P}{L_{eff}} + \frac{\beta_P}{W_{eff}} \tag{17.48}$$

where P_0 is the value of the parameter for a long, wide transistor ($P = P_0$ if $L_{eff}, W_{eff} \to \infty$), and α_P and β_P are fitting factors. For example, the mobility is computed as

$$\mu = \mu_0 + \frac{\alpha_\mu}{L_{eff}} + \frac{\beta_\mu}{W_{eff}} \tag{17.49}$$

The formulation of (17.48) nonetheless becomes less accurate at small dimensions [1].

The device equations and fitting parameters used in BSIM are beyond the scope of this book. Using approximately 50 parameters, this model provides the following improvements over the Level 3 version [1]: (1) the dependence of mobility upon the vertical field includes the substrate voltage; (2) the threshold voltage is modified for substrates with nonuniform doping; (3) the currents in the weak and strong inversion regions are derived such that their values and first derivatives are continuous; and (4) to simplify the drain current equations, new expressions are devised for velocity saturation, dependence of mobility upon the lateral field, and the saturation voltage.

Measured results in a 0.7-μm technology [1] indicate that BSIM avoids gross errors in the I/V characteristics for various device dimensions, but its accuracy for narrow, short transistors is somewhat poor.

In addition to shortcomings at channel lengths below approximately 0.8 μm, BSIM suffers from other subtle inaccuracies. For example, at large drain-source voltages, BSIM predicts a *negative* output resistance for saturated MOSFETs. Furthermore, in the deep triode region, BSIM still exhibits slight discontinuities in the drain current [1].

The next model in the BSIM series is BSIM2. Requiring approximately 70 parameters, this version employs new expressions for mobility, drain current, and subthreshold conduction. It also represents the output impedance more accurately by incorporating both channel-length modulation and drain-induced barrier lowering. Nevertheless, measured results indicate that the overall accuracy of the model is only marginally higher than that of BSIM. For short, narrow transistors, BSIM2 suffers from large errors in the triode region and even substantial “kinks” in the saturation region [1].

The trend in BSIM and BSIM2, namely, expressing the device behavior by means of empirical equations that bear little relation to the physical phenomena, eventually created difficulties in modeling short-channel devices. Parameter extraction, modeling process variations, and the need for extensive use of polynomials made the generation and application of these models quite difficult. Consequently, the next generation, BSIM3, has returned to the physical principles of device operation while maintaining many of the useful features of BSIM and BSIM2. BSIM3 itself has rapidly gone through several versions, requiring approximately 180 parameters in the third one. For channel lengths as low as 0.25 μm, BSIM3 provides reasonable accuracy for subthreshold and strong inversion operation while still suffering from large errors in predicting the output impedance. BSIM4 has overcome many of these issues and serves modeling needs in 40-nm and 28-nm generations.

17.3.5 Other Models

In addition to the Level 1–3 models and the four generations of BSIM, a number of other MOS models have been introduced. Among these, HSPICE Level 28, MOS9, and the Enz-Krummenacher-Vittoz (EKV) model are the most notable, for they provide new approaches to representing the behavior of MOSFETs [1]. For example, the HSPICE Level 28 model improves the dependence of accuracy upon device dimensions by expressing the parameters as

$$P = P_0 + \alpha\left(\frac{1}{L} - \frac{1}{L_{ref}}\right) + \beta\left(\frac{1}{W} - \frac{1}{W_{ref}}\right) + \gamma\left(\frac{1}{L} - \frac{1}{L_{ref}}\right)\left(\frac{1}{W} - \frac{1}{W_{ref}}\right) \tag{17.50}$$

where L_{ref} and W_{ref} denote the dimensions of a "reference" device, i.e., a transistor whose characteristics have been measured. Thus, the dependence is expressed in terms of *increments* with respect to characterized transistors rather than the absolute value of the dimensions, yielding a potentially higher accuracy. Also, the term proportional to the product of the length and width increments facilitates curve fitting.

The EKV model [11] substantially departs from traditional views of MOSFET operation by considering the *bulk*, rather than the source, as the reference point for all voltages. This approach thus avoids distinguishing between the source and drain terminals and, more important, introduces a single drain-source current equation that is valid for both subthreshold and saturation regions.

The reader is referred to [1] for an extensive study of these models.

17.3.6 Charge and Capacitance Modeling

The simple gate capacitance model described in Chapter 2 for the Level 1 model, called the Meyer capacitance model [1], suffers from many shortcomings even for long-channel devices. In transient SPICE analyses, such a model does not conserve charge (!), thereby introducing errors in the simulation. For example, as illustrated in Fig. 17.16, a periodic rectangular waveform applied to a voltage divider consisting of an ideal capacitor and a MOSFET experiences "droop" at the output because in every period, some charge at node X is lost. This effect arises from the calculation of charge by integrating capacitor voltages with respect to time, an operation that accumulates small errors in the simulation.[6] To minimize this type of error, the simulation algorithm can be modified such that it first computes the charge in the inversion layer and the depletion region and subsequently partitions the charge among the device capacitances.

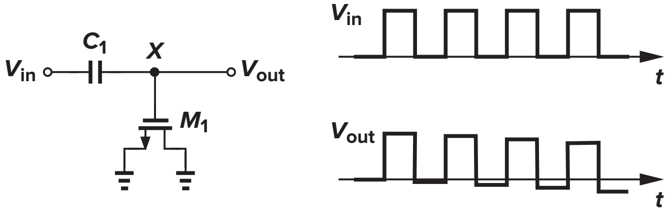


Figure 17.16 Annihilation of charge in simulation.

Another issue in the Meyer charge model relates to partitioning of the channel charge between the source and drain terminals. The assumption that in the triode region, $C_{GS} = C_{GD} = (1/2)WLC_{ox} + WC_{ov}$, and in the saturation region, $C_{GS} = (2/3)WLC_{ox} + WC_{ov}$ and $C_{GD} = WC_{ov}$ is inaccurate for

[6] Another source of error here is the assumption that the device capacitances are reciprocal, e.g., $C_{GS} = C_{SG}$ [1].

short-channel devices, requiring flexible partitioning for ease of curve fitting. In BSIM and BSIM3, for example, three different charge partitioning scenarios (40%/60%, 50%/50%, and 0%/100%) are available.

Recent efforts have created more sophisticated charge and capacitance models for MOS devices so as to improve the accuracy, especially for analog applications. However, as with many other modeling improvements, the resulting equations are quite cumbersome, imparting little intuition. The reader is referred to [1] for details.

17.3.7 Temperature Dependence

Many parameters of MOS transistors vary with temperature, making it difficult to maintain a reasonable fit between measured and simulated behavior across a wide temperature range. In the Level 1–3 models as well as BSIM and BSIM2, the following parameters have temperature dependence: V_{TH}, built-in potential of S/D junctions, the intrinsic carrier concentration of silicon (n_i), the bandgap energy (E_g), and the mobility. Most equations are empirical, e.g.,

$$E_g = 1.16 - \frac{7.02 \times 10^{-4} T^2}{T + 1108} \tag{17.51}$$

and

$$\mu = \mu_0 \left(\frac{300}{T} \right)^{3/2} \tag{17.52}$$

where $\mu_0 = \mu(T = 300\ \text{K})$.

BSIM3 incorporates a few more parameters to represent the temperature dependence of phenomena such as velocity saturation and the effect of subthreshold voltage on V_{TH}. It is unclear at this point how accurately BSIM3 expresses the temperature variation of MOS devices and circuits.

17.4 ■ Process Corners

Unlike bipolar transistors, MOSFETs suffer from substantial parameter variations from wafer to wafer and from lot to lot. Despite decades of technology advancement, the large variability of CMOS circuits remains a fact with which digital and analog designers must cope.

In order to facilitate the task of circuit design to some extent, process engineers guarantee a performance envelope for the devices, in essence tightening the anticipated parameter variations by discarding wafers that fall out of the envelope (Fig. 17.17). Of course, in their eternal battle, circuit designers insist on a tighter variability space so that they can design more aggressively, whereas process engineers tend

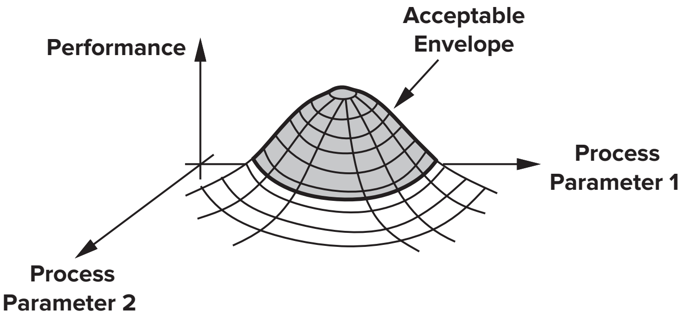


Figure 17.17 Performance envelope as a function of process parameters.

to enlarge the envelope as much as possible so as to increase the yield. For example, it is common in today's CMOS technologies to obtain a gate delay that varies by a factor of two to one with process and temperature.

The performance envelope furnished to designers has traditionally been one suited to digital circuits and constructed in the form of "process corners." Illustrated in Fig. 17.18, the idea is to constrain the speed envelope of the NMOS and PMOS transistors to a rectangle defined by four corners: fast NFET and fast PFET; slow NFET and slow PFET; fast NFET and slow PFET; and slow NFET and fast PFET. For example, transistors having a thinner gate oxide and lower threshold voltage fall near the fast corner. The device models corresponding to each corner are extracted from wafers whose NMOS or PMOS test structures display a large or small gate delay, and the actual corners are chosen so as to obtain an acceptable yield. Thus, only wafers satisfying these specifications are considered acceptable. Simulation of circuits for various process corners and temperature extremes is essential to determining the yield.

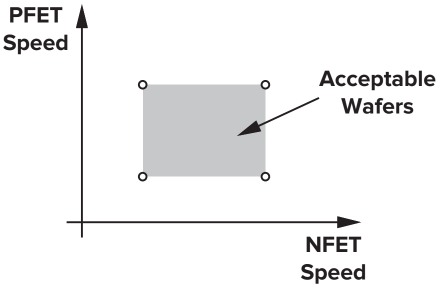


Figure 17.18 Process corners based on speed of NMOS and PMOS devices.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

17.1. Silicon dioxide breaks down at high electric fields. Explain what happens if ideal scaling is performed while keeping the gate oxide thickness constant.

17.2. The maximum doping level that can be established in the source and drain regions is limited by the "solid solubility" of silicon. Explain what happens to the S/D junction capacitance and series resistance as ideal scaling occurs, but the S/D doping level remains constant. Does DIBL become more or less significant?

17.3. Suppose the supply voltage of a switched-capacitor amplifier is reduced by a factor of two and so is the maximum allowable output voltage swing. In order to maintain the dynamic range constant, the noise voltage must scale down by the same factor.

(a) If the noise is only of kT/C type, how should the capacitors in the circuit be scaled?

(b) If the time constant is given by G_m/C, where G_m denotes the transconductance of a one-stage op amp, how should G_m be scaled to maintain the same small-signal time constant?

(c) How should the dimensions and tail current of the input differential pair of the op amp be scaled?

(d) Repeat parts (b) and (c) where the slew rate must remain constant.

17.4. Explain how each parameter in Eq. (17.16) scales in an ideal constant-field scaling scenario. What happens to the subthreshold slope?

17.5. A common-gate stage designed for an input impedance of 50 Ω undergoes ideal scaling. If $\lambda = \gamma = 0$, what is the input impedance?

17.6. Repeat Problem 17.5 if $\lambda \neq 0$, $\gamma \neq 0$, and the load is a MOS current source that is also scaled.

17.7. For power-conscious applications, a figure of merit is defined as the transconductance of devices normalized to their bias current. Determine this quantity for long-channel devices operating in strong inversion or the subthreshold region. At what drain current are these two equal?

17.8. Explain why the mobile charge density cannot drop to exactly zero at any point along the channel. What happens beyond the pinch-off point?

17.9. Using Eq. (17.21), calculate the transconductance of a MOSFET. What happens if the overdrive voltage is very small or very large?

17.10. Using Eq. (17.30), calculate the transconductance of a MOSFET. Prove that

$$g_m = \frac{I_D}{V_{GS} - V_{TH}} \left[1 + \frac{1}{1 + \left(\dfrac{\mu_0}{2v_{sat}L} + \theta \right)(V_{GS} - V_{TH})} \right] \tag{17.53}$$

17.11. Suppose the channel-length modulation coefficient λ is modified as $\lambda/(1 + \kappa V_{DS})$, where κ is a constant, to represent the dependence of the output impedance upon V_{DS}. Calculate r_O. Explain how a current source with such behavior introduces distortion in the voltage across it.

17.12. Assuming that the devices in Fig. 17.19 experience complete velocity saturation, derive expressions for the voltage gain of each circuit in terms of W and v_{sat}. Assume that $\lambda = \gamma = 0$.

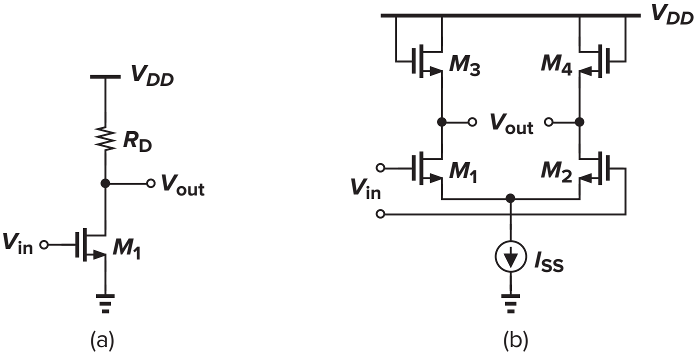


Figure 17.19

17.13. Using Eq. (17.37), calculate g_{mb} and compare the result with that derived in Chapter 2.

17.14. From Eq. (17.51), determine $\partial E_g/\partial T$ at room temperature and explain how it affects bandgap reference voltages.

17.15. Suppose the fast corners of a process result from a higher μC_{ox}. Explain what happens to the voltage gain and the input thermal noise of the circuits shown in Fig. 17.20 at the four corners of the process if the transistors are biased at a constant current in saturation.

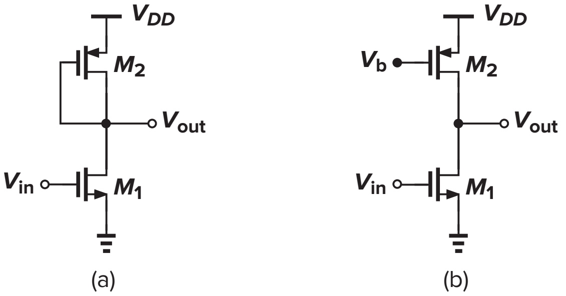


Figure 17.20

17.16. Repeat Problem 17.15 if each transistor is biased with a fixed V_{GS}.

References

[1] D. P. Foty, *MOSFET Modeling with SPICE* (Upper Saddle River, NJ: Prentice-Hall, 1997).

[2] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. (Boston: McGraw-Hill, 1999).

[3] P. Antognetti and G. Massobrio, eds., *Semiconductor Device Modeling with SPICE* (New York: McGraw-Hill, 1988).

[4] R. H. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. of Solid-State Circuits*, vol. 9, pp. 256–268, October 1974.

[5] G. E. Moore, "Progress in Digital Integrated Circuits," *IEDM Tech. Dig.*, pp. 11–14, December 1975.

[6] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices* (New York: Cambridge University Press, 1998).

[7] C. G. Sodini, P. K. Ko, and J. L. Moll, "The Effect of High Fields on MOS Device and Circuit Performance," *IEEE Tran. on Electron Devices*, vol. 31, pp. 1386–1393, October 1984.

[8] P. K. Ko, "Approaches to Scaling," in N. G. Einspruch and G. Gildenblat, eds., *Advanced MOS Device Physics*, pp. 1–35 (San Diego: Academic Press, 1998).

[9] S. Wong and A. T. Salama, "Impact of Scaling on MOS Analog Performance," *IEEE J. of Solid-State Circuits*, vol. 18, pp. 106–114, February 1983.

[10] H. Shichman and D. A. Hodges, "Modeling and Simulation of Insulated Field Effect Transistor Switching Circuits," *IEEE J. of Solid-State Circuits*, vol. 3, pp. 285–289, September 1968.

[11] C. C. Enz, F. Krummenacher, and E. Vittoz, "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low Voltage and Low Current Applications," *Analog Integrated Circuits and Signal Processing*, vol. 8, pp. 83–114, 1995.

[12] Y. Tsividis and K. Suyama, "MOSFET Modeling for Analog Circuit CAD: Problems and Prospects," *IEEE J. of Solid-State Circuits*, vol. 29, pp. 210–216, March 1994.

[13] B. Razavi, "CMOS Technology Characterization for Analog and RF Design," *IEEE J. of Solid-State Circuits*, vol. 34, pp. 268–276, March 1999.

CHAPTER

18

CMOS Processing Technology

With the high-order effects of MOS devices covered in Chapter 17, we now study the fabrication of CMOS technologies. A solid understanding of device processing proves essential in the design and layout of ICs because many limitations imposed on the performance of circuits are related to fabrication issues. Furthermore, today's semiconductor technology demands that process engineers and circuit designers interact regularly so as to understand each other's needs, necessitating a good knowledge of each discipline.

In this chapter, we deal with the processing technology of CMOS devices, aiming to provide a simple view of the fabrication steps and their relevance to circuit design and layout. We begin with a brief description of basic fabrication steps, such as wafer processing, photolithography, oxidation, ion implantation, deposition, and etching. Next, we study the fabrication sequence of MOS transistors in detail. Finally, we describe the processing of passive devices and interconnections.

18.1 ■ General Considerations

Before delving into a detailed study of fabrication, it is instructive to consider the basic structure of NMOS and PMOS transistors and predict the required processing steps. As shown in Fig. 18.1, a p-type substrate (wafer) serves as the foundation upon which n-wells, source/drain regions, gate dielectric, polysilicon, n-well and substrate ties, and metal interconnects are built. Considering both the side view and the top view, we may raise the following questions: (1) How are various regions defined so accurately? For example, how is a gate polysilicon line with a minimum dimension of 0.25 μm fabricated while maintaining a distance of 0.25 μm from another polysilicon line? (2) How are the n-wells and S/D regions built? (3) How are the gate oxide and polysilicon fabricated? (4) How are the gate oxide and polysilicon *aligned* with the S/D regions? (5) How are the contact windows created? (6) How are the metal interconnect layers deposited?

Modern CMOS technologies involve more than 200 processing steps, but for our purposes, we can view the sequence as a combination of the following operations: (1) wafer processing to produce the proper type of substrate; (2) photolithography to precisely define each region; (3) oxidation, deposition, and ion implantation to *add* materials to the wafer; and (4) etching to *remove* materials from the wafer. Many of these steps require "heat treatment," i.e., the wafer must undergo a thermal cycle inside a furnace.

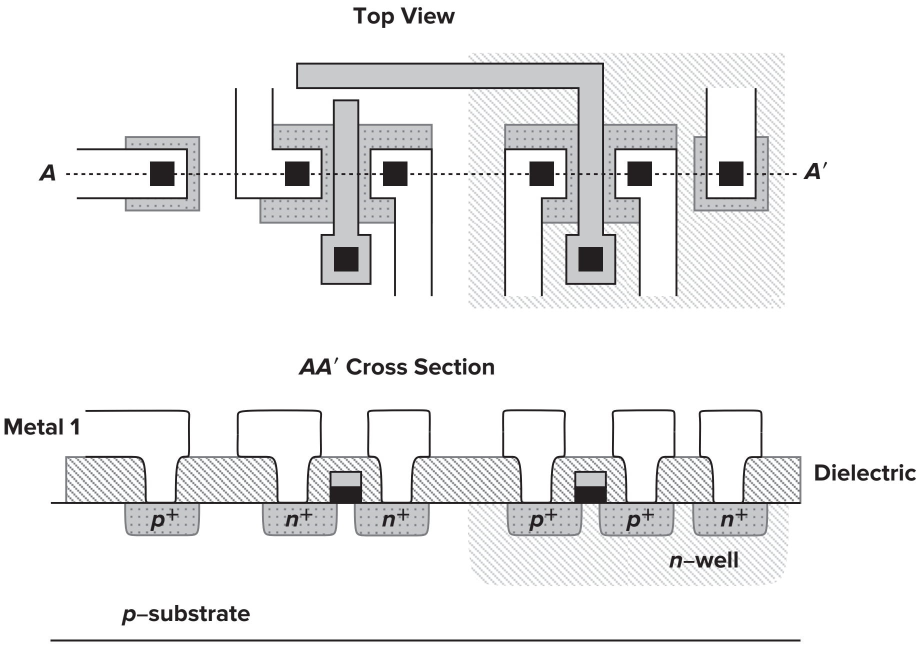


Figure 18.1 Side view and top view of MOS devices.

In semiconductor processing and characterization, we often refer to the "sheet resistance" of a layer. The total resistance of a rectangular bar is $R = \rho L/(W \cdot t)$, where ρ is the resistivity of the material and L, W, and t denote the length, width, and thickness of the bar, respectively. In integrated circuits, the resistivity and thickness of the layers are set by fabrication materials and processing steps and cannot be changed in the layout. The quantity $R_\square = \rho/t$ is thus defined as the sheet resistance, combining two constants of the technology. Since $R = R_\square$ for $W = L$, i.e., for a square geometry, we express $R_\square$ in terms of ohms per square. For example, for a sheet resistance of 10 $\Omega/\square$, a geometry with $W = 2$ μm and $L = 20$ μm has a resistance of $R = 10\ \Omega/\square \times (20/2) = 100\ \Omega$. In fact, we may say "this line is 10 squares long," meaning that $L/W = 10$ and $R = 10R_\square$.

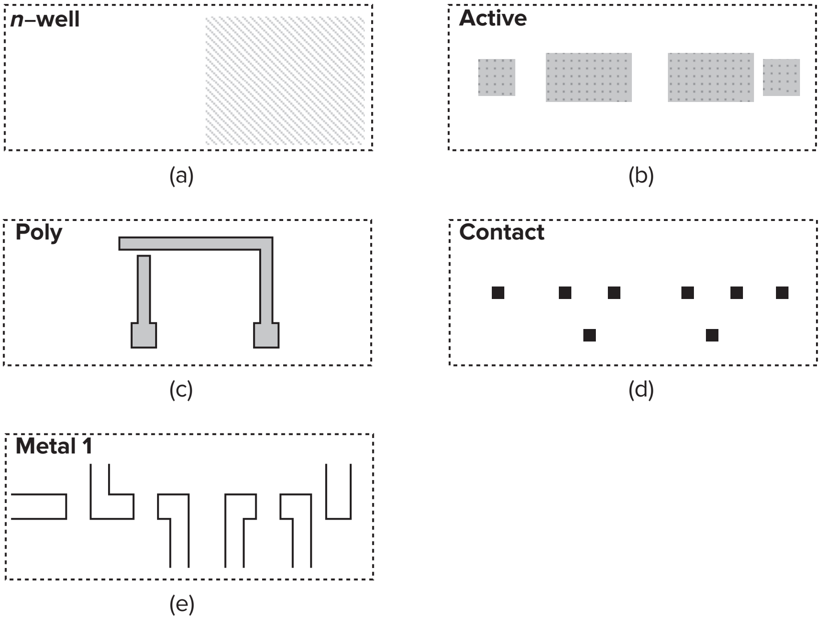
18.2 ■ Wafer Processing

The starting wafer in a CMOS technology must be created with a very high quality. That is, the wafer must be grown as a single-crystal silicon body having a very small number of "defects," e.g., dislocations in the crystal or unwanted impurities. Furthermore, the wafer must contain the proper type and level of doping so as to achieve the required resistivity.

This is accomplished by the "Czochralski method," whereby a seed of crystalline silicon is immersed in molten silicon and gradually pulled out while rotating. As a result, a large single-crystal cylindrical "ingot" is formed that can be sliced thin into wafers. The diameter of the wafer has scaled up with new technology generations, exceeding 30 cm (12 in) today. Note that dopants are added to the molten silicon to obtain the desired resistivity. The wafers are then polished and chemically etched, thereby removing damages on the surface that are created during slicing. In most CMOS technologies, the wafer has a resistivity of 0.05 to 0.1 Ω·cm and a thickness of approximately 500 to 1,000 μm (which is reduced to a few hundred microns after all of the processing steps).

18.3 ■ Photolithography

Photolithography, or simply lithography, is the first step in transferring the circuit layout information to the wafer. As shown in the top view of Fig. 18.1 and explained in Chapter 19 in more detail, the layout consists of polygons representing different types of "layers," e.g., n-well, S/D regions, polysilicon, contact windows, etc. For fabrication purposes, we decompose the layout into these layers. For example, the layout of Fig. 18.1 can be viewed as the five different layers shown in Fig. 18.2, each of which must be created on the wafer with a very high precision. Note that the "active" (or "diffusion") layer includes the source/drain regions and the p^+ and n^+ openings serving as the substrate and well ties.



(a)

(b)

(c)

(d)

(e)

Figure 18.2 Layers comprising the structures of Fig. 18.1.

To understand how a layer is transferred from the layout to the wafer, let us consider the n-well pattern of Fig. 18.2(a) as an example. This pattern is "written" to a transparent glass "mask" by a precisely controlled electron beam [Fig. 18.3(a)]. Also, as depicted in Fig. 18.3(b), the wafer is covered by a thin layer of "photoresist," a material whose etching properties change upon exposure to light.[1] Subsequently, the mask is placed on top of the wafer and the pattern is projected onto the wafer by ultraviolet (UV) light [Fig. 18.3(c)]. The photoresist "hardens" in the regions exposed to light and remains "soft" under the opaque rectangle. The wafer is then placed in an etchant that dissolves the "soft" photoresist area, thereby exposing the silicon surface [Fig. 18.3(d)]. Now, an n-well can be created in the exposed area. We call this set of operations a lithography sequence.

In summary, the sequence associated with the lithography of each layer involves one mask and three processing steps: (1) cover wafer with photoresist; (2) align mask on top and expose to light; (3) etch exposed photoresist. The example of Fig. 18.2 therefore requires at least five masks and hence five lithography sequences.

[1] In practice, a thin layer of oxide is grown before depositing the photoresist to protect the surface.

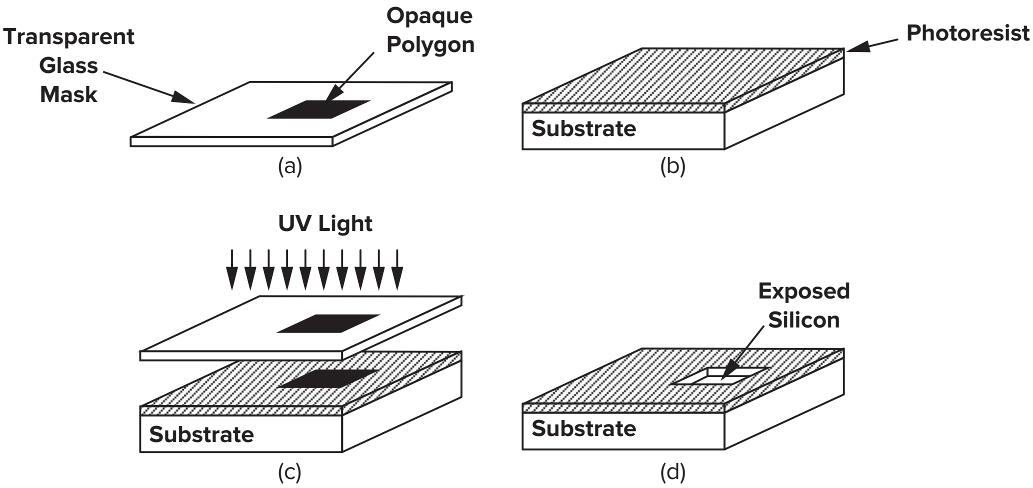


Figure 18.3 (a) Glass mask used in lithography; (b) coverage of wafer by photoresist; (c) selective exposure of photoresist to UV light; (d) exposed silicon after etching.

We should mention that two types of photoresists are used in processing. A "negative" photoresist hardens in the areas exposed to light, and a "positive" photoresist hardens in the areas not exposed to light. As explained later in this chapter, both types prove useful in fabrication.

The number of masks in a process heavily affects the overall cost of fabrication, eventually influencing the unit price of the chip. This is so for two reasons: each mask costs tens of thousands of dollars, and, owing to the necessary precision, lithography is a slow and expensive task. In fact, CMOS technology originally became attractive by virtue of the relatively small number of masks—about seven—that it required. Although in modern CMOS processes this number is around 30, the cost of each IC has nonetheless remained low because both the number of transistors per unit area and the size of the wafer have steadily increased.

18.4 ■ Oxidation

A unique property of silicon is that it can produce a very uniform oxide layer on the surface with little strain in the lattice, allowing the fabrication of gate-oxide layers as thin as a few tens of angstroms (only several *atomic* layers). In addition to serving as the gate dielectric, silicon dioxide can act as a protective coating in many steps of fabrication. Also, in areas between the devices, a thick layer of SiO_2, called the "field oxide" (FOX) is grown, providing the foundation for interconnect lines that are formed in subsequent steps (Fig. 18.4).

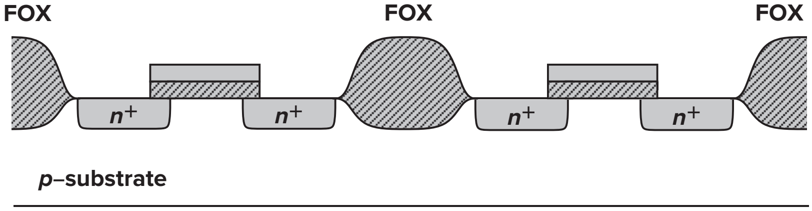


Figure 18.4 Field oxide.

Silicon dioxide is “grown” by placing the exposed silicon in an oxidizing atmosphere such as oxygen at a temperature around $1,000°\text{C}$. The rate of growth depends on the type and pressure of the atmosphere, the temperature, and the doping level of the silicon.

The growth of the gate oxide is a very critical step in the process. Since the oxide thickness, t_{ox}, determines both the current handling and the reliability of the transistors, it must be controlled to within a few percent. For example, the oxide thicknesses of two transistors separated by 20 cm on a wafer must differ by less than a few angstroms, requiring extremely high uniformity across the wafer and hence a slow growth of the oxide. Also, the “cleanness” of the silicon surface under the oxide affects the mobility of the charge carriers and thus the current drive, transconductance, and noise of the transistors.

18.5 ■ Ion Implantation

In many fabrication steps, dopants must be selectively introduced into the wafer. For example, after the lithography sequence of Fig. 18.3 is completed, the n-well is formed by entering dopants into the exposed silicon area. Similarly, the source and drain regions of transistors require selective addition of dopants to the wafer.

The most common method of introducing dopants is “ion implantation,” whereby the doping atoms are accelerated as a high-energy focused beam, hitting the surface of the wafer and penetrating the exposed areas [Fig. 18.5(a)]. The doping level (dosage) is determined by the intensity and duration of the implantation, and the depth of the doped region is set by the energy of the beam. As shown in Fig. 18.5, with a high energy, the peak of the doping concentration in fact occurs well below the surface, thereby creating a “retrograde” profile. Such a profile is desirable for the n-well because it establishes a low resistivity near the bottom, reducing susceptibility to latch-up (Sec. 18.8), and a low doping level at the surface, decreasing the S/D junction capacitance of PMOS devices.

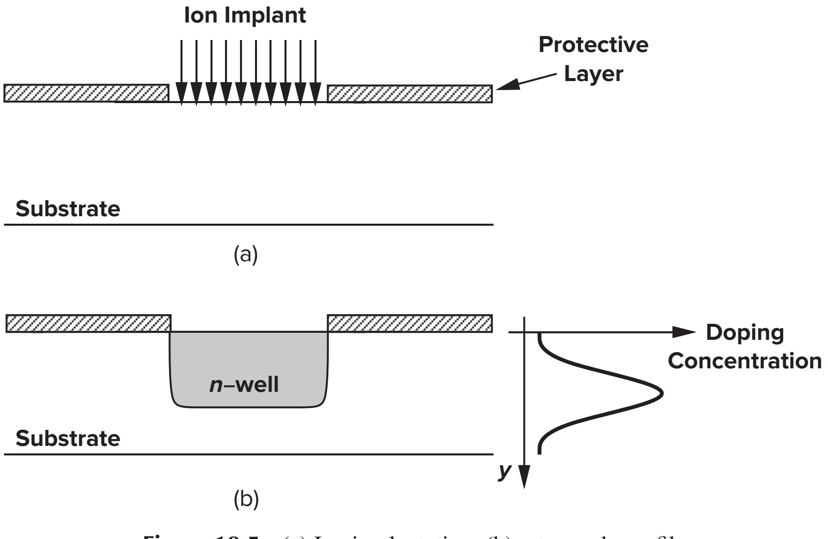


Figure 18.5 (a) Ion implantation; (b) retrograde profile.

Another important application of implantation is to create “channel-stop” regions between transistors. Consider the field oxide and the S/D junctions of M_1 and M_2 in Fig. 18.6(a), assuming that an interconnect line passes on top of the field oxide. Interestingly, the two n^+ regions and the FOX form a MOS transistor having a thick gate oxide and hence a large threshold voltage. Nonetheless, with a sufficiently positive potential on the interconnect line, this transistor may turn on slightly, creating a leakage path between

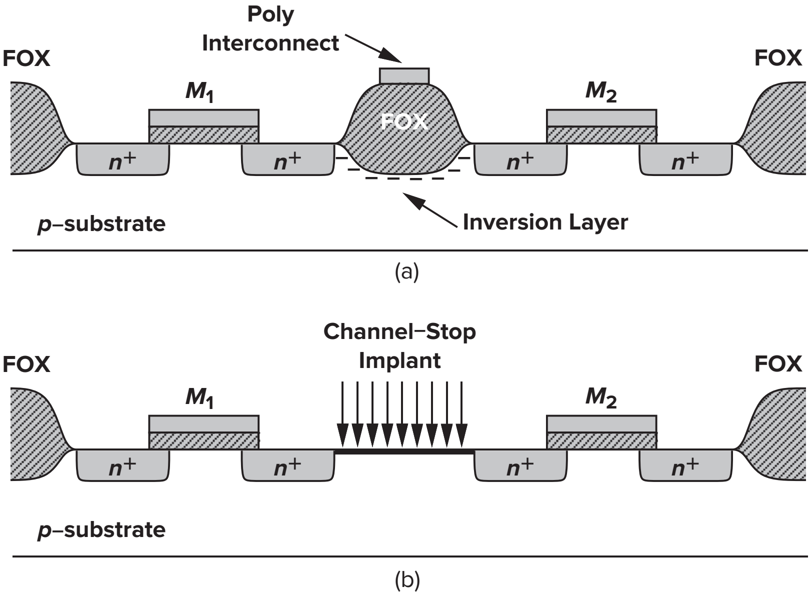


Figure 18.6 (a) Unwanted conduction due to inversion of field area; (b) channel-stop implant.

M_1 and M_2. To resolve this issue, a channel-stop implant (also called a field implant) is performed before the field oxide deposition [Fig. 18.6(b)], thereby raising the threshold voltage of the field oxide transistor to a very large value.

Ion implantation damages the silicon lattice extensively. For this reason, the wafer is subsequently heated to approximately 1, 000°C for 15 to 30 minutes, allowing the lattice bonds to form again. Called "annealing," this operation also leads to diffusion of dopants, broadening the profile in all directions. For example, annealing results in side diffusion of S/D regions, creating overlap with the gate area. The wafer is therefore usually annealed only once, after all implantations have been completed.

An interesting phenomenon in ion implantation is "channeling." As shown in Fig. 18.7(a), if the implant beam is aligned with the crystal axis, the ions penetrate the wafer to a great depth. For this reason, the implant (or the wafer) is tilted by 7–9° [Fig. 18.7(b)], avoiding such an alignment and ensuring a predictable profile. As explained in Chapter 19, this tilt affects the matching of transistors, necessitating precautions in the layout.

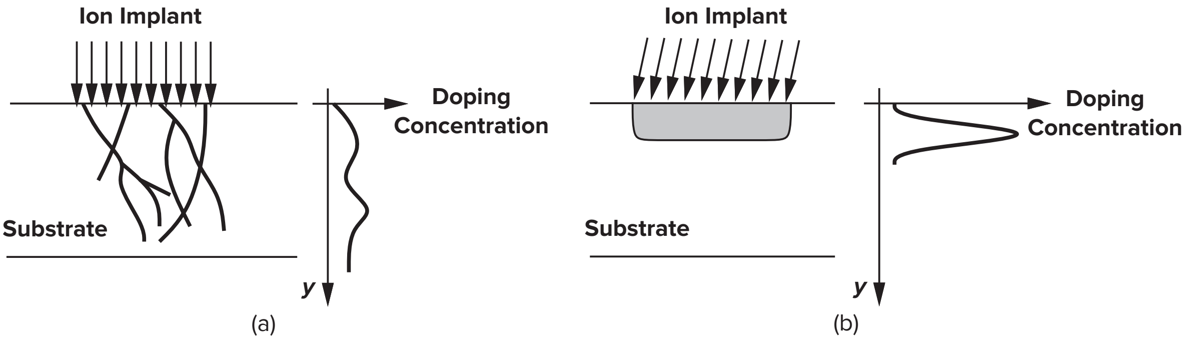


Figure 18.7 (a) Effect of channeling; (b) tilt in implant to avoid channeling.

18.6 ■ Deposition and Etching

As suggested by the structures of Fig. 18.1, device fabrication requires the deposition of various materials. Examples include polysilicon, dielectric materials separating interconnect layers, and metal layers serving as interconnects.

A common method of forming polysilicon on thick dielectric layers is "chemical vapor deposition" (CVD), whereby wafers are placed in a furnace filled with a gas that creates the desired material through a chemical reaction. In modern processes, CVD is performed at a low pressure to achieve more uniformity.

The etching of the materials is also a crucial step. For example, contact windows with very small dimensions, e.g., 0.3 μm × 0.3 μm, and relatively large depths, e.g., 2 μm, must be etched with high precision. Depending on the speed, accuracy, and selectivity required in the etching step, and the type of material to be etched, one of these methods may be used: (1) "wet" etching, i.e., placing the wafer in a chemical liquid (low precision); (2) "plasma" etching, i.e., bombarding the wafer with a plasma gas (high precision); (3) reactive ion etching (RIE), where ions produced in a gas bombard the wafer.

18.7 ■ Device Fabrication

With the processing operations described in the previous section, we now study the fabrication sequence and device structures in typical CMOS technologies. We consider three categories: active devices, passive devices, and interconnects.

18.7.1 Active Devices

Basic Transistor Fabrication The fabrication begins with a p-type silicon wafer approximately 1 mm thick. Following the cleaning and polishing steps, a thin layer of silicon dioxide is grown as a protective coating on top of the wafer [Fig. 18.8(a)]. Next, to create the n-wells, a lithography sequence consisting of photoresist deposition, exposure to UV light using the n-well mask, and selective etching is carried out, and the n-wells are implanted [Fig. 18.8(b)]. The remaining photoresist and oxide layers are then removed [Fig. 18.8(c)].

Recall from the previous section that a field implant and a field oxide growth are necessary in the areas between the transistors. At this point in the sequence, a stack consisting of a silicon oxide layer, a silicon nitride (Si_3N_4) layer, and a *positive* photoresist layer is created. Next, the "active" mask is used for lithography so that only the regions between the transistors are exposed [Fig. 18.8(d)].[2] Subsequently, the channel-stop implant is performed, the photoresist is removed, and a thick oxide layer is grown in the exposed silicon areas, producing the field oxide. The protective nitride and oxide layers are then removed [Fig. 18.8(e)], thereby exposing all areas where transistors are to be formed. In the subsequent diagrams, the channel-stop implant will be omitted for the sake of clarity.

The next step involves the growth of the gate oxide, a critical operation requiring slow, low-pressure CVD [Fig. 18.8(f)]. As explained in Chapter 2, the "native" threshold voltage of the transistors is typically far from the desired value, necessitating a threshold-adjust implant. (The native threshold of both PMOS and NMOS is usually more negative than desired, e.g., $V_{THN} \approx 0$, and $V_{THP} \approx -1$ V.) Such an implant is performed following the growth of the gate oxide, creating a thin sheet of dopants near the surface and making the threshold of both NMOS and PMOS devices more positive than their nominal values.

[2] The n-wells are not shown for clarity.

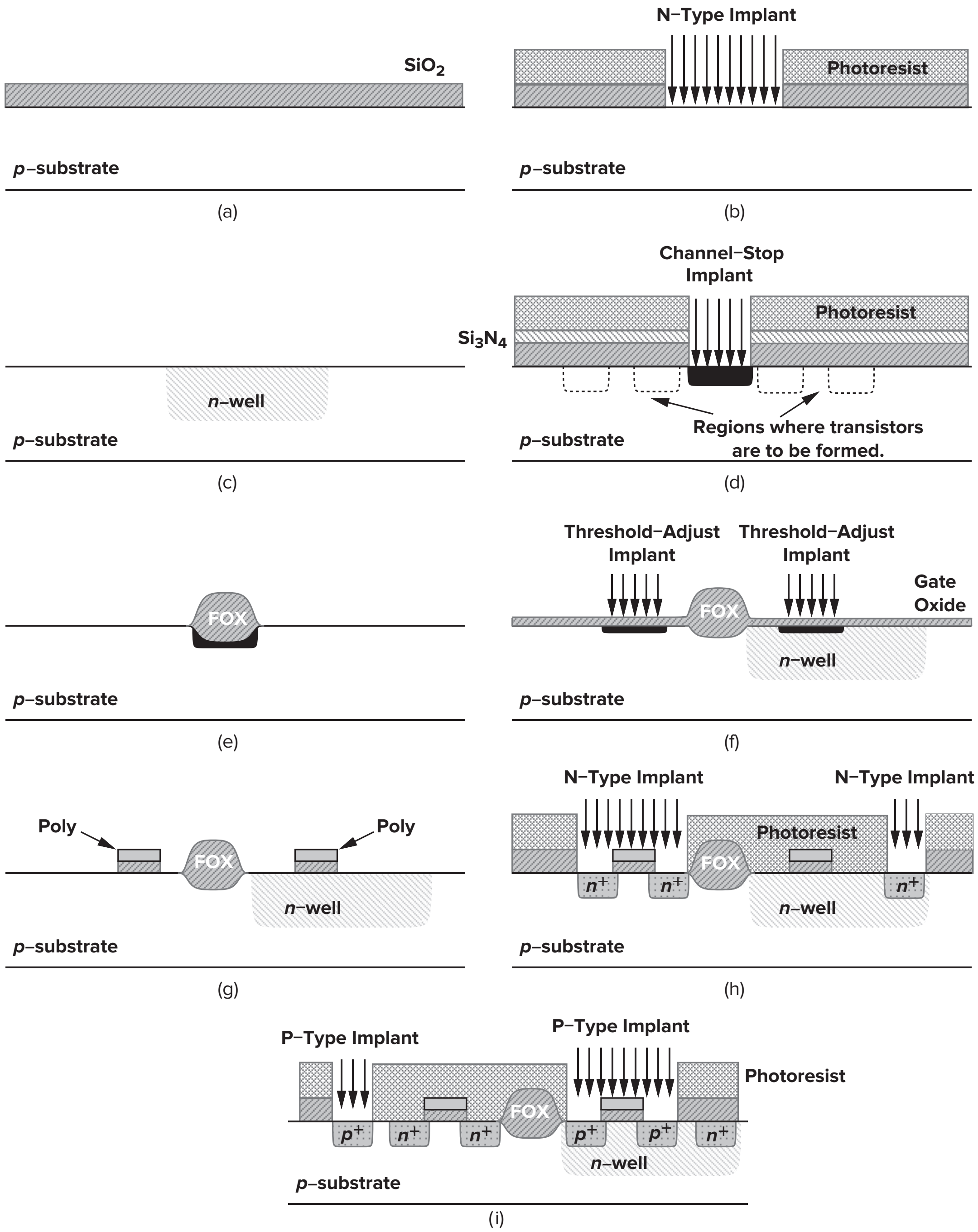


Figure 18.8 Fabrication sequence of MOS devices.

With the gate oxide in place, the polysilicon layer is deposited and the "poly mask" lithography is carried out, resulting in the structure shown in Fig. 18.8(g). We should note that polysilicon is simply noncrystalline ("amorphous") silicon, a property that arises because this layer grows on top of silicon dioxide and hence cannot form a crystal. Since polysilicon serves as a conductor, its amorphous nature is unimportant. To reduce the resistivity of this layer, an additional implant is typically used, yielding a sheet resistance of a few tens of ohms per square.

In the next step, the source/drain junctions of the transistors and the substrate and n-well ties are formed by ion implantation. This step requires a "source/drain mask" and two lithography sequences. As illustrated in Fig. 18.8(h), the first sequence incorporates a negative photoresist, exposing the areas to receive an n^+ implant (the S/D junctions of NMOS transistors and the n-well ties). In the second sequence [Fig. 18.8(i)], the same mask and a positive photoresist are used, exposing the areas to receive a p^+ implant (the S/D junctions of PMOS transistors and the substrate ties). Note that these implants also dope the polysilicon layer, reducing its sheet resistance. This step completes the fabrication of the basic transistors.

The reader may wonder why the source/drain junctions are formed *after* the gate oxide and polysilicon. Suppose, as depicted in Fig. 18.9(a), these junctions were created first. Then, the alignment of the gate poly mask with respect to the S/D areas would become extremely critical. Even if the misalignment were a small fraction of the minimum channel length, a gap might appear between the source (or drain) and the gate area, prohibiting the formation of a continuous channel in the transistor. By contrast, the sequence shown in Fig. 18.8 yields a "self-aligned" structure because the source/drain regions are implanted at precisely the edges of the gate area and a misalignment in lithography simply makes one junction slightly narrower than the other [Fig. 18.9(b)]. Interestingly, the first few generations of CMOS technology were based on the approach shown in Fig. 18.9(a), but it was soon discovered that the self-aligned structure would lend itself to scaling much more easily.

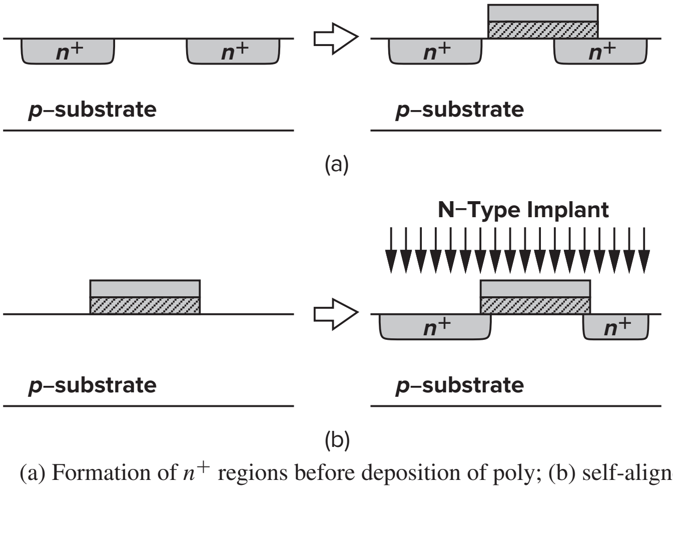


Figure 18.9 (a) Formation of n^+ regions before deposition of poly; (b) self-aligned structure.

Back-End Processing With the basic transistors fabricated, the wafers must next undergo "back-end" processing, a sequence primarily providing various electrical connections on the chip through contacts and wires. The first step in this sequence is "silicidation." Since the sheet resistance of doped polysilicon and S/D regions is typically several tens of ohms per square, it is desirable to reduce their resistance by about an order of magnitude. Silicidation accomplishes this by covering the polysilicon layer and active areas (S/D regions and substrate and n-well ties) with a thin layer of a highly conductive material, e.g., titanium silicide or tungsten. Illustrated in Fig. 18.10, this step in fact begins with creating an "oxide spacer" at

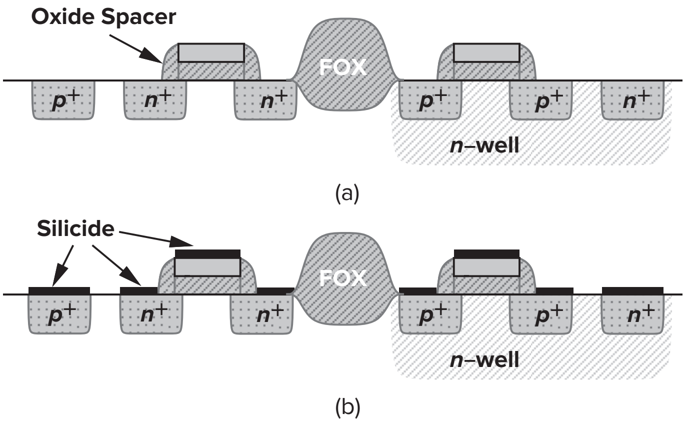


Figure 18.10 (a) Oxide spacers and (b) silicide.

the edges of the polysilicon gate such that the deposition of the silicide becomes a self-aligned process as well.[3] Without the spacer, the silicide layer on the gate may be shorted to that on the source/drain.

The next step in back-end processing is to produce contact windows on top of the polysilicon and active regions. This is carried out by first covering the wafer with a relatively thick (0.3- to 0.5-μm) layer of oxide and subsequently performing a lithography sequence using the "contact mask." The contact holes are then created by plasma etching [Fig. 18.11(a)]. Owing to reliability issues, contacts to the gate polysilicon are not placed on top of the gate area.

Following the contact windows, the first layer of metal interconnect (called "metal 1") (using aluminum or copper) is deposited over the entire wafer. A lithography sequence using the "metal 1 mask" is then carried out, and the metal layer is selectively etched [Fig. 18.11(b)].

The higher levels of interconnect are fabricated using the same procedure [Fig. 18.11(c)]. For each additional metal layer, two masks are required: one for the contact windows and another for the metal itself. Thus, a CMOS process having five layers of metal contains 10 masks for the back end. The contact windows between metal layers are sometimes called "vias" to distinguish them from the first level of contacts to active areas and polysilicon.

We should mention that if a large area must be contacted, many small windows—rather than a large window—are usually used. Dictated by reliability issues, the dimensions of each contact or via are fixed and cannot be decreased or increased by the layout designer. An interesting phenomenon related to large active areas is "contact spiking." If a large contact window allows aluminum to touch the active area, then, as depicted in Fig. 18.12(a), the metal may "eat" and penetrate the doped region, eventually crossing the junction to the bulk and shorting the diode. With small windows, on the other hand, this effect is avoided [Fig. 18.12(b)].

The final step in back-end processing is to cover the wafer with a "glass" or "passivation" layer, protecting the surface against damages caused by subsequent mechanical handling and dicing. After a lithography sequence using the "passivation mask," the glass is opened only on top of the bond pads to allow connection to the external environment (e.g., the package).

18.7.2 Passive Devices

Passive components such as resistors and capacitors find wide usage in analog design, making it desirable to add these devices to standard CMOS technologies. In practice, however, CMOS processes target

[3] Self-aligned silicide is sometimes called "salicide."

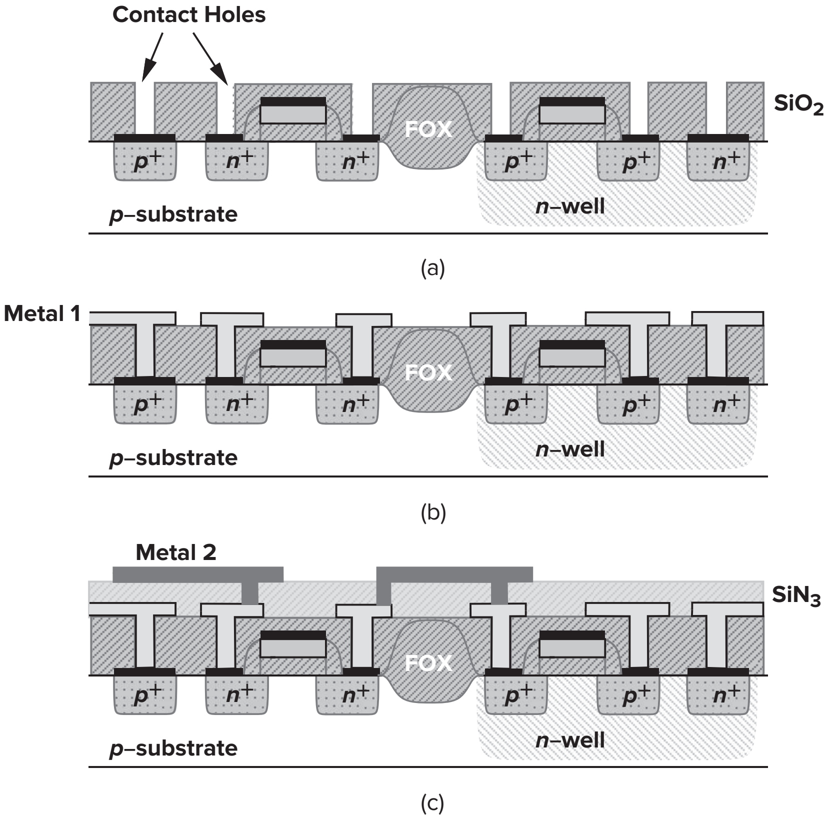


Figure 18.11 Contact and metal fabrication.

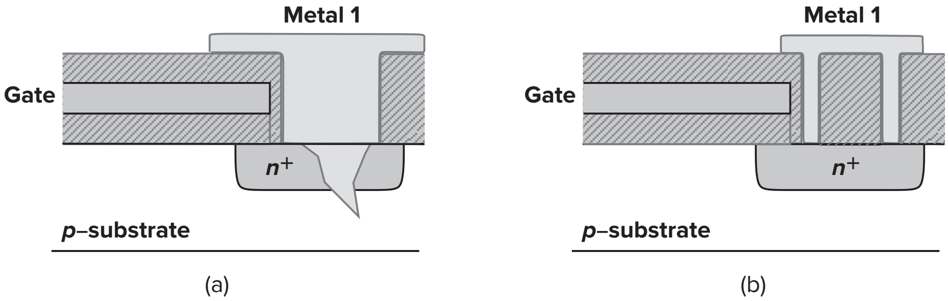


Figure 18.12 (a) Spiking due to large contact areas; (b) use of small contacts to avoid spiking.

primarily digital applications and hence provide only NMOS and PMOS transistors. A new generation of CMOS technology may take one to two years and many iterations before it becomes an "analog process," i.e., one offering high-quality passive devices. If a digital CMOS process is to be used for analog design, we must seek structures that can serve as passive components. The principal issue in using such structures is the *variability* of the component value from wafer to wafer because the process flow does not assume such structures are used in circuits.

Resistors A CMOS process may be modified so as to provide resistors suited to analog design. A common method is to selectively "block" the silicide layer that is deposited on top of the polysilicon, thereby creating a region having the resistivity of the doped polysilicon (Fig. 18.13). This means that the fabrication requires an additional mask and a corresponding lithography sequence. Since the poly doping level is determined by various implants in the process, the resistivity obtained here is not necessarily a target value, but it usually falls in the range of fifty to a few hundred ohms per square. For the same reason, the resistance value may vary by as much as ±20% from wafer to wafer or lot to lot.

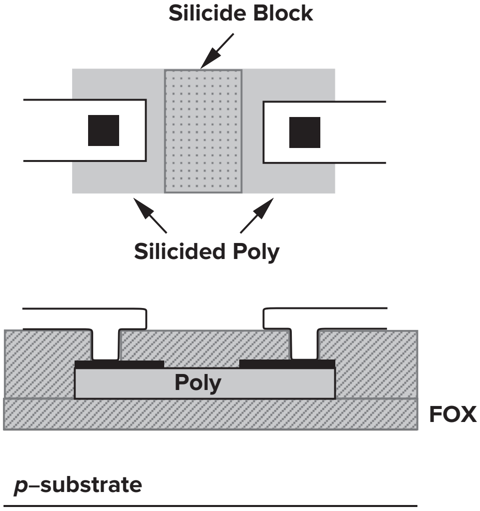


Figure 18.13 Poly resistor using silicide block.

The use of silicide on the two ends of the resistor in Fig. 18.13 results in a much lower contact resistance than that obtained by directly connecting the metal layer to doped polysilicon. This improves both the definition of the resistor value and the matching with identical structures. Also, for a given resistance, poly resistors typically exhibit much less capacitance to the substrate than other types—on the order of 90 aF/μm^2 for the bottom plate capacitance and 100 aF/μm for the fringing capacitance. These resistors are quite linear, especially if they are long. The primary difficulties with silicide-block poly resistors are variability, mask cost, and process complexity.

In a purely digital process, silicided poly, silicided p^+ or n^+ active areas, n-well, and metal layers can be used as resistors. An n-well resistor can be formed as shown in Fig. 18.14, but the n-well resistivity may vary by several tens of percent with process. With typical sheet resistivities of about 1 kΩ/□, n-well resistors can prove useful where their absolute value is not critical. For example, Fig. 18.15 shows a common-source stage that is biased by means of M_0 and I_0 while employing C_1 to block the dc level of the preceding stage. In order to isolate the signal path from the low impedance (and the noise) introduced by M_0, resistor R_1 is inserted between X and Y. Here, the value of R_1 is not critical so long as it is sufficiently large.

We should mention that, due to the depletion region formed between the n-well and the p-substrate, n-well resistors suffer from both a large parasitic capacitance and significant voltage dependence. Figure 18.16 illustrates a typical case, where one terminal of the n-well resistor is tied to V_{DD}. Since the capacitance to the substrate is distributed (nonuniformly) along the resistor, a lumped model may not be accurate enough, but as a rough approximation, we place half of the total capacitance on each side of the resistor. We also note that as V_{out} varies, so do the width of the depletion region and hence the value of the resistor.

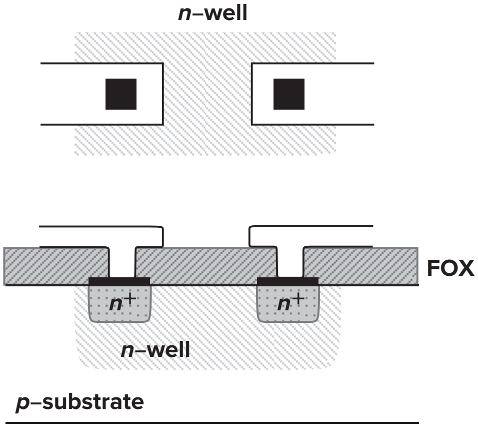


Figure 18.14 Resistor made of *n*-well.

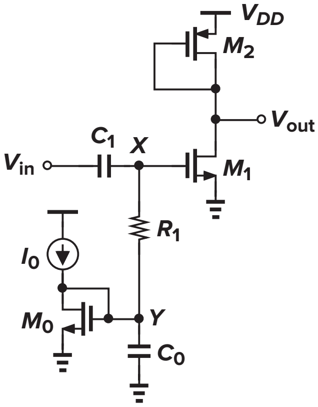


Figure 18.15 Use of an *n*-well resistor in a coupling network.

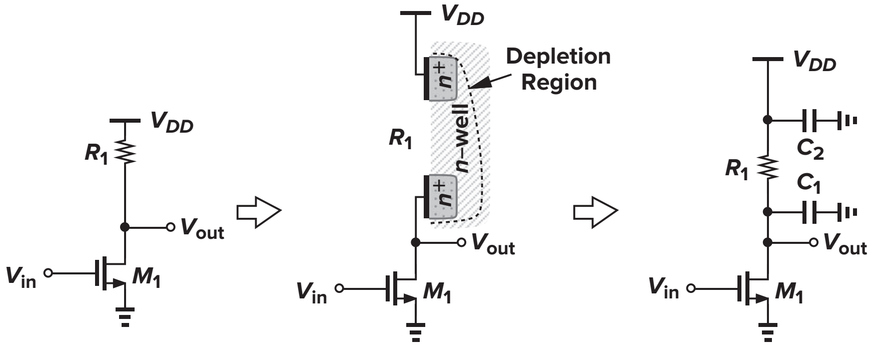


Figure 18.16 Common-source stage using *n*-well resistors.

The metal layers available in CMOS technologies exhibit sheet resistances on the order of 100 mΩ/□ (for bottom layers) to 30 mΩ/□ (for top layers). Thus, for resistor values common in analog design, metal layers are rarely used.

Capacitors Capacitors prove indispensible in most of today's analog CMOS circuits. Several parameters of capacitors are critical in analog design: parasitic capacitance to the substrate, capacitance per unit area (density), and nonlinearity.

Perhaps the simplest capacitor structure in CMOS technology is that implemented by a MOSFET. Illustrated in Fig. 18.17(a), the device has a capacitance that varies from a small value at low voltages (where no channel exists and the equivalent capacitance is the series combination of the oxide capacitance and the depletion region capacitance) to a large value (C_{ox}) if the voltage difference exceeds V_{TH}. Since the gate oxide is typically the thinnest layer in the process, MOS capacitors biased in strong inversion are quite dense, saving substantial area if large values are required. For the same reason, the bottom-plate parasitic, i.e., that due to drain and source junctions, is a relatively small percentage of the gate capacitance—typically 10 to 20%.

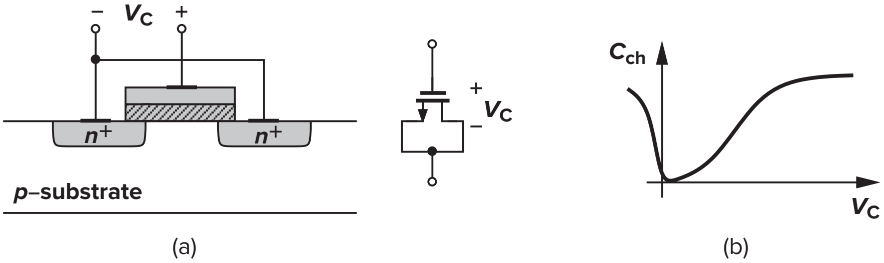


Figure 18.17 (a) MOSFET configured as a capacitor; (b) nonlinear C/V characteristic.

Unfortunately, the voltage dependence of MOS capacitors, even in strong inversion, makes the structure less attractive for precision charge transfer.

▶ **Example 18.1**

Consider the multiply-by-two amplifier of Sec. 13.3.3, shown in Fig. 18.18(a) as an implementation using a MOS capacitor C_1 and a linear capacitor C_2. Explain how the output voltage in the amplification mode is distorted.

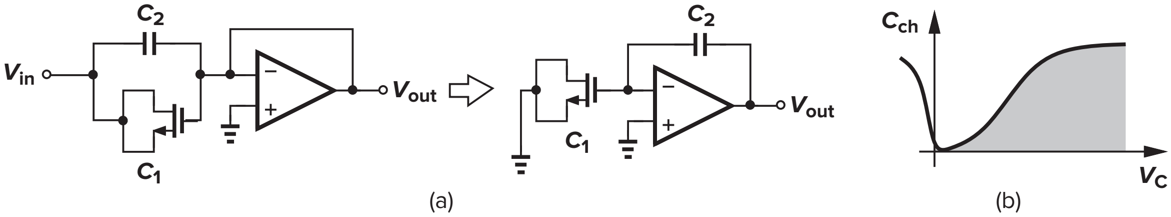


Figure 18.18 Precision multiply-by-two circuit using a MOS capacitor.

Solution

Suppose for simplicity that V_{in} is below ground by more than V_{TH}, so that the NMOS capacitors are in strong inversion during sampling. As the circuit enters the amplification mode, the voltage across C_1 approaches zero and the total charge stored on C_1 is transferred to C_2. How much is this charge? If C_1 were linear, we would have $Q = C_1 V$, but here we must write $dQ = C_1 dV$. Thus, as shown in Fig. 18.18(b), the total transferred charge when the voltage across the capacitor goes from V_{in} to zero is equal to the area under the C/V characteristic, a value substantially less than that in the linear case. The output voltage is then given by

$$V_{out} \approx V_{in} + \frac{1}{C_2}\int_0^{V_{in}} C_1 dV \tag{18.1}$$

◀

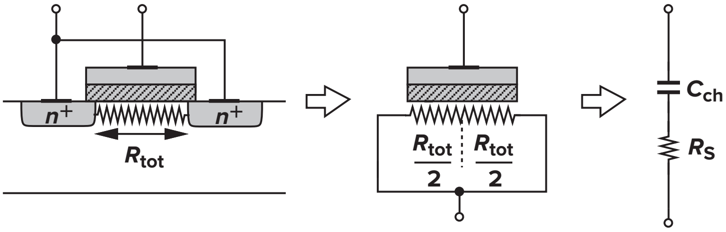


Figure 18.19 Channel resistance of MOS capacitor.

Another issue related to MOS capacitors is their series resistance, an effect arising from the gate material and, more important, the channel resistance. Assuming that proper layout minimizes the gate resistance, we view the channel resistance as shown in Fig. 18.19, estimating the equivalent series resistance as $(R_{tot}/2)\|(R_{tot}/2) = R_{tot}/4$, where $R_{tot} = [\mu C_{ox}(W/L)(V_{GS} - V_{TH}]^{-1}$. The intrinsic time constant of the capacitor is therefore equal to

$$\tau = \frac{R_{tot}}{4} C_{ch} \tag{18.2}$$

$$= \frac{1}{4\mu C_{ox}(W/L)(V_{GS} - V_{TH})} \cdot WLC_{ox} \tag{18.3}$$

$$= \frac{L^2}{4\mu(V_{GS} - V_{TH})} \tag{18.4}$$

In reality, the distributed nature of the resistance and the capacitance along the channel results in a time constant equal to one-third of that given above [2]. Another figure of merit for such a capacitor is $Q = [1/(C\omega)]/R_S$. As a rule of thumb, we choose $R_S < 0.1/(C\omega)$.

Equation (18.4) indicates that for a given overdrive, to minimize the series resistance of a MOS capacitor, L must be minimized. Consequently, MOS capacitors are usually designed as a parallel combination of wide, short devices rather than as a *square* block (Fig. 18.20). The penalty is a higher junction capacitance to the substrate and somewhat greater area.

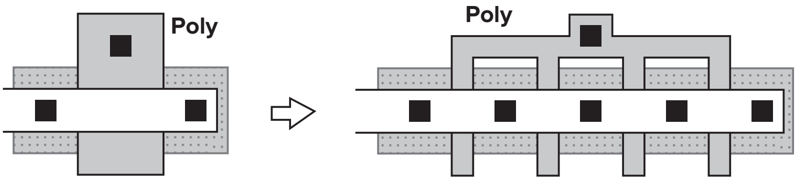


Figure 18.20 Use of wide, short MOS fingers to reduce channel resistance.

In applications requiring linear capacitors, a "sandwich" of conductive layers can be formed in CMOS technology. Shown in Fig. 18.21 is an example, where the capacitance between metal layers is exploited to increase the density. Since the dielectrics between the layers are relatively thick, this structure still requires a large area. More important, the bottom-plate parasitic (e.g., the capacitance between the lowest layer and the substrate in Fig. 18.21) is significant, about 5 to 10% of the total interplate capacitance. This structure is studied in detail in Chapter 19.

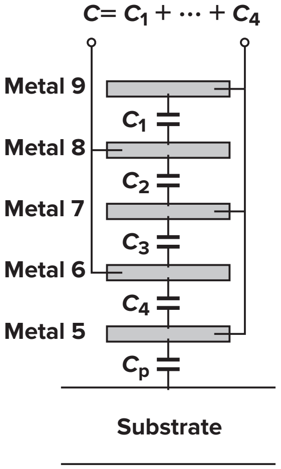


Figure 18.21 Linear capacitor made of native conductive layers.

▶ **Example 18.2**

An amplifier with an input capacitance of C_{in} is to be ac-coupled to a preceding stage having an output resistance R_{out}. Considering both of the topologies depicted in Fig. 18.22 and allowing a maximum signal attenuation of 20%, determine the minimum value of the coupling capacitor and the resulting time constant if $C_P = 0.5C_C$ or $C_P = 0.2C_C$.

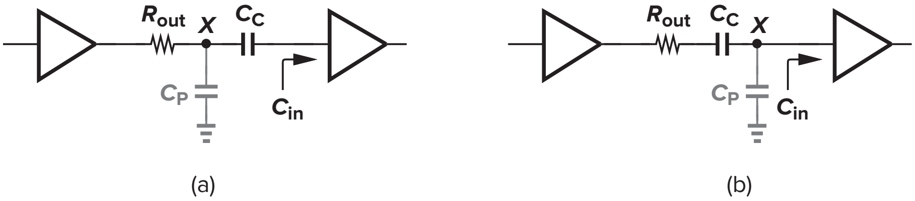


Figure 18.22

Solution

In Fig. 18.22(a), the attenuation is given by $A_v = C_C/(C_C + C_{in})$, yielding $C_C \geq 4C_{in}$ for a 20% signal loss. The total capacitance seen from node X to ground is therefore equal to $C_P + C_C C_{in}/(C_C + C_{in}) = C_P + 0.8C_{in}$. It follows that the time constant is $2.8R_{out}C_{in}$ for $C_P = 0.5C_C$ and $1.6R_{out}C_{in}$ for $C_P = 0.2C_C$.

In Fig. 18.22(b), C_P itself attenuates the signal: $A_v = C_C/(C_C + C_{in} + C_P)$, indicating that no value of C_C can yield a signal loss of 20% if $C_P \geq 0.25C_C$.

These calculations yield two important results. First, the topology of Fig. 18.22(a) is generally preferable. Second, the addition of a coupling capacitor, e.g., to isolate the bias levels, substantially degrades the speed. ◀

18.7.3 Interconnects

The performance of today's complex integrated circuits heavily depends on the quality of the available interconnects, requiring more metal layers in new generations of the technology.[4] Proper modeling of interconnects in a high-performance circuit is still a topic of active research, but our objective is to provide a basic understanding of the interconnect issues.

[4] At the time of this writing, five layers of metal are in production.

Two properties of interconnects, namely, series resistance and parallel capacitance, affect the performance, often calling for iteration between layout and circuit design. The series resistance becomes especially problematic in supply and ground lines, creating dc and transient voltage drops. Also, for long signal lines, the distributed resistance and capacitance of the wire may result in a significant delay.

The resistance of metal wires can be easily estimated at low frequencies, at which skin effect is negligible. Typical sheet resistances are 30 mΩ/□ for the topmost (thickest) layer and 100 mΩ/□ for lower layers. The finite resistance of wires influences the choice of line widths for high-current interconnects such as supply and ground buses, as illustrated by the following example.

▶ Example 18.3

A D/A converter incorporates N equal current sources implemented as NMOS devices, each having an aspect ratio of W/L [Fig. 18.23(a)]. Assuming that the interconnect between every two consecutive current sources has a small resistance, r, estimate the mismatch between I_N and I_1.

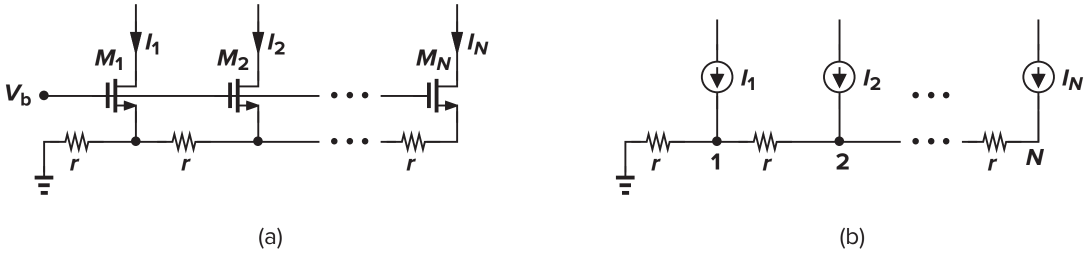


Figure 18.23 Effect of ground resistance in a D/A converter.

Solution

If r is sufficiently small, the circuit can be modeled as shown in Fig. 18.23(b), where $I_1 \approx I_2 \approx \cdots \approx I_N = I$. The voltage at node N is obtained by superposition of currents:

$$V_N = Ir + I(2r) + \cdots + I(Nr) \tag{18.5}$$

$$= \frac{N(N+1)}{2} Ir \tag{18.6}$$

If V_N is relatively small, the assumption that $I_1 \approx I_2 \approx \cdots \approx I_N$ used in the above calculation is reasonable and M_1–M_N exhibit roughly equal transconductances. Thus,

$$I_N = I - g_m V_N \tag{18.7}$$

$$= I - g_m r \frac{N(N+1)}{2} I \tag{18.8}$$

$$= I \left[1 - g_m r \frac{N(N+1)}{2}\right] \tag{18.9}$$

Since $V_1 \approx N \cdot I \cdot r$, we have $I_1 = I - g_m N \cdot I \cdot r$, and the relative mismatch between I_1 and I_N is

$$\left|\frac{I_1 - I_N}{I}\right| = g_m r \frac{N(N-1)}{2} \tag{18.10}$$

The key point here is that the error grows in proportion to N^2. The ground bus must therefore be sufficiently wide to minimize r.

Another factor determining the width of interconnects is “electromigration.” At high current densities, the aluminum atoms in a wire tend to “migrate,” leaving a void that eventually (after some years of operation) grows to a discontinuity. For this reason, long-term reliability considerations restrict the maximum current density of interconnects. As a rule of thumb, a current density of 2 mA per micron of width is acceptable, but the actual value varies according to the thickness of the metal. Also, for transient currents, the peak value may be quite a lot higher.

The problem of interconnect capacitance is much more complicated. We begin with a single wire on top of a substrate (Fig. 18.24), identifying a “parallel-plate” capacitance and a “fringe” capacitance. For narrow lines, the two are comparable.

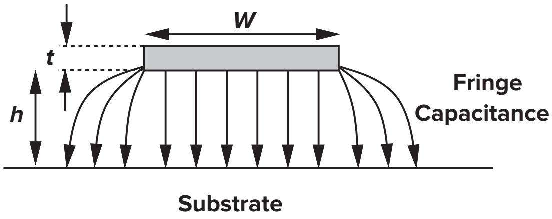


Figure 18.24 Parallel-plate and fringe capacitance of an interconnect.

A simple empirical relationship for calculating the total wire capacitance per unit length on top of a conducting substrate is

$$C = \epsilon \left[\frac{W}{h} + 0.77 + 1.06 \left(\frac{W}{h} \right)^{0.25} + 1.06 \left(\frac{t}{h} \right)^{0.5} \right] \tag{18.11}$$

where W, h, and t denote the dimensions shown in Fig. 18.24 [3]. For typical dimensions, this equation predicts the capacitance with a few percent of error.

While upper levels of metal in a process exhibit less capacitance per unit width and length, their minimum allowable width is usually greater than that of the lower layers. Thus, the minimum capacitance for a given length may be only slightly smaller for the topmost layer(s). Table 18.1 depicts typical values of minimum widths and parallel-plate and fringe capacitances (to the substrate) in a four-metal 0.25-μm process.

Table 18.1 Minimum widths and capacitances of interconnects in a 0.25-μm technology.

	Poly	Metal 1	Metal 2	Metal 3	Metal 4
Minimum Width (μm)	0.25	0.35	0.45	0.50	0.60
Bottom–Plate Capacitance (aF/μm^2)	90	30	15	9.0	7.0
Fringe Capacitance (Two Sides) (aF/μm)	110	80	50	40	30

Wires also suffer from parallel and fringe capacitances between them. Illustrated in Fig. 18.25, this effect is difficult to quantify for a complex layout, often necessitating the use of computer programs. In

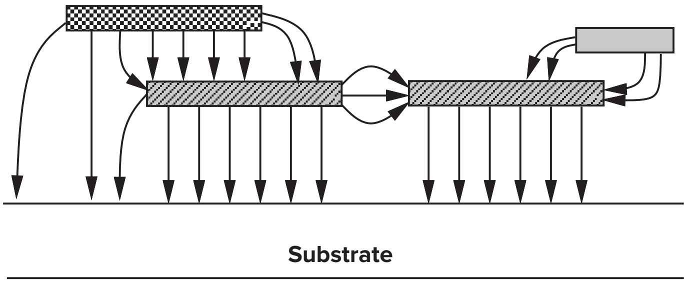


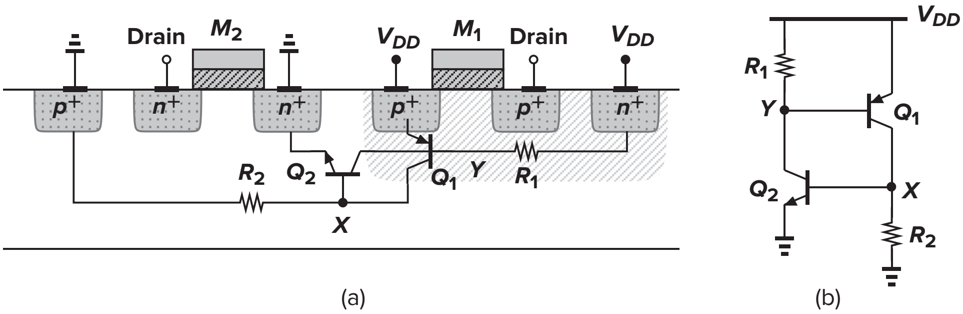
Figure 18.25 Complex interconnect structure.

practice, the capacitances between the layers are calculated by "electromagnetic field solvers," measured experimentally, and tabulated in the process design manual.

18.8 ■ Latch-Up

Owing to manufacturing difficulties, the first few generations of MOS technologies provided only NMOS devices. In fact, many of the early microprocessors and analog circuits were fabricated in NMOS processes, but they consumed substantial power. The advent of CMOS technology was motivated by the zero static power dissipation of CMOS logic—although CMOS devices required a greater number of masks and fabrication steps. Another issue that did not exist in NMOS implementations but arose in CMOS circuits was latch-up.

Consider the NMOS and PMOS devices shown in Fig. 18.26(a). Recall from Chapter 12 that a parasitic *pnp* bipolar transistor, Q_1, is associated with the PFET, the n-well, and the substrate. By the same token, a parasitic *npn* device, Q_2, can be identified in conjunction with the NFET. We make two observations: (1) the base of each bipolar transistor is inevitably tied to the collector of the other; and (2) owing to the finite resistance of the n-well and the substrate, the bases of Q_1 and Q_2 see a nonzero resistance to V_{DD} and ground, respectively. The parasitic circuit can therefore be drawn as in Fig. 18.26(b), revealing a *positive* feedback loop around Q_1 and Q_2. In fact, if a current is injected into node X such that V_X rises, then I_{C2} increases, V_Y falls, $|I_{C1}|$ increases, and V_X rises further. If the loop gain is greater than or equal to unity, this phenomenon continues until both transistors turn on completely, drawing an enormous current from V_{DD}. We say that the circuit is latched up.



(a) (b)

Figure 18.26 (a) Parasitic bipolar transistors in a CMOS process; (b) equivalent circuit.

The initial current required to trigger latch-up may be produced by various sources in an integrated circuit. For example, in Fig. 18.26(a), the bases of Q_1 and Q_2 are capacitively coupled to the drains of M_1 and M_2, respectively. A large voltage swing at the drains can therefore inject a significant displacement current into the n-well or the substrate, initiating latch-up.

A common case of latch-up occurs with the use of large digital output buffers (inverters). These circuits inject high currents into the substrate through the large drain junction capacitance of the transistors and by forward-biasing the source-bulk junction diodes. The latter arises because of the substantial transient voltages produced across the bond wires connected to the ground (Chapter 19).

In order to prevent latch-up, both process engineers and circuit designers take precautions to ensure that the loop gain of the equivalent circuit shown in Fig. 18.26(b) remains well below unity. Proper choice of the doping levels and profiles as well as layout design rules ensure a low value for both the parasitic resistances and the current gain of the bipolar transistors. Furthermore, the layout of the circuit incorporates substrate and n-well contacts with sufficiently small spacing to minimize the resistance. The design manual of each technology typically provides an extensive set of layout rules recommended for latch-up prevention.

References

[1] C. Kaya et al., "Polycide/Metal Capacitors for High Precision A/D Converters," *IEDM Dig. of Tech. Papers*, pp. 782–785, December 1988.

[2] P. Larsson, "Parasitic Resistance in an MOS Transistor Used as On-Chip Decoupling Capacitor," *IEEE J. Solid-State Circuits*, vol. 32, pp. 574–576, April 1997.

[3] E. Barke, "Line-to-Ground Capacitance Calculations for VLSI: A Comparison," *IEEE Trans. on Commputer-Aided Design*, vol. 7, pp. 195–298, February 1988.

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

18.1. A MOS technology is designed to provide only n-type transistors and two metal layers. Sketch the fabrication steps and determine the minimum number of masks required in this technology.

18.2. During a threshold-adjust implant, the wafer was not tilted, leading to severe channeling. Explain whether the resulting threshold voltage is higher or lower than the target value.

18.3. The circuits of Fig. 18.27 have been fabricated with a longer-than-expected gate oxidation cycle. If the threshold voltages are still equal to the desirable value, sketch V_{out} versus V_{in} and compare the results to the target case.

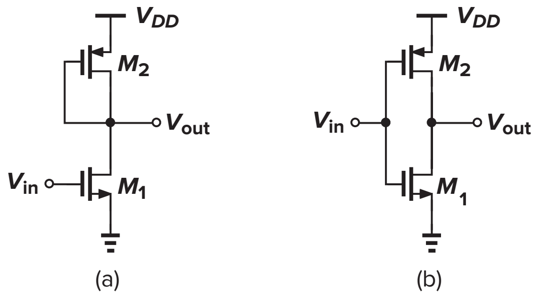


Figure 18.27

18.4. The circuits of Fig. 18.27 have been fabricated without a threshold-adjust implant. Sketch V_{out} versus V_{in} and compare the results to the target case.

18.5. Due to a layout error, the circuit shown in Fig. 18.28 suffers from contact spiking in one of the junctions. Identify the faulty junction if (a) the voltage gain is higher than expected, (b) the output voltage is near V_{DD}.

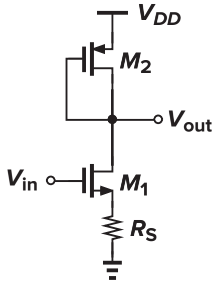


Figure 18.28

18.6. An NMOS cascode current source used in a large circuit exhibits a substantially lower output impedance than expected. Determine which fabrication error may have led to this effect: (a) channeling during S/D implant, (b) omission of the channel-stop implant, or (c) insufficient gate-oxide growth.

18.7. An NMOS cascode current source has a zero output current. If a single (small) lithography misalignment has caused this error, determine in which fabrication step(s) this may have occurred.

18.8. A differential pair using an active current mirror as load suffers from a low small-signal voltage gain. If the bias current is equal to the target value, determine which fabrication error may have led to this effect: (a) heavy n-well implantation, (b) heavy threshold-adjust implantation, or (c) long gate oxidation cycle.

18.9. The switched-capacitor amplifier of Fig. 18.29 exhibits a large gain error. If the bias current of the op amp is equal to the desired value, which fabrication error is likely to have happened: (a) heavy threshold-adjust implantation, (b) very heavy doping in the bottom plate of C_1 (placed at node P), or (c) channeling during the S/D implantation?

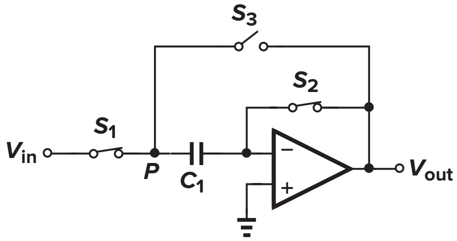


Figure 18.29

18.10. In Fig. 18.30, the digital circuit draws large transient currents from V_{DD}. Without M_1, the inductor L_b would sustain a large transient voltage $L_b dI_{DD}/dt$. Transistor M_1 with $W/L = 100/0.5$ is added to suppress this effect.

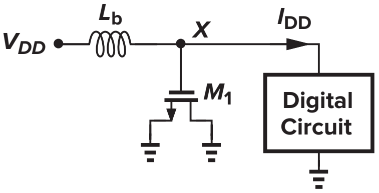


Figure 18.30

(a) Calculate the equivalent series resistance of M_1.

(b) Calculate the maximum value of L_b that results in a critically-damped response at node X. Model the digital circuit by a transient current source.

18.11. In the circuit of Fig. 18.23, $V_b = 1.2$ V, $N = 32$, and $(W/L)_{1-N} = 20/0.5$. Determine the maximum value of r for a maximum current mismatch of 1%.

18.12. Suppose that in Eq. (18.11), $t = 1\ \mu$m and $h = 3\ \mu$m. For what value of W are the parallel-plate and fringe capacitances equal? What if $h = 5\ \mu$m?

CHAPTER

19

Layout and Packaging

In the past 40 years, analog CMOS circuits have evolved from low-speed, low-complexity, small-signal, high-voltage topologies to high-speed, high-complexity, low-voltage "mixed-signal" systems containing a great deal of digital circuitry. While device scaling has enhanced the raw speed of transistors, unwanted interaction between different sections of integrated circuits as well as nonidealities in the layout and packaging are increasingly limiting both the speed and the precision of such systems. Today's analog circuit design is very heavily influenced by layout and packaging.

In this chapter, we study principles of layout and packaging, emphasizing the effects that manifest themselves when analog and digital circuits coexist on a chip. For the sake of brevity, we use the term "analog" to mean both "analog" and "mixed-signal." Beginning with an overview of layout design rules, we study a number of topics related to the layout of analog circuits, including multifinger transistors, symmetry, reference distribution, passive device layout, and interconnects. Next, we deal with the problem of substrate coupling. Finally, we describe packaging issues, analyzing the effect of self- and mutual inductance and capacitance of external connections to integrated circuits.

19.1 ■ General Layout Considerations

The layout of an integrated circuit defines the geometries that appear on the masks used in fabrication. From Chapter 18, the geometries include n-well, active, polysilicon, n^+ and p^+ implants, interlayer contact windows, and metal layers.

Figure 19.1 shows an example, where the mask geometries required for a PMOS transistor are drawn. It is important to note the following: (1) the n-well surrounds the device with enough margin to ensure that the transistor is contained in the well for all expected misalignments during fabrication; (2) each

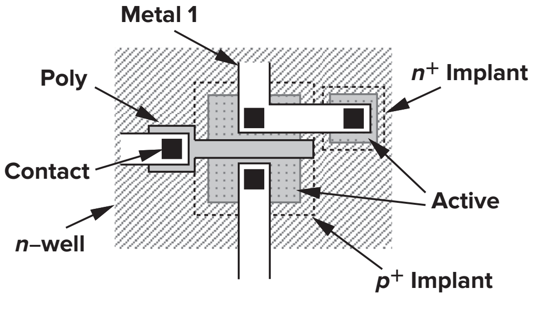


Figure 19.1 Layout of a PMOS transistor.

"active" area (S/D regions and n^+ contact to the well) is surrounded by a proper implant geometry with enough margin; (3) from the fabrication steps described in Chapter 18, the gate requires its own mask; (4) the contact windows mask provides connection from active and poly regions to the first layer of metal.

In most modern layout tools, the implants and even the n-wells are automatically generated from the remainder of the transistor geometries, reducing the number of layers that the layout designer draws or sees on the computer screen and simplifying the task.

19.1.1 Design Rules

While the width and length of each transistor are determined by circuit design, most of the other dimensions in a layout are dictated by "design rules," i.e., a set of rules that guarantees proper transistor and interconnect fabrication despite various tolerances in each step of processing. Most design rules can be categorized under one of the four groups described here.

Minimum Width The widths (and lengths) of the geometries defined on a mask must exceed a minimum value imposed by both lithography and the processing capabilities of the technology. For example, if a polysilicon rectangle is excessively narrow, then, owing to fabrication tolerances, it may simply break or at least suffer from a large local resistance (Fig. 19.2). In general, the thicker a layer, the greater its minimum allowable width, indicating that as technologies scale, the thickness must be decreased proportionally. Figure 19.3 depicts examples of minimum widths in a 40-nm technology. Note that the thickness of the layers is not under the control of the layout designer.

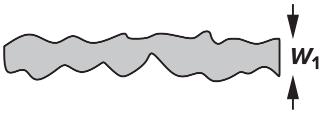


Figure 19.2 Excessive width variation in a narrow poly line.

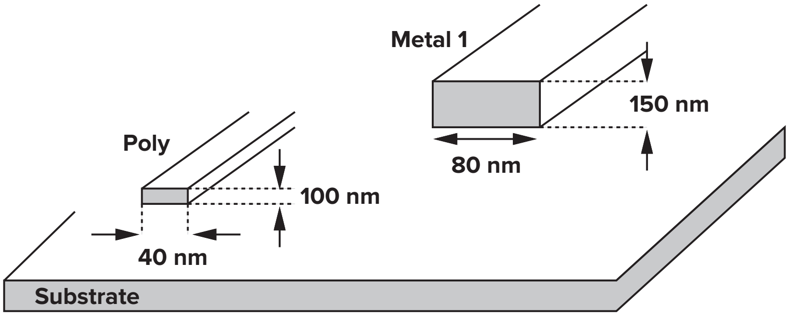


Figure 19.3 Widths and thicknesses of poly and metal lines.

Minimum Spacing The geometries built on the same mask or, in some cases, different masks must be separated by a minimum spacing. For example, as shown in Fig. 19.4(a), if two polysilicon lines are placed too close to each other, they may be shorted. As another example, consider the case shown in Fig. 19.4(b), where a polysilicon line runs close to the S/D area of a transistor. A minimum spacing is required here to ensure that the implant surrounding the transistor does not overlap with the poly line.

Minimum Enclosure We mentioned earlier that in the layout of Fig. 19.1, the n-well and the p^+ implant must surround the transistor with sufficient margin to guarantee that the device is contained by these geometries despite tolerances. These are examples of minimum enclosure rules. Figure 19.5 depicts

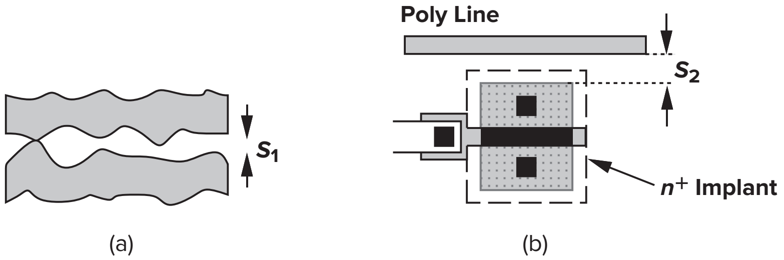


Figure 19.4 (a) Short between two excessively close poly lines; (b) minimum spacing between active and poly.

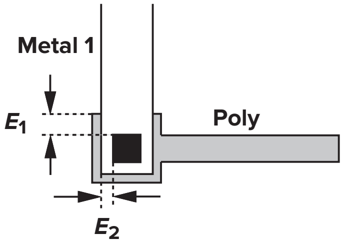


Figure 19.5 Enclosure rule for poly and metal surrounding a contact.

another example, where a poly contact window connects a poly line to a metal 1 line. To ensure that the contact remains inside the poly and metal 1 squares, both geometries must enclose the contact with enough margin.

Minimum Extension Some geometries must extend beyond the edge of others by a minimum value. For example, as shown in Fig. 19.6, the gate polysilicon must have a minimum extension beyond the active area to ensure proper transistor action at the edge.

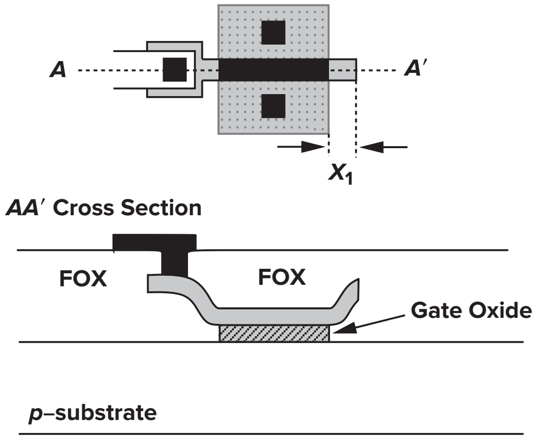


Figure 19.6 Extension of poly beyond the gate area.

In addition to the minimum dimensions specified in the previous four categories, some *maximum allowable* dimensions may also be enforced. For example, for long metal wires, the minimum width is typically larger than that for short wires to avoid “liftoff” problems. Other such rules relate to the “antenna effect,” described in the next section.

Figure 19.7 summarizes a small subset of design rules governing the layout of an NMOS differential pair with PMOS current-source loads. Modern CMOS technologies typically involve several hundred layout design rules.

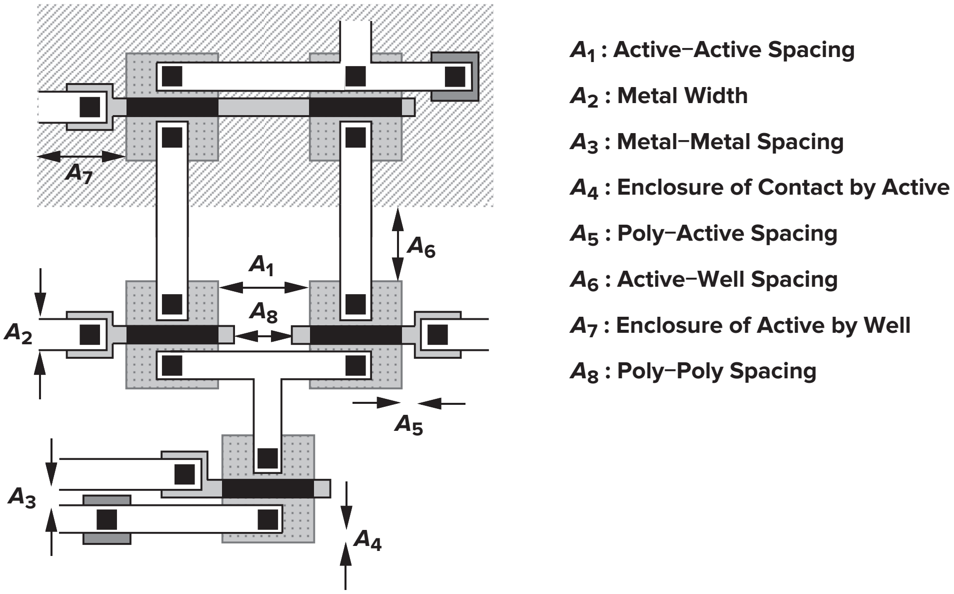

A_1 : Active–Active Spacing
A_2 : Metal Width
A_3 : Metal–Metal Spacing
A_4 : Enclosure of Contact by Active
A_5 : Poly–Active Spacing
A_6 : Active–Well Spacing
A_7 : Enclosure of Active by Well
A_8 : Poly–Poly Spacing
A_1
A_2
A_3
A_4
A_5
A_6
A_7
A_8

Figure 19.7 Layout of a differential pair with PMOS current-source loads.

19.1.2 Antenna Effect

Suppose the gate of a small MOSFET is tied to a metal 1 interconnect having a large area [Fig. 19.8(a)]. During the etching of metal 1, the metal area acts as an "antenna," collecting ions and rising in potential. It is therefore possible that the gate voltage of the MOS device increases so much that the gate oxide breaks down (irreversibly) during fabrication.

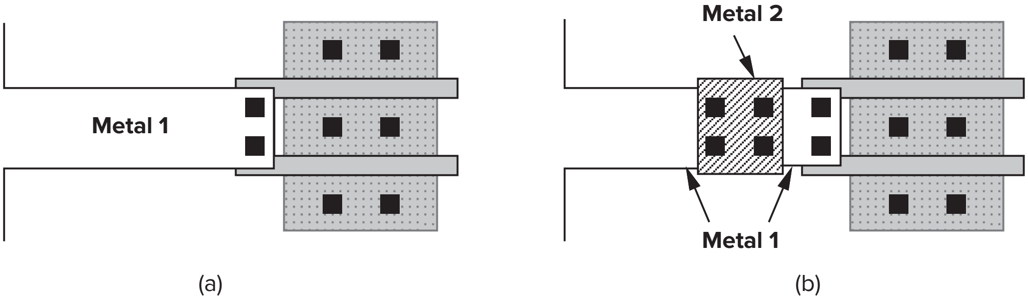


Figure 19.8 (a) Layout susceptible to antenna effect; (b) discontinuity in metal 1 layer to avoid antenna effect.

The antenna effect may occur for any large piece of conductive material tied to the gate, including polysilicon itself. For this reason, submicron CMOS technologies typically limit the total area of such geometries, thereby minimizing the probability of gate-oxide damage. If large areas are inevitable, then a discontinuity can be created as illustrated in Fig. 19.8(b) so that, when metal 1 is being etched, the large area is not connected to the gate.

19.2 ■ Analog Layout Techniques

The extensive sets of design rules enforced by mainstream CMOS processes aim to maximize the yield of digital ICs while allowing moderately aggressive circuit design. Analog systems, on the other hand, demand many more layout precautions so as to minimize effects such as crosstalk, mismatches, noise, etc.

19.2.1 Multifinger Transistors

As mentioned in Chapter 2, wide transistors are usually "folded" so as to reduce both the S/D junction area and the gate resistance. A simple folded structure such as that in Fig. 19.9(a) may prove inadequate for very wide devices, necessitating the use of multiple "fingers" [Fig. 19.9(b)]. As a rule of thumb, the width of each finger is chosen such that the resistance of the finger is less than the inverse transconductance associated with the finger. In low-noise applications, the gate resistance must be one-fifth to one-tenth of $1/g_m$.

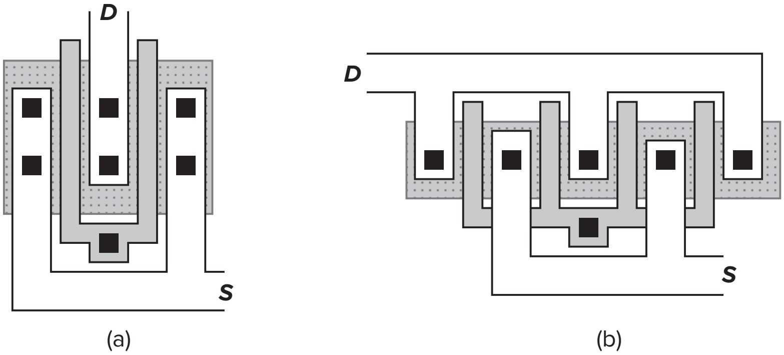


Figure 19.9 (a) Simple folding of a MOSFET; (b) use of multiple fingers.

Example 19.1

A 5-μm/40-nm MOSFET biased at 1 mA exhibits a transconductance of $1/(100\ \Omega)$. If the sheet resistance of the gate polysilicon is equal to $30\ \Omega/\square$, what is the widest finger that the structure can incorporate while ensuring that the gate thermal noise voltage is one-fifth of the gate-referred channel thermal noise voltage?

Solution

If the transistor is laid out as N parallel fingers, each finger exhibits a distributed resistance of $30\ \Omega \times (5/0.04)/N$. Using the gate-referred channel thermal noise from Chapter 7, we have for the overall transistor

$$\text{Channel Noise} = \sqrt{4kT\gamma(100)}\ \text{V}/\sqrt{\text{Hz}} \tag{19.1}$$

$$\text{Gate Noise} = \sqrt{4kT\frac{150}{0.04N^2}\frac{1}{3}}\ \text{V}/\sqrt{\text{Hz}} \tag{19.2}$$

where the factor 1/3 on the right-hand side of (19.2) accounts for the distributed nature of the resistance (Chapter 7). Equating (19.1) to five times (19.2) and assuming that $\gamma = 1$, we have

$$N = 17.7 \tag{19.3}$$

Thus, a minimum of 18 fingers is required.

While the gate resistance can be reduced by decomposing the transistor into more parallel fingers, the capacitance associated with the perimeter of the source/drain areas increases. As exemplified by the structures depicted in Fig. 19.10,[1] with three fingers, the total perimeter of the source or the drain is equal to $2(2E + 2W/3) = 4E + 4W/3$, whereas with five fingers, it is equal to $3(2E + 2W/5) = 6E + 6W/5$.

[1] The use of multiple fingers is sometimes called "interdigitization."

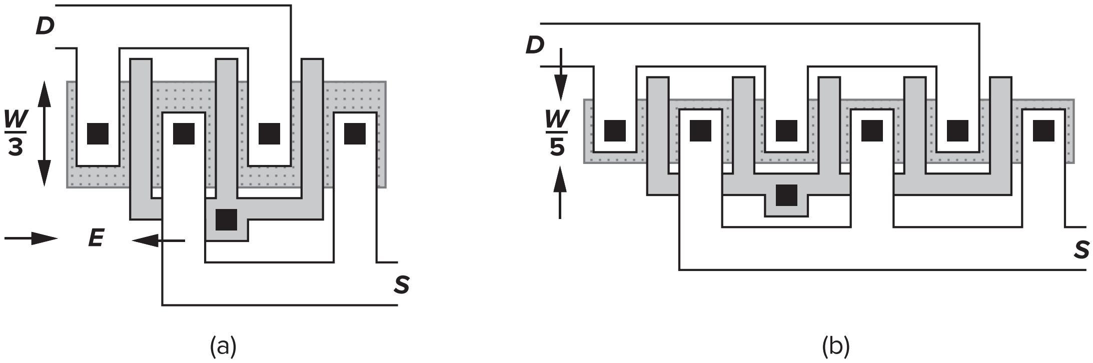


Figure 19.10 Layout of a transistor using (a) three fingers and (b) five fingers.

In general, for an odd number of fingers N, the S/D perimeter capacitance is given by

$$C_P = \frac{N+1}{2}\left(2E + \frac{2W}{N}\right)C_{jsw} \tag{19.4}$$

$$= \left[(N+1)E + \frac{N+1}{N}W\right]C_{jsw} \tag{19.5}$$

Thus, the number of fingers multiplied by E must be much less than W so as to minimize the S/D perimeter capacitance contribution. In practice, this requirement may conflict with that for minimizing the gate resistance noise, demanding a compromise between the two or contacting the gate on both ends to reduce the resistance.

For transistors having a large number of gate fingers, the structure may be modified to that shown in Fig. 19.11, thereby avoiding long geometries and hence disproportionate dimensions in the layout of the overall circuit.

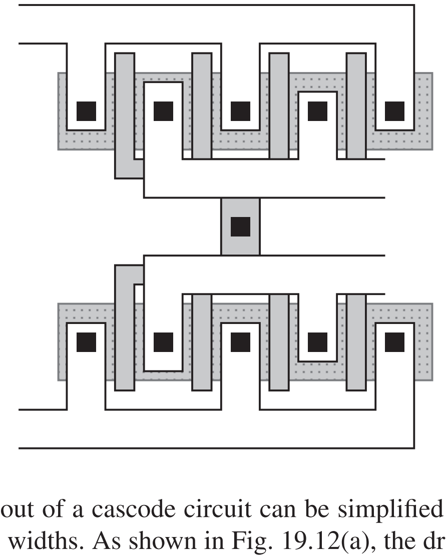


Figure 19.11 Layout of a wide transistor with many fingers.

The layout of a cascode circuit can be simplified if the input device M_1 and the cascode device M_2 have equal widths. As shown in Fig. 19.12(a), the drain of M_1 and the source of M_2 can share the same junction. More important, since this junction is not connected to any other node, it need not accommodate a contact window and can therefore be quite a lot smaller [Fig. 19.12(b)]. Consequently, the capacitance at

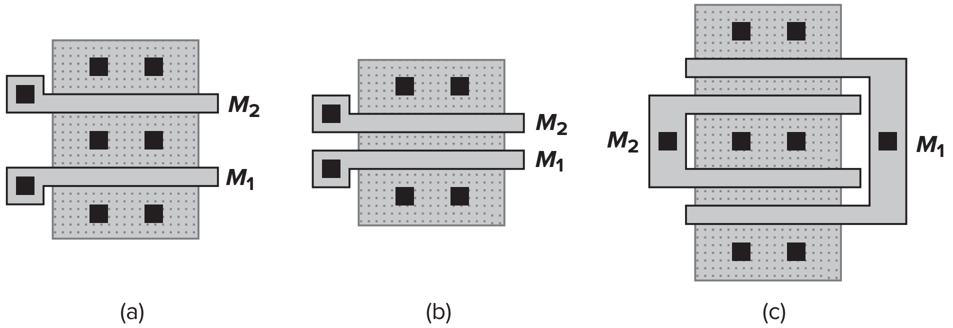


Figure 19.12 Layout of cascode devices having the same width.

the drain of M_1 is reduced substantially, improving the high-frequency performance. For wide transistors, each transistor may use two or more fingers [Fig. 19.12(c)].

19.2.2 Symmetry

Recall from Chapter 14 that asymmetries in fully differential circuits introduce input-referred offsets, thus limiting the minimum signal level that can be detected. While some mismatch is inevitable, inadequate attention to symmetry in the layout may result in large offsets—much greater than the values predicted by the statistical treatment of Chapter 14. Symmetry also suppresses the effect of common-mode noise and even-order nonlinearity. It is important to note that symmetry must be applied to both the devices of interest and their surrounding environment. We return to this point later.

Let us consider the differential pair of Fig. 19.13(a) as the starting point. If, as depicted in Fig. 19.13(b), the two transistors are laid out with different orientations, the matching suffers greatly because many steps in lithography and wafer processing behave differently along different axes.

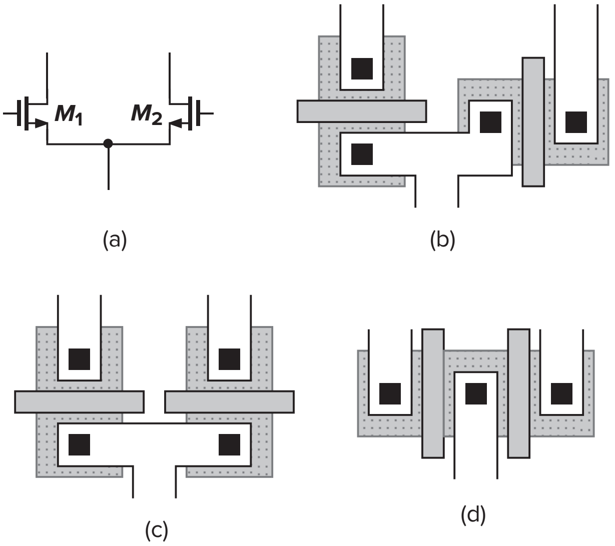


Figure 19.13 (a) Differential pair; (b) layout of M_1 and M_2 with different orientations; (c) layout with gate-aligned devices; (d) layout with parallel-gate devices.

Thus, one of the configurations in Fig. 19.13(c) and (d) provides a more plausible solution. The choice between these two is determined by a subtle effect called "gate shadowing." Illustrated in Fig. 19.14, the

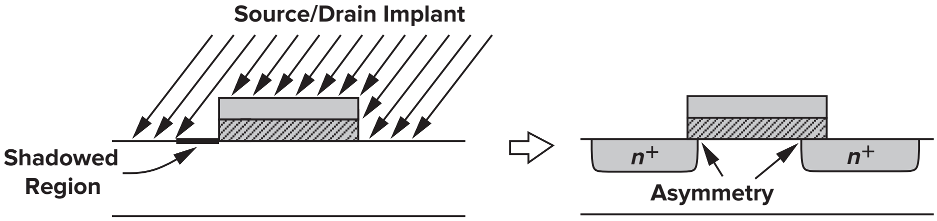


Figure 19.14 Shadowing due to implant tilt.

shadowing is caused by the gate polysilicon during the source/drain implantation because the implant (or the wafer) is tilted by about 7° to avoid channeling (Chapter 18). As a result, a narrow strip in the source or drain region receives less implantation, creating a small asymmetry between the source and drain side diffusions after the implanted areas are annealed.

Now consider the structures of Figs. 19.13(c) and (d) in the presence of gate shadowing (Fig. 19.15). In Fig. 19.15(a), if the shadowed terminal is distinguished as the drain (or the source), then the two devices sustain no asymmetry resulting from shadowing. In Fig. 19.15(b), on the other hand, the transistors are not identical even if the shadowed terminals are distinguished because the source region of M_1 "sees" M_2 to its right, whereas the source region of M_2 sees only the field oxide. Similarly, the drains of M_1 and M_2 see different structures to their left. In other words, the surrounding environment of M_1 is not identical to that of M_2. For this reason, the topology of Fig. 19.15(a) is preferable.

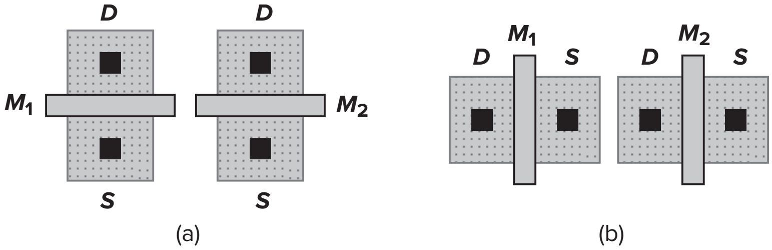


Figure 19.15 Effect of shadowing on (a) gate-aligned and (b) parallel-gate transistors.

The asymmetry inherent in the structures of Fig. 19.15(b) can be ameliorated by adding "dummy" transistors to the two sides so that M_1 and M_2 see approximately the same environment (Fig. 19.16). However, in more complex circuits, e.g., in a folded-cascode op amp, such measures cannot be easily applied. We will see later that a simpler version of dummies proves useful and essential in today's technologies.

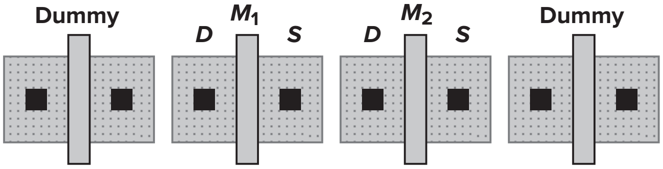


Figure 19.16 Addition of dummy devices to improve symmetry.

We should emphasize the importance of maintaining the same environment on the two sides of the axis of symmetry. For example, in the structure of Fig. 19.17, an unrelated metal line passing over only

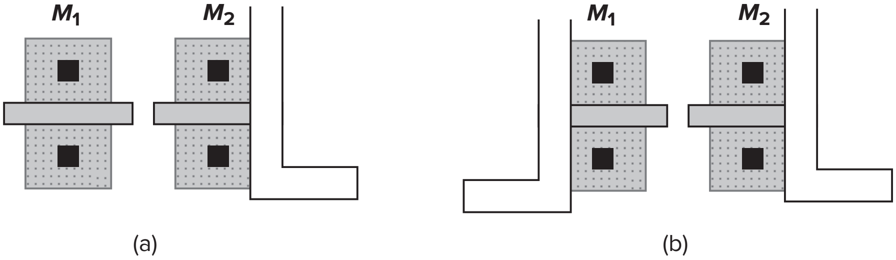


Figure 19.17 (a) Asymmetry resulting from a metal line passing over M_2; (b) removing the asymmetry by replicating the line on top of M_1.

one transistor indeed degrades the symmetry, increasing the mismatch between M_1 and M_2. In such cases, either a replica must be produced on the other side [Fig. 19.17(b)] (even though the replica may be grounded) or, preferably, the source of asymmetry must be removed.

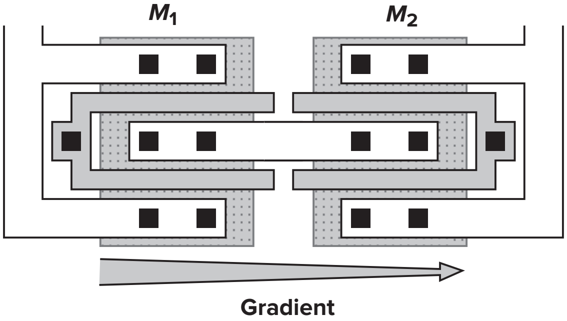


Figure 19.18 Effect of gradient in a differential pair.

Symmetry becomes more difficult to establish for large transistors. In the differential pair of Fig. 19.18, for example, the two transistors have a large width so as to achieve a small input offset voltage, but gradients along the x axis give rise to appreciable mismatches. To reduce the error, a "common-centroid" configuration may be used such that the effect of first-order gradients along both axes is canceled. Illustrated in Fig. 19.19, the idea is to decompose each transistor into two halves that are placed diagonally

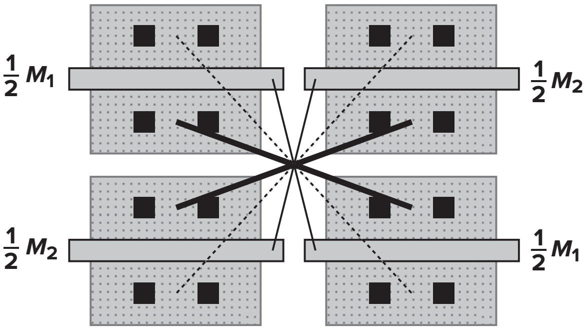
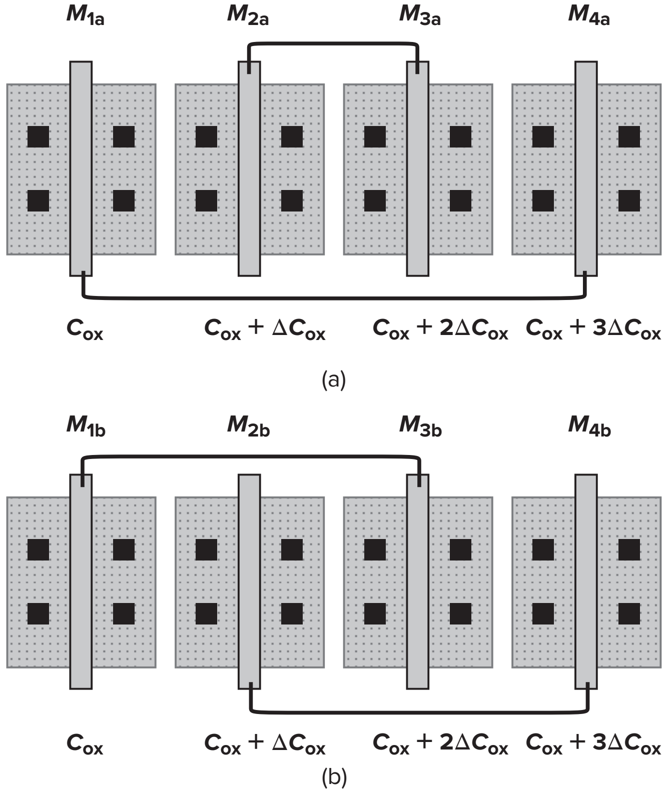


Figure 19.19 Common-centroid layout.



(a)

(b)

Figure 19.20 One-dimensional cross-coupling.

opposite each other and connected in parallel.[2] However, the routing of interconnects in this layout is quite difficult, often leading to systematic asymmetries of the type depicted in Fig. 19.17(a) or in the capacitances from the wires to ground and between the wires. For a larger circuit, e.g., an op amp, the routing may become prohibitively complex. We thus seek simpler solutions.

The effect of linear gradients can be suppressed by "one-dimensional" cross-coupling, as depicted in Fig. 19.20. Here, all four half transistors are placed along the same axis, and M_1 and M_2 are formed by connecting either the near ones and the far ones [Fig. 19.20(a)] or every other one [Fig. 19.20(b)]. (For clarity, the connections between the sources and the drains are not shown.) To analyze the effect of gradients in these structures, let us assume that, for example, the gate-oxide capacitance varies by ΔC_{ox} from each half transistor to the next.[3] Placing M_{1a} and M_{4a} in parallel, we have

$$I_{D1a} + I_{D4a} = \frac{1}{2}\mu_n(C_{ox} + C_{ox} + 3\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{19.6}$$

and for M_{2a} and M_{3a},

$$I_{D2a} + I_{D3a} = \frac{1}{2}\mu_n(C_{ox} + \Delta C_{ox} + C_{ox} + 2\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{19.7}$$

[2] The interconnect lines shown in this figure are only conceptually correct.

[3] In reality, variation of C_{ox} influences the threshold voltage as well. We neglect this effect here.

This type of cross-coupling therefore cancels the effect of the gradient. Now, for the configuration of Fig. 19.20(b), we have

$$I_{D1b} + I_{D3b} = \frac{1}{2}\mu_n(C_{ox} + C_{ox} + 2\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{19.8}$$

and

$$I_{D2b} + I_{D4b} = \frac{1}{2}\mu_n(C_{ox} + \Delta C_{ox} + C_{ox} + 3\Delta C_{ox})\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{19.9}$$

Equations (19.8) and (19.9) suggest that this approach removes the error to a lesser extent.

The reader can prove that for small gradients in other device parameters, similar results are obtained, concluding that the topology of Fig. 19.20(a) contains smaller errors than that of Fig. 19.20(b). However, since the environment seen by $M_{2a} + M_{3a}$ differs from that seen by $M_{1a} + M_{4a}$, dummy transistors must be added to the left of M_{1a} and the right of M_{4a}.

19.2.3 Shallow Trench Isolation Issues

Modern MOS devices are surrounded by a shallow "trench" so as to avoid the formation of a channel between adjacent transistors [Fig. 19.21(a)]. Called "shallow trench isolation" (STI) and created automatically, this structure is filled with oxide and exhibits a different thermal expansion coefficient from that of silicon. As a result, during fabrication steps, the STI and the enclosed silicon area expand and contract differently. This STI-induced "stress" alters the electrical properties of the MOS transistor, introducing substantial error in its I/V characteristics.

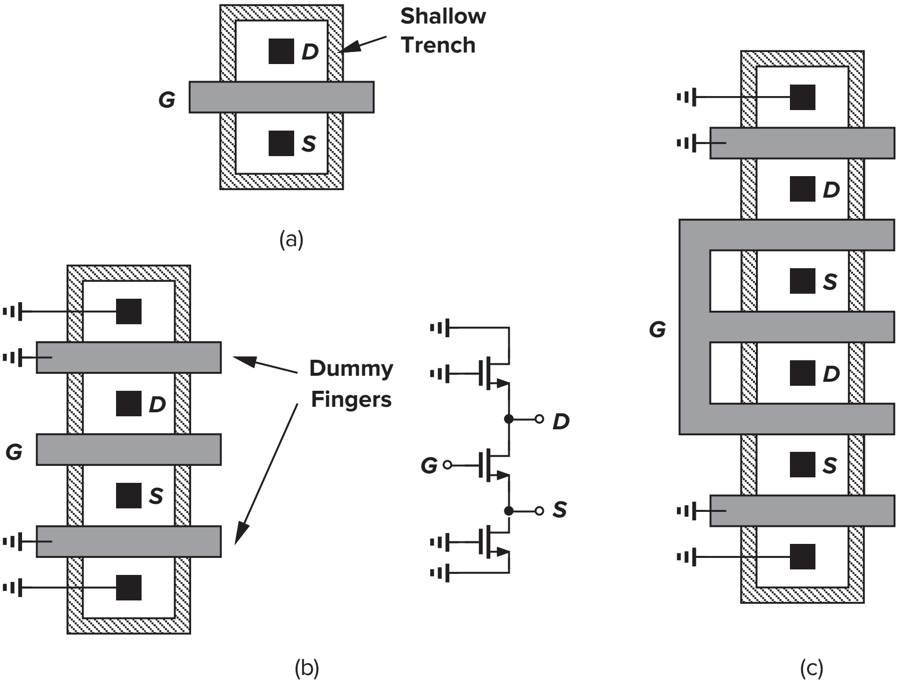


Figure 19.21 (a) Shallow trench isolation surrounding a device, (b) use of dummy fingers to reduce STI-induced stress, and (c) multifinger transistor example.

In order to alleviate this issue, we must minimize the propagation of the stress toward the gate area. To this end, we insert two fingers on the two sides of the main device [Fig. 19.21(b)]. These "dummy" fingers and their associated S/D junctions are typically grounded to ensure that they do not interfere with the operation of the main transistor. Note, however, that the dummy gates increase the S/D capacitances to ground.

For transistors employing multiple fingers, the dummy gates can be simply added to the two ends of the array [Fig. 19.21(c)].

19.2.4 Well Proximity Effects

As explained in Chapter 18, an n-well is formed by an N-type implant onto the exposed areas of silicon. The unexposed areas are covered by a thick layer consisting of oxide and photoresist [Fig. 19.22(a)]. Unfortunately, the implant does not occur at a 90° angle with respect to the wafer, thus reflecting from the walls formed by oxide and photoresist and creating *nonuniform* doping in the n-well. That is, the border areas of the n-well receive a different doping density from those in the middle of the n-well. Consequently, the PMOS devices located near the *edges* of the n-well have different I/V characteristics compared to those in the middle. We call this effect the "well proximity" error. For example, the current mirror arrangement shown in Fig. 19.22(b) exhibits mismatches between M_1 and M_2 or M_3 because M_1 is more heavily influenced by the implant reflections.

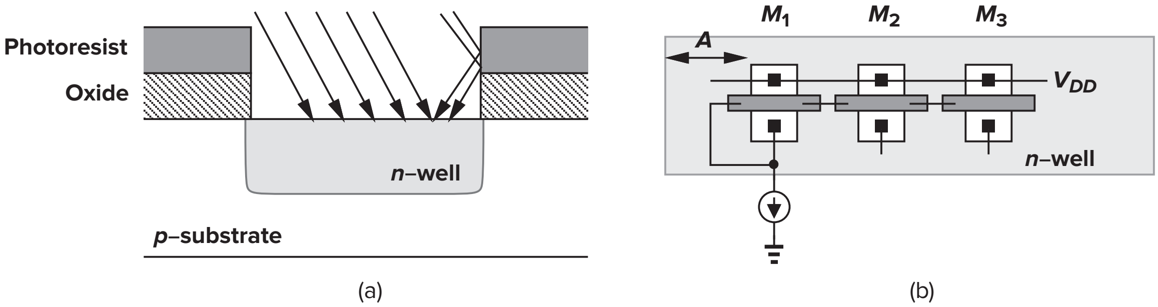


Figure 19.22 (a) Effect of implant reflection on n-well doping uniformity, and (b) current mirror arrangement showing the effect of n-well edges.

To reduce the well proximity effect, the n-well must extend well beyond the PMOS devices. For example, A in Fig. 19.22(b) can be chosen greater than several microns.

19.2.5 Reference Distribution

In analog systems, the bias currents and voltages of various building blocks are derived from one or more bandgap reference generators. The distribution of such references across a large chip entails a number of important issues. Consider the example depicted in Fig. 19.23, where I_{REF} is produced by a bandgap reference and M_1–M_n serve as bias current sources of building blocks that are located far from M_{REF} and from each other. If the matching between I_{D1}–I_{Dn} and I_{REF} is critical, then the voltage drop along the ground line must be taken into account. In fact, for a large number of circuits connected to the same ground line, the systematic mismatch between the current sources and I_{REF} may be unacceptable.

To remedy the above difficulty, the reference can be distributed in the current domain rather than in the voltage domain. Illustrated in Fig. 19.24, the idea is to route the reference current to the vicinity of the

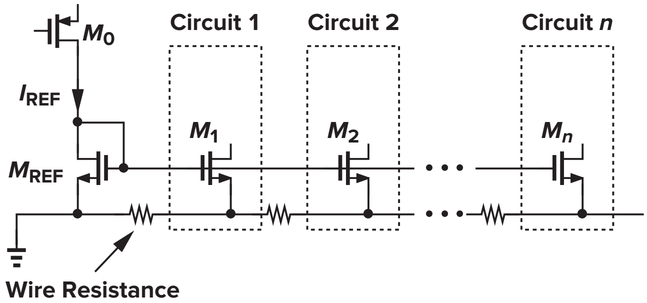


Figure 19.23 Distribution of a reference voltage for current mirror biasing.

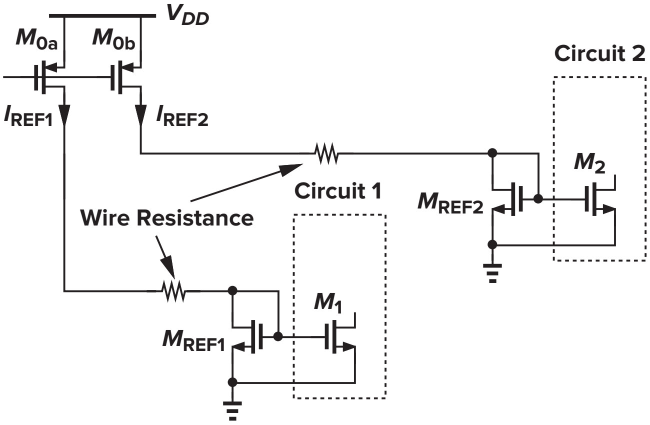


Figure 19.24 Distribution of current to reduce the effect of interconnect resistance.

building blocks and perform the current mirror operation *locally*. At the destination, bypass capacitors suppress any noise that the long interconnects may pick up. Placing the interconnect resistance in series with current sources, this approach lowers systematic errors if the building blocks appear in dense groups in different regions on the chip. However, mismatches between I_{REF1} and I_{REF2} and between M_{REF1} and M_{REF2} introduce error. In large systems, it may be advantageous to employ several local bandgap reference circuits so as to alleviate routing problems.

Another issue in the circuits of Figs. 19.23 and 19.24 relates to the orientation of the transistors. As mentioned in Sec. 19.2.2, if, for example, M_{REF} and M_1–M_n in Fig. 19.23 have different orientations, then substantial mismatches arise. Since circuits $1, 2, \ldots, n$ may be laid out individually, particular attention must be paid to the orientation of their current sources while the entire chip is assembled.

The scaling of currents in Figs. 19.23 and 19.24 also demands careful choice of device dimensions and layout. Suppose the circuit of Fig. 19.23 requires $I_{D1} = 0.5I_{REF}$ and $I_{D2} = 2I_{REF}$. How do we choose $(W/L)_1$ and $(W/L)_2$ with respect to $(W/L)_{REF}$? Recall from Chapter 2 that, owing to the side diffusion of the source/drain regions, the effective channel length is less than the drawn length by $2L_D$, a poorly controlled quantity. Thus, to avoid large mismatches, the lengths of the transistors must be equal and the currents must be scaled by proper choice of the widths. We then choose that $W_1 = 0.5W_{REF}$ and $W_2 = 2W_{REF}$. Figure 19.25 shows how M_{REF}, M_1, and M_2 in this example are laid out to ensure reasonable matching. Note that all equivalent widths are integer multiples of a unit value, W_u. Transistor M_1 is identical to M_{REF} except that half of its source remains floating (or connected to the drain). To improve the matching, the array can be surrounded by dummy devices.

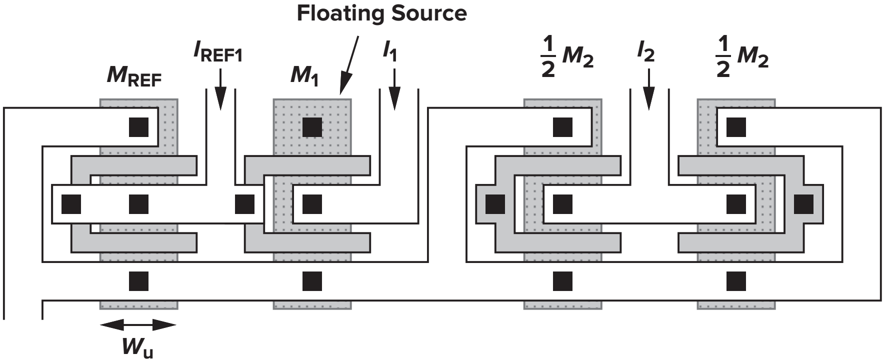


Figure 19.25 Proper scaling of device dimensions for adequate matching of current sources.

19.2.6 Passive Devices

Resistors Polysilicon resistors using a silicide block exhibit high linearity, low capacitance to the substrate, and relatively small mismatches. The linearity of these resistors in fact depends on their length [1], necessitating accurate measurement and modeling for high-precision applications. Figure 19.26 depicts an example in which the nonlinearity of the resistor is critical. Since $V_{out} = -I_{in}R_F$, the accuracy of current-to-voltage conversion depends on the linearity of R_F. In practice, however, the op amp limits the linearity.

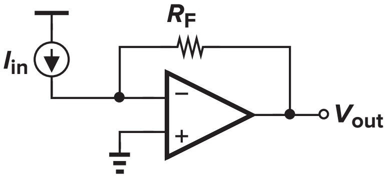


Figure 19.26 Feedback amplifier converting a voltage to current.

As with other devices, the matching of polysilicon resistors is a function of their dimensions. For example, resistors having a length and width of a few microns display typical mismatches on the order of 0.2%. Most of the symmetry rules described for the layout of MOS devices apply to resistors as well. For example, resistors that are required to bear a well-defined ratio must consist of identical units placed in parallel or in series (with the same orientation).

▶ **Example 19.2**

Consider the bandgap circuit shown in Fig. 19.27. Choose the values of n, R_1, and R_2 such that V_{out} exhibits a zero temperature coefficient and the layout can be designed for high precision.

Solution

Since $V_{out} = V_{BE3}+V_T(R_2/R_1)\ln n$, we must find convenient values of n, R_1, and R_2 such that $(R_2/R_1)\ln n \approx 17.2$ (Chapter 12). If $n = 31$, then $R_2/R_1 \approx 5$, yielding the layout of Fig. 19.28(a). Note that R_1 is placed in the middle to partially cancel the effect of gradients.

Now suppose we choose $n = 25$, obtaining $R_2/R_1 = 5.34$. Such a value cannot be accurately established by simply adjusting the dimensions of R_2 and R_1. Rather, we write $R_2/R_1 = 16/3$ and construct the resistors as shown in Fig. 19.28(b).

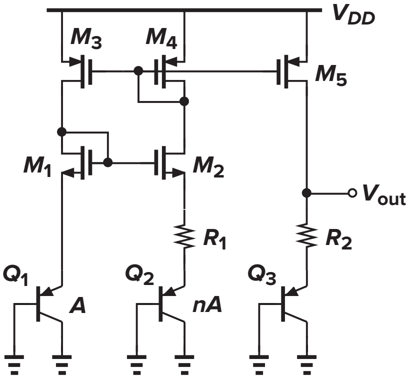


Figure 19.27

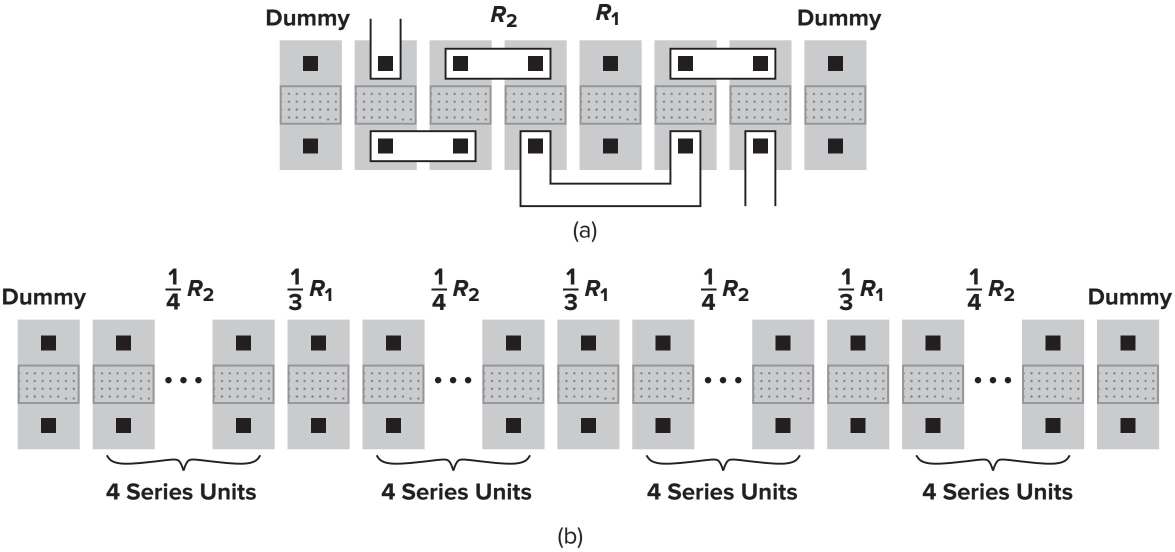


Figure 19.28 Layout of R_1 and R_2 with (a) $R_2/R_1 = 5$ and (b) $R_2/R_1 = 5.34 \approx 16/3$.

The resistance of the polysilicon structure studied above consists of two components: that due to the unsilicided region and the resistance associated with the two contacts. As depicted in Fig. 19.29(a), the

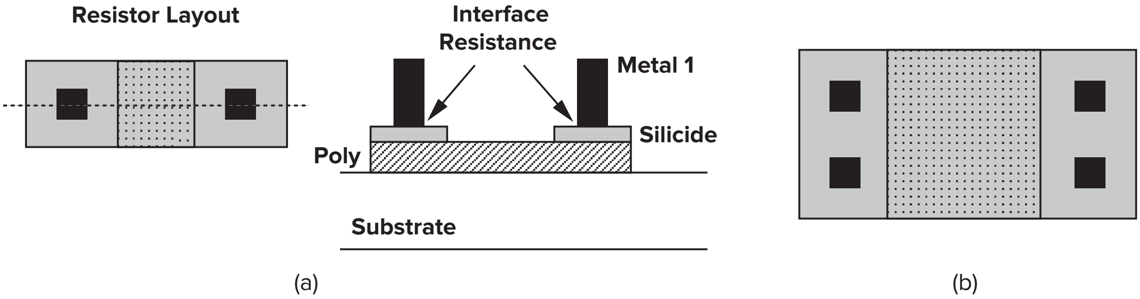


Figure 19.29 (a) Top view and cross section of a poly resistor, and (b) doubling the width and length to reduce the effect of contact resistance.

narrow contact window (about 80 nm×80 nm in 40-nm technology) results in a high interface resistance between metal 1 and the silicide area. This component is poorly controlled and must preferably remain much less than the first. For example, the length and width of the structure in Fig. 19.29(a) can be doubled so as to halve the total contact resistance while keeping the unsilicided region's resistance approximately constant [Fig. 19.29(b)].

For large values, resistors are usually decomposed into shorter units that are laid out in parallel and connected in series [Fig. 19.30(a)]. From the viewpoint of matching and reproducibility, this structure is preferable to "serpentine" topologies [Fig. 19.30(b)], where the corners contribute significant resistance.

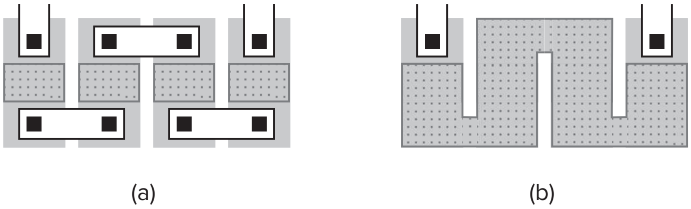


Figure 19.30 (a) Layout of large resistors; (b) serpentine topology.

The sheet resistance, $R_\square$, of polysilicon resistors varies with temperature and process, necessitating provisions in the design for this variation. The temperature coefficient depends on the doping type and level and must be measured for each technology. Typical values are $+0.1\%$ /°C and -0.1% /°C for p^+ and n^+ doping, respectively. The variation with process is usually less than ±20%.

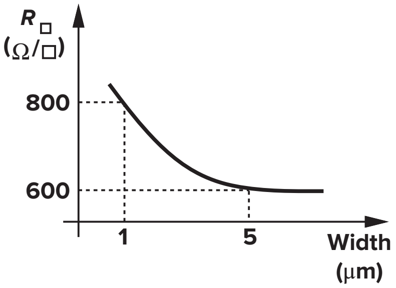


Figure 19.31 Dependence of n-well sheet resistance upon resistor width.

In technologies lacking a silicide block mask, resistors may be made of n-well, source/drain p^+ or n^+ material, silicided polysilicon, or metal, with $R_\square$ decreasing in this order. The sheet resistance of n-well is typically around 1 kΩ, but it may vary by a large fraction, e.g., ±40%, with process. Furthermore, $R_\square$ depends on the *width* of the resistor, as exemplified by the plot of Fig. 19.31. This is because, with a depth of several microns, n-well regions exhibit width-dependent diffusion at the edges. Also, $R_\square$ is a strong function of the n-well–substrate voltage difference, giving rise to both nonlinearity and poor definition of the value of the resistor. For example, in the circuit of Fig. 19.32, resistors R_S and R_D suffer from large mismatches in $R_\square$ because the depletion region below R_S is quite a lot narrower than that below R_D. Also, as V_{out} varies, so does the sheet resistance of R_D, introducing nonlinearity. Resistors made of n-well display a TC of $+0.2\%$ to $+0.5\%$ /°C.

▶ Example 19.3

An A/D converter incorporates a resistor ladder consisting of 128 units made of n-well to generate equally-spaced reference voltages (Fig. 19.33). If the two ends of the ladder are connected to $V_1 = +1$ V and $V_2 = +2$ V, calculate the ratio R_{128}/R_1.

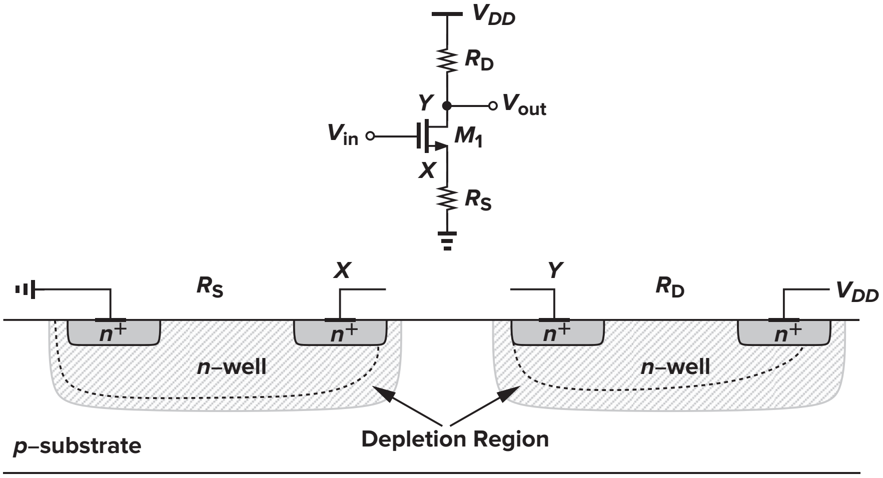


Figure 19.32 Common-source stage using n-well resistors.

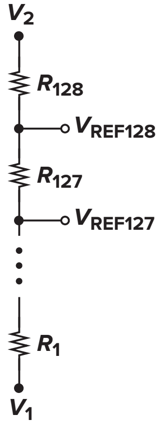


Figure 19.33 Resistor ladder used in an A/D converter.

Solution

The width of the depletion region inside the n-well is given by $x_d = \sqrt{2\epsilon_{si}(\phi_B + V_R)/(qN_{well})}$, where N_{well} denotes the n-well doping level and V_R the reverse bias voltage. Assuming that the zero-bias depth of the n-well is equal to t_0, we have

$$\frac{R_{128}}{R_1} = \frac{t_0 - \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}(\phi_B + V_1)} + \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}\phi_B}}{t_0 - \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}(\phi_B + V_2)} + \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}\phi_B}} \tag{19.10}$$

$$= \frac{t_0 + \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \dfrac{V_1}{\phi_B}}\right)}{t_0 + \sqrt{\dfrac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \dfrac{V_2}{\phi_B}}\right)} \tag{19.11}$$

If the difference between R_1 and R_{128} is small, we can divide the numerator and denominator of (19.11) by t_0 and approximate the result as

$$\frac{R_{128}}{R_1} \approx \left[1 + \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_1}{\phi_B}}\right)\right]\left[1 - \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(1 - \sqrt{1 + \frac{V_2}{\phi_B}}\right)\right] \tag{19.12}$$

$$\approx 1 + \frac{1}{t_0}\sqrt{\frac{2\epsilon_{si}}{qN_{well}}\phi_B}\left(\sqrt{1 + \frac{V_2}{\phi_B}} - \sqrt{1 + \frac{V_1}{\phi_B}}\right) \tag{19.13}$$

For example, if $t_0 = 2\ \mu$m, $N_{well} = 10^{16}\ \text{cm}^{-1}$, and $\phi_B = 0.7$ V, the mismatch between R_{128} and R_1 is nearly 60%.

The p^+ and n^+ source/drain regions can also be used as resistors. With a sheet resistance of 20 to 30 ohms per square, silicided S/D regions are suited only to low-value resistors. Furthermore, the junction between these areas and the bulk introduces capacitance and voltage dependence.[4]

Silicided polysilicon has a sheet resistance of 20 to 30 ohms per square and can be utilized for low resistor values. While suffering from less capacitance to the substrate than n^+ or p^+ resistors, silicided polysilicon has a process-dependent $R_\square$, with variations as high as 20 to 30%. Thus, it can be used only if its absolute value is not critical, for example, in the resistor ladder of Fig. 19.33. The temperature coefficient of this type of resistor is between $+0.2$ and $+0.4\%/^\circ$C.

The metal layers in a process can provide very low resistor values. For example, in extremely high-speed A/D converters, the ladder of Fig. 19.33 may be constructed as simply a long metal line having equally spaced taps (Fig. 19.34). Note, however, that if the width of the metal resistor is small, matching suffers. The temperature coefficient of the resistance is about 0.3%/°C for aluminum.

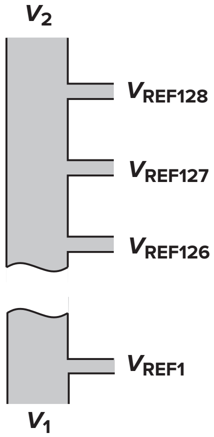


Figure 19.34 Resistor ladder made of metal.

Capacitors As explained in Chapter 18, linear capacitors are designed using sandwiches made of the available conductive layers. For example, in a process having nine layers of metal, the capacitors can be formed as shown in Fig. 19.35. The choice of one topology over another is determined by two factors: (1) the area occupied by the capacitor and (2) the ratio of the bottom-plate parasitic capacitance to the

[4] The nonlinearity of n-well resistors is much higher because the low doping level in the n-well results in a greater sensitivity to the voltage with respect to the substrate.

interplate capacitance, C_P/C. In typical technologies, the capacitance between consecutive metal layers (e.g., C_1 or C_2) in Fig. 19.35(d) is on the order of 35 to 40 aF/μm^2, and that between metal 1 and polysilicon is about 60 aF/μm^2. Thus, the structure of Fig. 19.35(d) provides more than nine times the density of that in Fig. 19.35(a). On the other hand, the value of C_P increases from Fig. 19.35(a) to Fig. 19.35(d). With typical values, C_P/C reaches a minimum—about 5 to 10%—for the structure of Fig. 19.35(b) or (c) and increases to about 20% for the sandwich of Fig. 19.35(d).

Since the absolute value of interlayer capacitances is poorly controlled in digital technologies, the capacitors of Fig. 19.35 may experience process variations as high as 20%. By contrast, the gate-oxide capacitance is typically controlled with less than 5% error. Interestingly, the structure of Fig. 19.35(d) may suffer from less variation than the others because random variations in the capacitances between various layers tend to "average out."

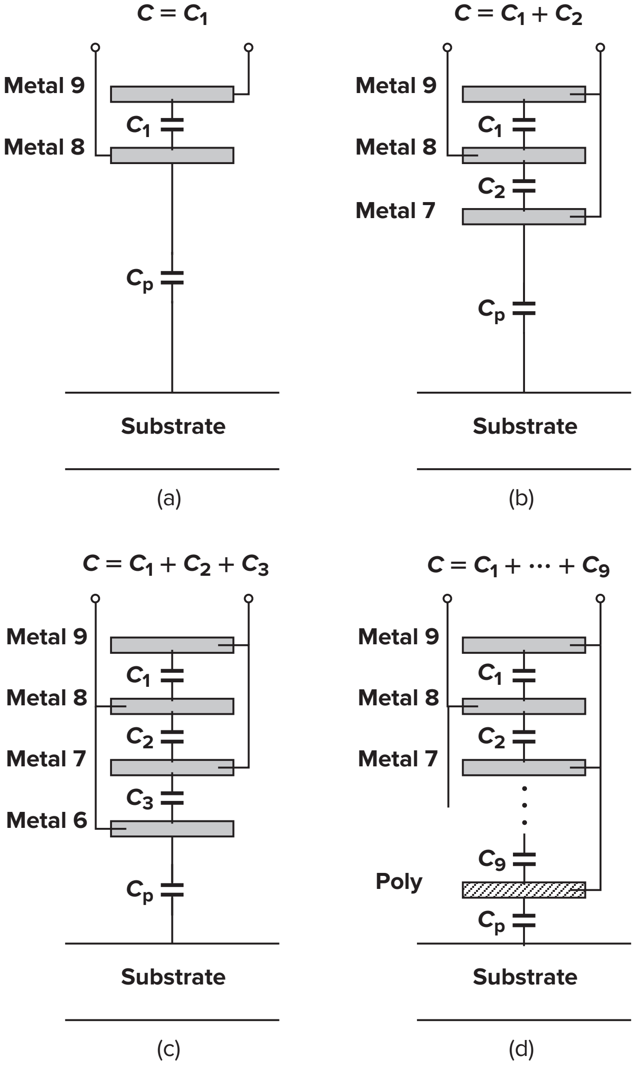


Figure 19.35 Capacitor structures using various conductive layers.

We have thus far neglected the fringe capacitance. As depicted in Fig. 19.36, the electric field lines emanating from the edge of each plate must terminate on the edge of the other plate or on the substrate,

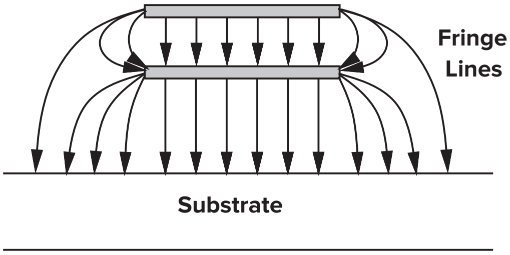


Figure 19.36 Fringe component of capacitance.

giving rise to a fringe capacitance that must be taken into account. The fringe capacitance can be calculated using Eq. (18.11) or from tabulated values in the process design manual.

As explained in Chapter 18, a MOS transistor with its source and drain tied together can act as a capacitor if the gate-source potential is sufficient to establish an inversion layer. However, the voltage dependence of the capacitance limits the use of this structure.

The layout of capacitors for high-precision circuits must follow the principles described earlier for transistors and resistors. For example, in applications where an array of well-matched capacitors is required, dummy devices must be placed on the perimeter of the array.

▶ Example 19.4

The circuit of Fig. 19.37(a) is designed for a nominal gain of $C_1/C_2 = 8$. How should C_1 and C_2 be laid out to ensure precise definition of the gain?

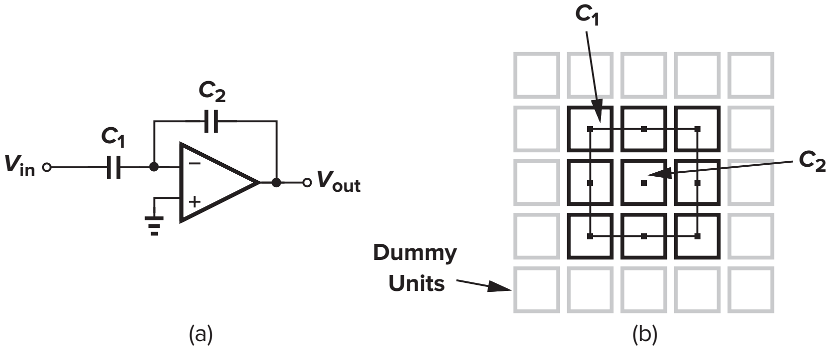


Figure 19.37

Solution

We form C_1 as 8 unit capacitors, each equal to C_2, and place all of the units in a square array [Fig. 19.37(b)]. Note that (1) C_2 is symmetrically surrounded by the units comprising C_1 so that the effect of vertical or horizontal gradients is canceled to the first order; and (2) dummy capacitor units are placed around the main array, creating approximately the same environment for the units of C_1 as that seen by C_2. ◀

For large capacitor arrays, cross-coupling techniques such as those illustrated in Figs. 19.20 and 19.27 can be applied. However, unlike transistors and resistors, capacitors are quite sensitive to the wiring capacitance, demanding great care in the interconnection of the units. Even in the simple array of Fig. 19.37(b), it is difficult to route all of the top-plate and bottom-plate connections while introducing no additional capacitance. As the layout of Fig. 19.38 exemplifies, the wiring inevitably leads to some error in the ratio C_1/C_2.

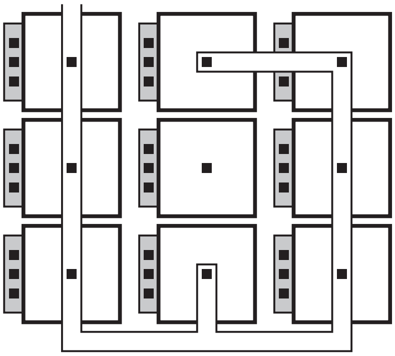


Figure 19.38 Layout of capacitors along with interconnections.

Diodes Two types of pn junctions can be formed in a standard CMOS technology: one in the p-substrate and another in an n-well (Fig. 19.39). The former must remain reverse biased and can therefore serve only as a voltage-dependent capacitor (varactor), e.g., in voltage-controlled oscillators.

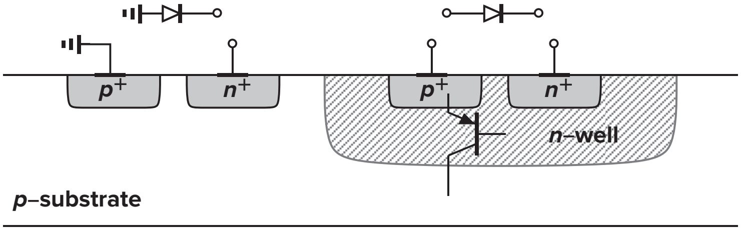


Figure 19.39 Diodes in CMOS technology.

The diode formed in an n-well also faces difficulties if forward biased. Recall from Chapter 12 that the p^+ region in the n-well, the n-well itself, and the p-substrate constitute a bipolar pnp transistor whose collector is typically grounded. Thus, if the pn junction in the n-well is forward biased, substantial current flows from the p^+ terminal to the substrate. In other words, the structure must not be viewed as merely a two-terminal floating diode. Nonetheless, if reverse-biased, the device can serve as a varactor.

Owing to these difficulties, analog CMOS circuits rarely incorporate forward-biased diodes (except in bandgap circuits).

19.2.7 Interconnects

Modern CMOS processes offer a dozen metal layers for interconnection, but the cost may dictate the use of eight or nine. Many effects related to wires must be taken into account when a high-precision and/or high-speed circuit is laid out.

Capacitance The parallel-plate and fringe capacitance of wires may degrade the speed if long interconnects are required. For example, in a mixed-signal system (e.g., using many switched-capacitor circuits), the clock signal must be distributed over long wires to access various building blocks, thereby experiencing significant line capacitance. More important, the capacitance between lines introduces substantial coupling of signals.

Figure 19.40 illustrates an example of cross talk between signals. Here, a common-source stage and a NAND gate are located next to each other, and the two inputs to the gate, V_A and V_B, cross over the analog signal, V_{in}. Furthermore, the clock wire, CK, is laid out in parallel with V_{in}, and the output of the NAND gate has some overlap with the output of the common-source stage. Each of the coupling capacitances in this layout may corrupt V_{in} or V_{out}. Note that, even though the coupling capacitances are

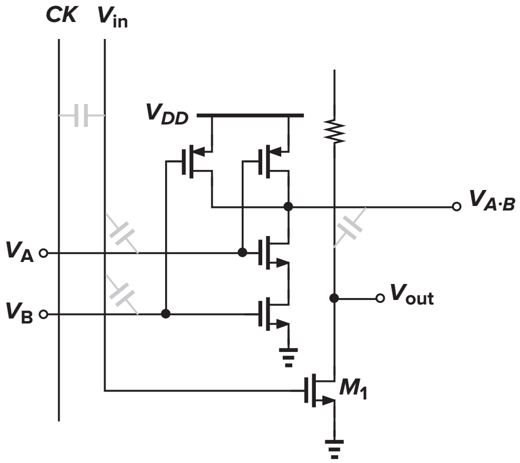


Figure 19.40 Capacitive coupling between various lines in a typical layout.

small, the signal corruption may be appreciable because the voltage swings on V_A, V_B, $V_{A \cdot B}$, and CK are large. For example, if the capacitance between CK and V_{in} is 50 aF, and the total capacitance seen from V_{in} to ground 10 fF, then a 1-V change in CK corrupts V_{in} by 5 mV.

Crosstalk can be reduced through the use of two techniques. First, differential signaling converts most of the crosstalk to common-mode disturbance. For example, if the circuit of Fig. 19.40 is modified to that shown in Fig. 19.41, the coupling of V_A and V_B to V_{in}^+ and V_{in}^- produces no differential error if $C_1 = C_1'$ and $C_2 = C_2'$. Even for 10% mismatch between the capacitances, the differential corruption is one order of magnitude less than that in Fig. 19.40. Note that a dummy wire is added to the layout so as to create an overlap capacitance between CK and V_{in}^- equal to that between CK and V_{in}^+. As mentioned in Chapter 4, it is desirable to employ differential clocks as well to suppress the net coupling further.

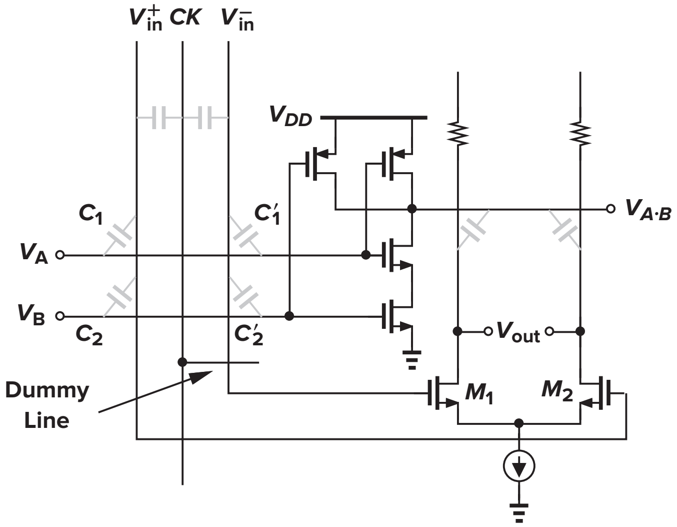


Figure 19.41 Reduction of capacitive coupling through the use of differential signaling.

Second, sensitive signals can be "shielded" in the layout. Depicted in Fig. 19.42(a), one approach places ground lines on the two sides of the signal, forcing most of the electric field lines emanating from the "noisy" lines to terminate on ground rather than on the signal. Note that this method proves more

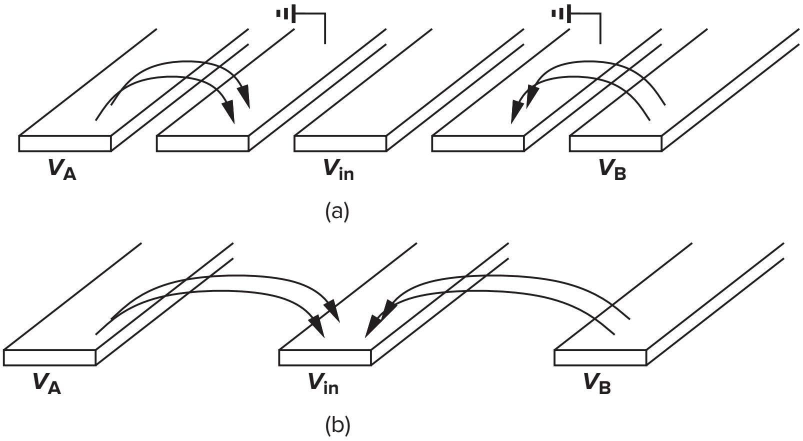


Figure 19.42 (a) Shielding sensitive signals by additional ground lines; (b) greater spacing between lines to reduce coupling.

effective than simply allowing more space between the signal and the noisy lines [Fig. 19.42(b)]. The shielding, however, is obtained at the cost of more complex wiring and greater capacitance between the signals and ground.

Another shielding technique is shown in Fig. 19.43. Here, the sensitive line in metal 6 is surrounded by a grounded shield consisting of a higher and a lower metal layer and hence fully isolated from external electric field lines.[5] However, the signal experiences higher capacitance to ground, and the use of three metal layers here complicates the routing of other signals.

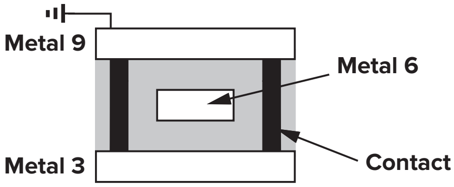


Figure 19.43 Shielding a sensitive line (metal 2) by lower and upper ground planes.

Resistance The resistance of interconnects also requires attention. In low-noise applications, long signal wires—with sheet resistances of 40 to 80 mΩ/□—may introduce substantial thermal noise. Furthermore, the contacts and vias also suffer from a high resistance. For example, an 80-nm × 80-nm metal contact to silicided polysilicon exhibits a resistance of 30 to 40 Ω, and a via between metal 1 and metal 2, a resistance of 5 to 10 Ω.

▶ **Example 19.5**

In the layout of Fig. 19.44, a 100-μm metal 4 line is connected to a sequence of vias and contacts to reach the gate of a transistor. Calculate the thermal noise contributed by the line and the contacts.

Solution

Assuming $R_\square = 40$ mΩ/□ for metal 4, a via resistance of 5 Ω, and a poly contact resistance of 30 Ω, we have $R_{tot} = 2 + 2.5 + 2.5 + 2.5 + 15 = 24.5$ Ω. The thermal noise voltage is thus equal to 0.64 nV/$\sqrt{\text{Hz}}$ at room

[5] We assume that the ground connection itself does not contain noise. We return to this issue in Sec. 19.4.

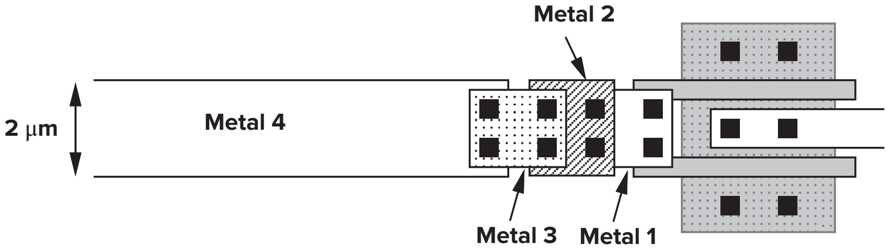


Figure 19.44

temperature. If guiding the input signal to a low-noise amplifier, this interconnect arrangement considerably raises the input-referred noise.

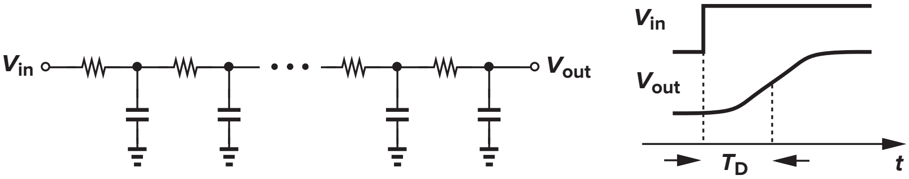


Figure 19.45 Delay and dispersion of a signal in a long line.

The distributed resistance and capacitance of long interconnects may introduce significant delay and "dispersion" in signals. Illustrated in Fig. 19.45, the delay can be approximated as

$$T_D = \frac{1}{2} R_u C_u L^2 \tag{19.14}$$

where R_u and C_u denote the resistance and capacitance per unit length, respectively, and L is the total length. For example, consider the circuit shown in Fig. 19.46, where an array of samplers senses the analog input V_{in} and is activated by CK. If the delays experienced by CK and V_{in} from the left side to the right side are unequal, then so are the levels sampled by $C_1, \ldots, C_n$, distorting the sampled waveform. Even if the clock and signal lines and their capacitive loading are identical, CK and V_{in} may still suffer from unequal delays because the former is a rectangular wave and the latter is not.

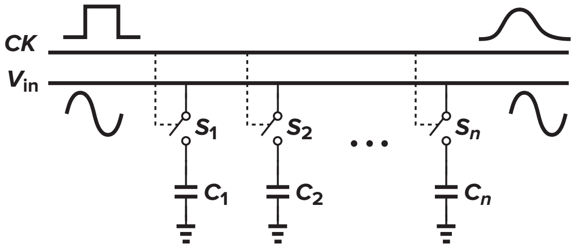


Figure 19.46 An array of sampling circuits sensing an input.

The term "dispersion" refers to the significant increase in the transition time of the signal as it propagates through a line, a particularly troublesome effect if a clock edge is to define a sampling point. In the example of Fig. 19.46, the clock waveform applied to S_n displays long rise and fall times, making the sampling susceptible to both noise and distortion [4]. The clock edges can be sharpened by inserting an inverter between CK and every switch, but at the cost of greater uncertainty in the delay difference between CK and V_{in}.

As mentioned in Chapter 18, the design of power and ground buses on a chip requires attention to a number of issues. In large ICs, the dc or transient voltage drop along the buses may be significant, affecting sensitive circuits supplied by the same lines. Furthermore, electromigration calls for a minimum line width to guarantee long-term reliability. With the multiple interconnect levels available in today's CMOS technology, it is possible to connect two or more layers in parallel, thereby reducing the series resistance and alleviating electromigration constraints. Since the thickness of the top metal layer is typically twice that of the lower ones, at least *three* layers must be placed in parallel to relax these issues by a factor of two. As a result, routing signals and bias lines across the buses may become difficult if only one or two more layers of metal are available.

If the bias currents drawn from a long bus are relatively well defined, then the bus width can be "tapered" from one end to the other so as to create a relatively constant voltage drop along the line. Illustrated in Fig. 19.47, this technique can be used if the metal resistance and its temperature coefficient are known.

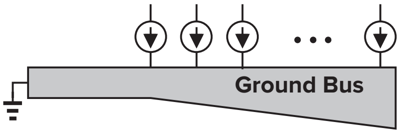


Figure 19.47 Tapered ground line for reduction of voltage drops.

19.2.8 Pads and ESD Protection

The interface between an integrated circuit and the external environment involves a number of important issues. In order to attach bond wires to the die, large "pads" are placed on the perimeter of the chip and connected to the corresponding nodes in the circuit (Fig. 19.48).

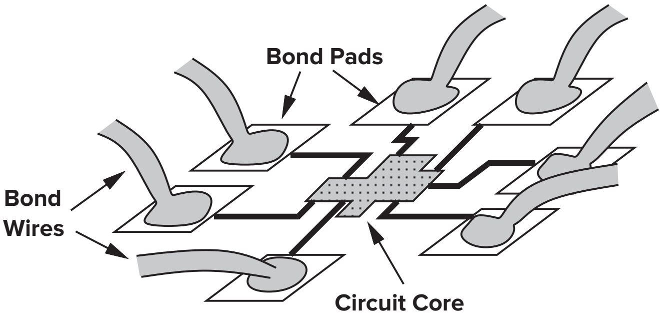


Figure 19.48 Addition of bonding pads to a chip.

The pad dimensions and structure are dictated by the reliability issues and margin for manufacturing tolerances in the wire-bonding process. With bond wire diameters ranging from 25 μm to 50 μm, the minimum pad size falls between roughly 70 μm × 70 μm and 100 μm × 100 μm. Adjacent pads are usually separated by at least 25 μm. From the circuit design point of view, the pad dimensions must be minimized so as to reduce both the capacitance of the pad to the substrate and the total die area.

A simple pad would consist of only a square made of the top metal layer. However, such a structure is susceptible to "lift-off" during bonding. For this reason, each pad is typically formed by the two topmost metal layers, connected to each other by many small vias on the perimeter (Fig. 19.49). Note that this structure suffers from a larger capacitance to the substrate than a pad made of only the top layer.

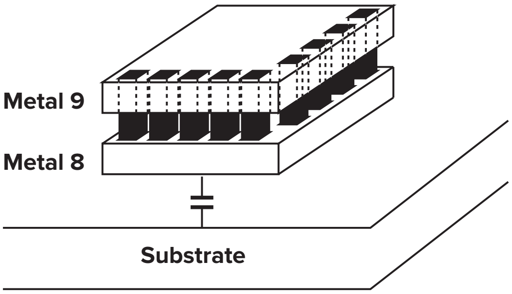


Figure 19.49 Structure of a typical bonding pad.

▶ **Example 19.6**

Calculate the capacitance of a metal 4 pad and a metal 4/metal 3 pad. Assume dimensions of 75 μm × 75 μm and use the capacitance data shown in Fig. 19.50.

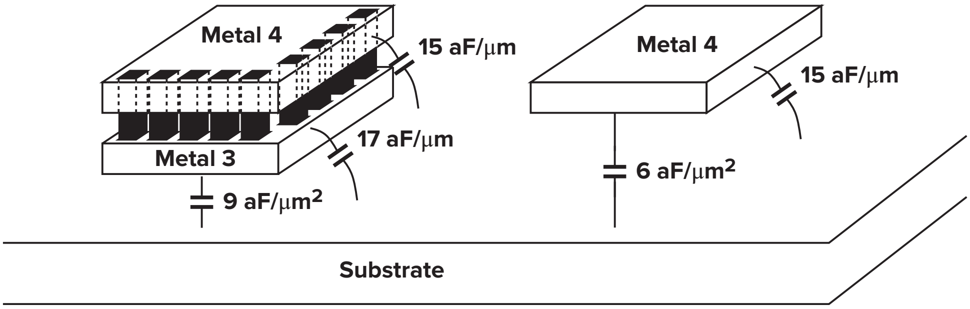


Figure 19.50

Solution

For a metal 4 pad,

$$C_{tot} = 75^2 \times 6 + 75 \times 4 \times 15 \tag{19.15}$$

$$= 38.25 \text{ fF} \tag{19.16}$$

For a metal 4/metal 3 pad,

$$C_{tot} = 75^2 \times 9 + 75 \times 4 \times (17 + 15) \tag{19.17}$$

$$= 60.22 \text{ fF} \tag{19.18}$$

Note that the fringe capacitances of metal 4 and metal 3 are directly added here. This is a rough approximation. ◀

The pads carrying high-frequency signals can be configured as octagons so as to reduce their capacitance. Depicted in Fig. 19.51, such a structure is obtained by removing the corner areas of a square pad—without making the task of bonding more difficult. If $a = b$, then both the area and the perimeter of the pad fall by about 20%.

The interface between an IC and the external world also entails the problem of electrostatic discharge (ESD). This effect occurs when an external object having a high potential touches one of the connections to the circuit. Since the capacitance seen at each input or output is small, the ESD produces a large voltage, damaging the devices fabricated on the chip.

A common case of ESD arises when ICs are handled by human beings. For this effect, the human body can be modeled by a capacitance of a few hundred picofarads in series with a resistance of a few kilohms.

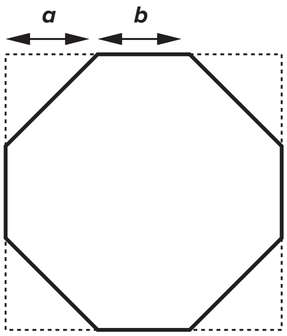


Figure 19.51 Use of octagonal pad to reduce capacitance.

Depending on the environment, the voltage across the capacitance ranges from a few hundred volts to several thousand volts. Thus, if a person touches a line connecting to the chip, the chip is easily damaged. Interestingly, electrostatic discharge may occur even without actual contact because at high electric fields, the person's finger "arcs" to the connection through the air if the finger is sufficiently close to the line.

It is important to note that ESD may occur even without human intervention. If not properly grounded, various objects in a typical chip assembly line accumulate charge, rising to high potential levels. Furthermore, charge in dry air may create substantial potential gradients with respect to ground.

MOS devices sustain two types of permanent damage as a result of ESD. First, the gate oxide may break down if the electric field exceeds roughly 10^7 V/cm (e.g., 2 V for an oxide thickness of 20 Å), typically leading to a very low resistance between the gate and the channel. Second, the source/drain junction diodes may melt if they carry a large current in forward or reverse bias, creating a short to the bulk. For today's short-channel devices, both of these phenomena are likely to occur.

In order to alleviate the problem of electrostatic discharge, CMOS circuits incorporate ESD protection devices. Illustrated in Fig. 19.52, such devices clamp the external discharge to ground or V_{DD}, thereby limiting the potential applied to the circuit. Resistor R_1 is usually necessary so as to avoid damaging D_1 or D_2 due to large currents that would otherwise flow from the external source.

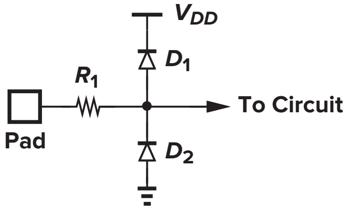


Figure 19.52 Simple ESD protection circuit.

The use of ESD protection structures involves three critical issues. First, the devices introduce substantial capacitances from the node to ground and V_{DD}, degrading the speed and the matching of impedances at the input and output ports of the circuit. Since the protection devices, such as D_1 and D_2 in Fig. 19.52, must be large enough so that the chip sustains a high ESD voltage without damage, their capacitance may reach several picofarads. The thermal noise of R_1 may also become significant.

Second, the parasitic capacitance of the ESD devices may couple noise on V_{DD} to the input of the circuit, corrupting the signal. We return to this issue in Sec. 19.4.

Third, if not properly designed, ESD structures may lead to latch-up in CMOS circuits when electrostatic discharge occurs during actual circuit operation (or even when the circuit is turned on). For this reason, process engineers fabricate and characterize many different ESD structures for each generation of a technology, eventually providing a few reliable configurations that can be used in circuits.[6]

[6] In general, a circuit designer should not use an ESD structure that has not been tested and qualified for the technology. Uncharacterized ESD devices are likely to cause latch-up.

19.3 ■ Substrate Coupling

Most modern CMOS technologies use a heavily-doped p^+ substrate to minimize latch-up susceptibility. However, the low resistivity of the substrate (on the order of 0.1 $\Omega\cdot$cm) creates unwanted paths between various devices in the circuit, thereby corrupting sensitive signals. Called "substrate coupling" or "substrate noise," this effect has become a serious issue in today's mixed-signal ICs [2].

To understand this phenomenon, suppose a CMOS inverter sensing a clock is laid out next to a common-source stage amplifying an analog signal [Fig. 19.53(a)]. Note that the substrate is connected

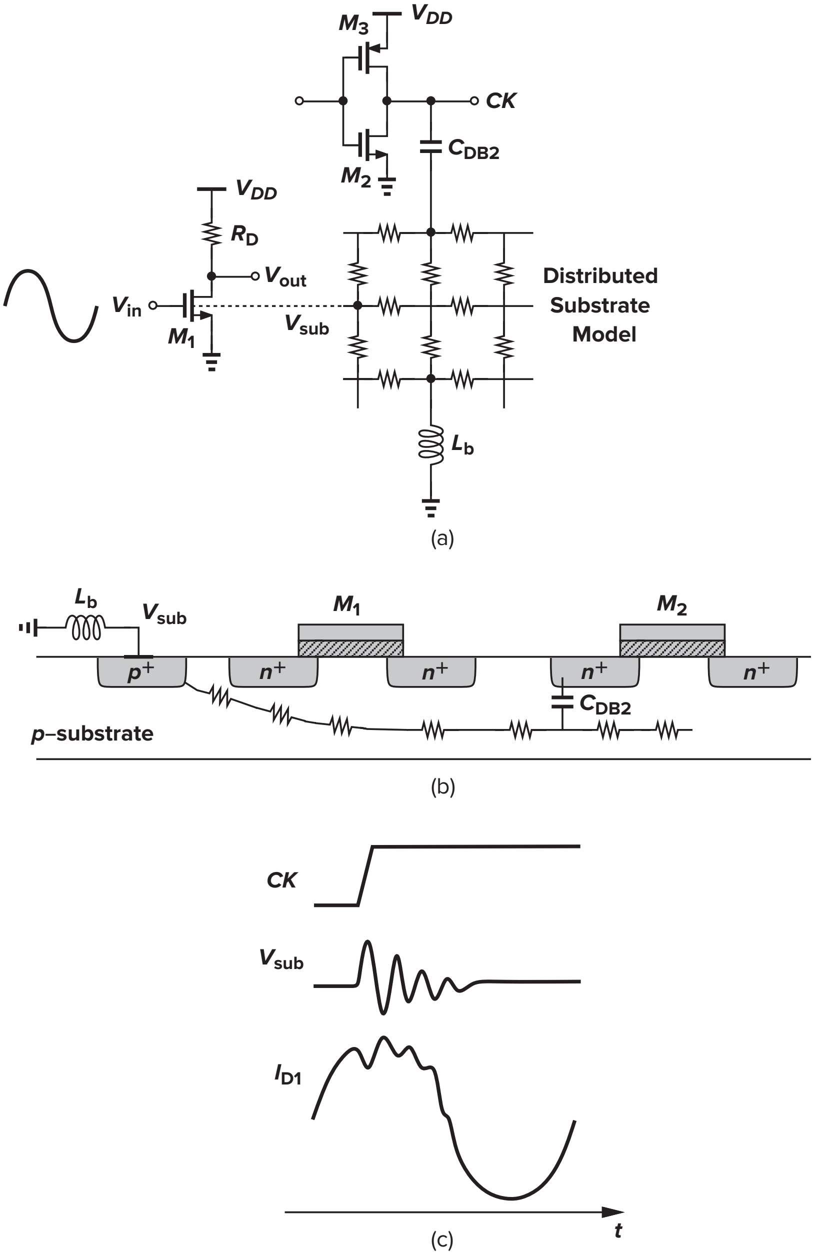


Figure 19.53 (a) Mixed-signal circuit including the effect of substrate coupling; (b) side view of device layout; (c) signal waveforms.

to ground through a bond wire that exhibits an (unwanted) inductance of L_b. With the aid of the cross section depicted in Fig. 19.53(b), we observe that the large voltage excursions at the drain of M_2 are coupled to the substrate through the drain junction capacitance, disturbing the substrate voltage because of the finite impedance of L_b.

How does the substrate noise influence M_1? The principal coupling mechanism here occurs through body effect, varying the threshold voltage of M_1 with the substrate voltage. Since the drain current of M_1 depends on $V_{in} - V_{TH1}$, variations in V_{TH1} are indistinguishable from those in V_{in}. In other words, as illustrated in Fig. 19.53(c), every transition of CK disturbs the analog output.

The problem of substrate coupling becomes more noticeable as the number of "noise" generators increases. In a mixed-signal environment, thousands of digital gates may inject noise into the substrate—especially during clock transitions—introducing hundreds of millivolts of disturbance in the substrate potential. The disturbance is also proportional to the size of the noise-injecting devices, an important issue if large transistors are used as buffers driving heavy external loads.

It may seem that substrate coupling can be decreased by increasing the physical spacing between sensitive building blocks and digital sections of a chip. In practice, however, this remedy may not be effective or feasible. If heavily doped, the substrate operates as a low-resistance plane, distributing a relatively uniform potential across the chip regardless of the position of the noise generators [3]. Furthermore, in many mixed-signal systems, the analog and digital functions are so heavily blended that it is difficult to separate their corresponding circuits. Figure 19.54 shows a slice of an A/D converter consisting of a comparator, a flipflop, a NAND gate, and a read-only memory (ROM). Various logical swings in the comparator and the digital circuits generate substrate noise, but increasing the distance between any two blocks necessitates long interconnects, degrading the performance.

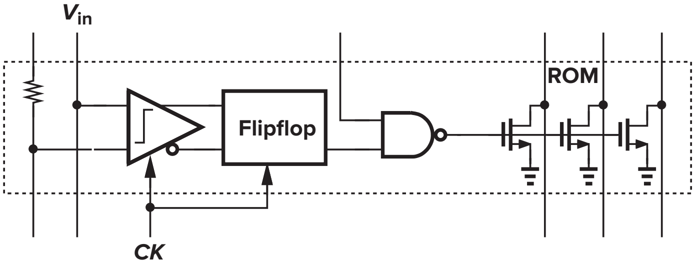


Figure 19.54 A slice of an A/D converter.

In order to minimize the effect of substrate noise, the following methods can be applied. First, differential operation should be used throughout the circuit, making the analog section less sensitive to common-mode noise. Second, digital signals and clocks should be distributed in complementary form, thereby reducing the net amount of the coupled noise. Third, critical operations, e.g., sampling a signal or transferring charge from one capacitance to another, should be performed well after clock transitions so that the substrate voltage settles. Fourth, the inductance of the bond wire connected to the substrate should be minimized (Sec. 19.4). Also, op amps using a PMOS differential input are preferred because the well of the transistors can be tied to their common source, reducing the effect of substrate noise.

In circuits fabricated on lightly-doped substrates, "guard rings" can be employed to isolate the sensitive sections from the substrate noise produced by other sections. A guard ring may be simply a continuous ring made of substrate ties that surrounds the circuit, providing a low-impedance path to ground for the charge carriers produced in the substrate. With its large depth, the n-well can also augment the operation of a guard ring by stopping the noise currents flowing near the surface (Fig. 19.55).

In large mixed-signal ICs, it may not be possible to avoid substrate "bounce" with respect to the external ground because of the high transient currents drawn by the devices and the finite impedance of

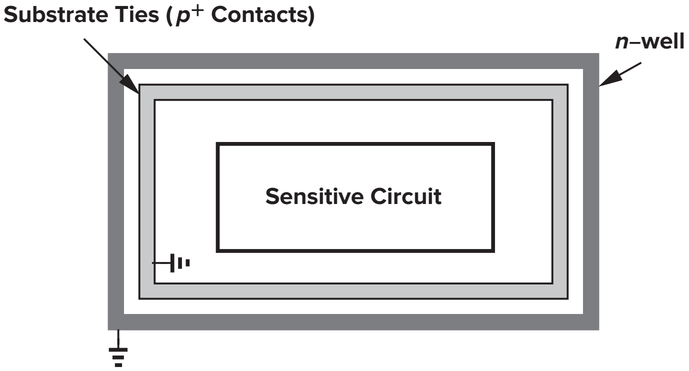


Figure 19.55 Use of guard ring to protect sensitive circuits.

the bond wire connected to the substrate. However, we recognize that if the ground of the chip bounces in unison with the substrate, then the transistors experience no noise. Illustrated in Fig. 19.56, this idea suggests that the ground and the substrate should be connected on the chip and brought out through a single wire.

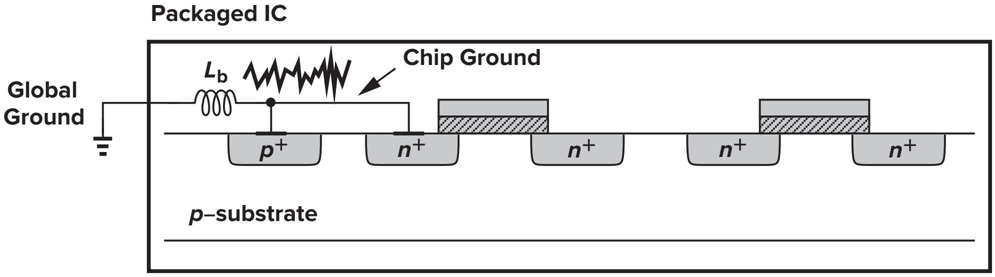


Figure 19.56 Substrate bounce.

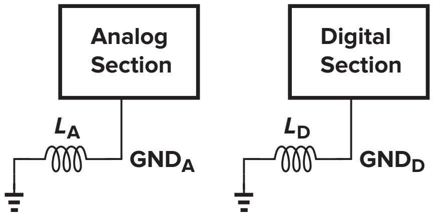


Figure 19.57 Analog and digital grounds.

The connection of the substrate to the chip ground nonetheless faces two difficulties. The first relates to "ground bounce." As shown in Fig. 19.57 and explained in Sec. 19.4, most mixed-signal circuits employ at least one "analog ground" and one "digital ground" so as to avoid corrupting the analog section by the large transient noise produced by the digital section. To which ground should the substrate be connected? If the analog ground is used, then the large substrate noise current must flow through L_A, creating noise on GND$_A$ [Fig. 19.58(a)], and if the digital ground is used, then the substrate voltage is heavily disturbed by the large noise on GND$_D$ [Fig. 19.58(b)]. Of course, connecting the substrate to both GND$_A$ and GND$_D$ gives rise to a low-resistance path between the two, defeating the purpose of separating the analog and digital grounds.

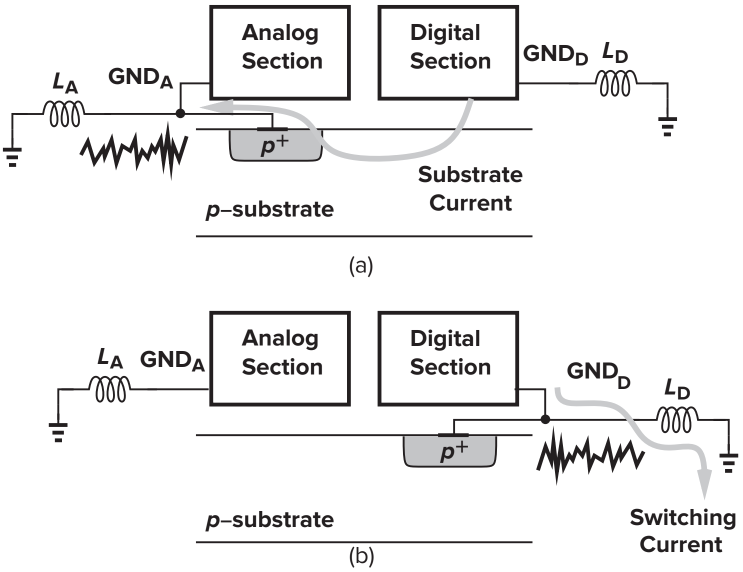


Figure 19.58 Connection of substrate contact to (a) analog ground and (b) digital ground.

The choice between the configurations shown in Figs. 19.58(a) and (b) depends on the transient currents drawn by the digital section from the substrate and the ground as well as the magnitudes of L_A and L_D. In most cases, the topology of Fig. 19.58(a) is preferred because it ensures that the analog ground voltage and the substrate potential vary in unison. As illustrated in Fig. 19.59(a), if the analog ground and the substrate experience unequal bounce, then the drain current of M_1 is corrupted by the substrate noise. The configuration of Fig. 19.59(b), on the other hand, introduces less noise in I_{D1}. In general, careful, realistic simulations of the overall environment (including the package) are necessary to determine which approach yields less noise.

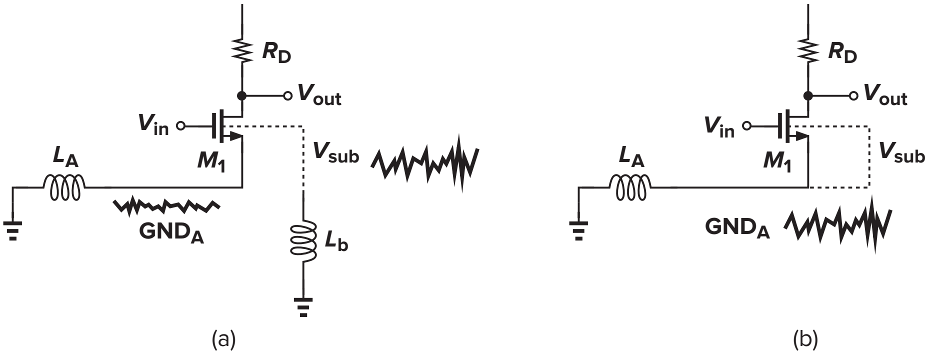


Figure 19.59 (a) Large source-bulk noise voltage due to separating substrate contact from analog ground; (b) suppression of the effect.

The second issue in allowing the substrate and a chip ground to bounce together is the difficulty in defining a reference potential for the input signals. As shown in Fig. 19.60(a), a single-ended input is heavily corrupted as its reference point changes from the off-chip ground to the on-chip ground. That is, $V'_{in} \neq V_{in}$, even though the substrate and the ground bounce in unison. For the differential structure of Fig. 19.60(b), the effect is much less pronounced, but in high-precision applications, asymmetries in the circuit and interconnections convert a fraction of the common-mode noise to a differential component.

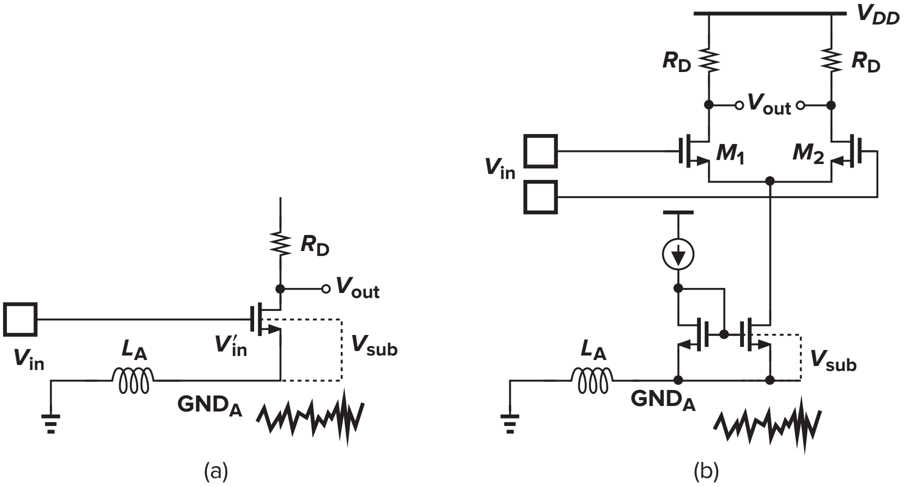


Figure 19.60 (a) Input signal corruption due to ground and substrate bounce; (b) less corruption in a differential environment.

19.4 ■ Packaging

After fabrication and dicing, integrated circuits are packaged. The parasitics associated with the package and connections to the chip introduce many difficulties in the evaluation of the actual performance of the circuit at high speeds and/or high accuracies.

Let us first consider a simple dual-in-line package (DIP) [Fig. 19.61(a)]. Here, the die is mounted in the center cavity and bonded to the pads on the perimeter of the cavity. These pads are in fact the tip of each trace that ends in each package pin. Such a structure exhibits the following parasitics: bond wire self-inductance, trace self-inductance, trace-to-ground capacitance, trace-to-trace mutual inductance, and trace-to-trace capacitance. Thus, as shown in Fig. 19.61(b), the connections between the circuit and the external world are far from ideal.

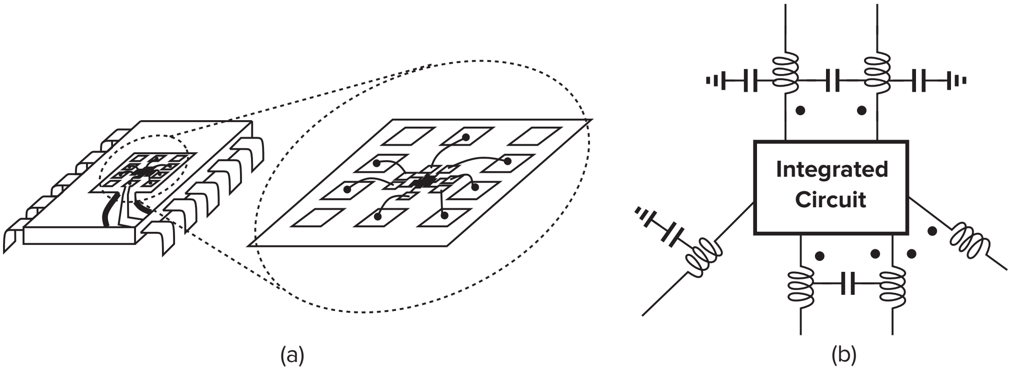


Figure 19.61 (a) Dual-in-line package; (b) electrical model of the package.

While, owing to both circuit innovations and device scaling, the speed and accuracy of integrated circuits have steadily increased, the performance of packages, especially for low-cost applications, has not improved significantly. This limitation originates from the unscalable nature of packages and the environment in which they are used. For example, the diameter of the bond wires, the width and spacing of package pins, and the width and spacing of the traces in printed circuit (PC) boards are determined by mechanical stress, ease and cost of assembly, series resistance at high frequencies (skin effect), etc. In the past 20 years, these dimensions have scaled by less than a factor of five, whereas the speed of many mixed-signal circuits has increased by two orders of magnitude. As a result, packaging continues to limit the achievable performance of today's high-performance ICs.

The foregoing issues dictate that the package parasitics be taken into account in the design of integrated circuits—sometimes from the very beginning. Thus, simulations must include a reasonable circuit model of the package, and the design and layout must take many measures to minimize the effect of package parasitics.

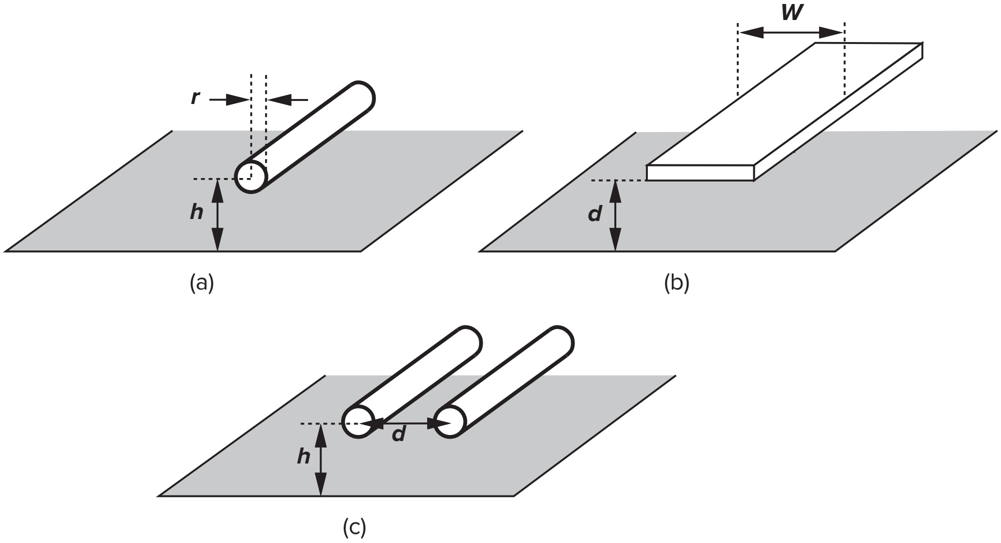


Figure 19.62 Common geometries in packaging.

Since some package manufacturers do not provide circuit models for their products, IC designers often develop the models themselves by calculations and measurements. Figure 19.62 depicts three common cases of self- and mutual inductance. From [6], we have for a round wire above a ground plane [Fig. 19.62(a)]

$$L \approx 0.2 \ln \frac{2h}{r} \;\; \text{nH/mm} \tag{19.19}$$

which amounts to roughly 1 nH/mm for typical bond wires. For a flat trace above a ground plane [Fig. 19.62(b)],

$$L \approx \frac{1.6}{K_f} \cdot \frac{d}{W} \;\; \text{nH/mm} \tag{19.20}$$

where K_f denotes the fringe factor and from the data in [6] can be approximated as $0.72(d/W) + 1$. For two round wires above a ground plane, the mutual inductance is [6]

$$L_m = 0.1 \ln \left[1 + \left(\frac{2h}{d} \right)^2 \right] \;\; \text{nH/mm} \tag{19.21}$$

The parasitic capacitances can be calculated with the simple interplate equation and Eq. (18.11). In practice, the bond wires are not simply straight, parallel lines, requiring electromagnetic field simulations for proper modeling.

Let us now study the effect of each type of package parasitic. We categorize the connections to the chip into five groups: power and ground lines, analog and clock inputs, outputs, reference lines, and substrate connection(s).

Self-Inductance Each bond wire and its corresponding package trace exhibit a finite self-inductance, with a total value between approximately 2 nH and 20 nH depending on the length of the wire and the type of the package. To understand how the self-inductance of supply and ground lines affects the performance, suppose a mixed-signal circuit incorporates a CMOS inverter as a clock buffer to drive a moderate on-chip capacitance, e.g., 0.5 pF (Fig. 19.63). Also, assume that the buffered clock must have transition times less than 0.5 ns, thereby demanding a current of $C\Delta V/\Delta t = 3$ mA. Since this current is drawn from V_{DD1} and GND$_1$ in 0.5 ns, we can estimate the voltage drop across L_D or L_G as[7] $L\Delta I/\Delta t = 6 \times 10^6 L$. For example, if $L_D = L_G = 5$ nH, then the transient voltage across each inductor equals 30 mV. This effect is called supply and ground "bounce" or "noise." Note that if the inverter is replaced by a differential pair, the supply bounce decreases substantially (why?), another advantange of differential operation.

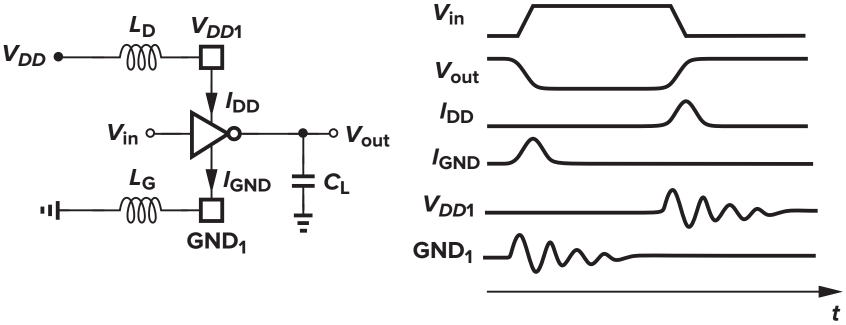


Figure 19.63 CMOS inverter driving a load capacitance.

A supply noise of 30 mV may seem quite benign, especially if the analog circuits feeding from the same supply line are fully differential. However, in a typical mixed-signal IC, hundreds or thousands of digital gates may switch during each clock transition, creating enormous noise on their supply and ground connections. For this reason, most such systems employ separate supply and ground lines for the analog and digital sections; hence the terminology "analog supply" and "digital supply."

Separating power lines into analog and digital groups is not always straightforward. As an example, suppose a sampling circuit is clocked by an inverter (Fig. 19.64). Should the inverter be supplied from analog or digital power lines? If the inverter is connected to the digital supply, then the large noise on V_{DD} couples through the gate-drain overlap capacitance of M_1, corrupting V_{out} when the transistor is off. On the other hand, if many such inverters are supplied from the analog V_{DD}, they collectively draw large transient currents, corrupting the supply voltage. These cases may require a third type of power line so that it remains less noisy than the digital supplies.

For characterization and troubleshooting purposes, it is sometimes desirable to monitor the supply noise. Figure 19.65 illustrates a simple method whereby a PMOS device sensing the noise between the on-chip supply and ground lines injects a current into an external 50-Ω transmission line and measurement

[7] This calculation is quite rough because the current produced by the buffer varies during the transition.

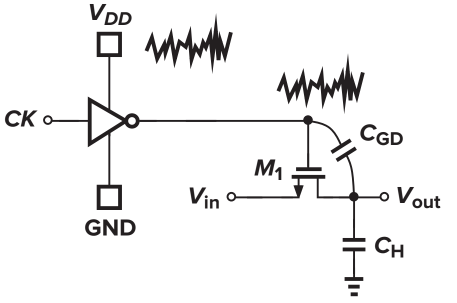


Figure 19.64 Noise in a sampling circuit resulting from the clock buffer's supply bounce.

apparatus [2]. Since the transconductance of M_1 can be determined by a small, static change in V_{DD}, the measurement readily reveals both the magnitude and the shape of the supply noise.

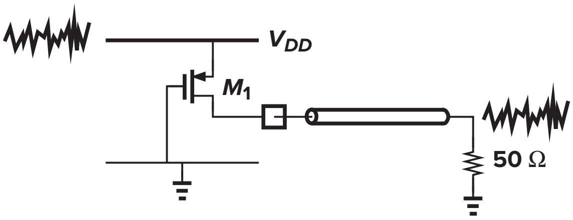


Figure 19.65 Measurement of supply noise.

In cases where a single connection to the chip sustains a prohibitively large transient voltage (e.g., if in Fig. 19.63 or 19.64 many inverters switch simultaneously), multiple pads, bond wires, and package pins are used so as to reduce the equivalent inductance (Fig. 19.66).

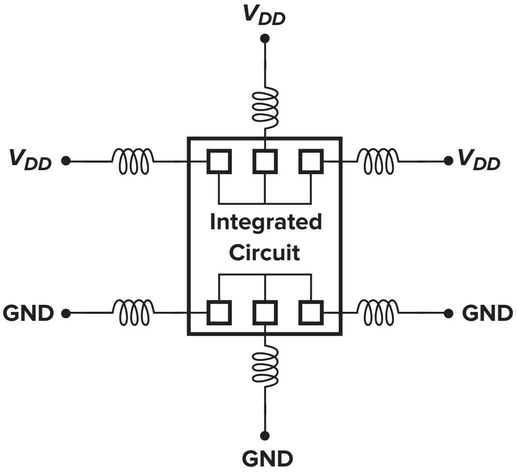


Figure 19.66 Use of multiple wires to reduce overall inductance.

▶ **Example 19.7**

In a 600-MHz, 2-V CMOS microprocessor containing 15 million transistors, the supply current varies by 25 A in approximately 5 ns [5]. If the processor provides 200 bond wires for ground and 200 for V_{DD}, estimate the resulting supply bounce.

Solution

Assuming a total inductance of 5 nH for each bond wire and its corresponding package trace and pin, we have

$$\Delta V = L\frac{\Delta I}{\Delta t} \tag{19.22}$$

$$= \frac{5 \times 10^{-9}}{200} \cdot \frac{25}{5 \times 10^{-9}} \tag{19.23}$$

$$= 125 \text{ mV} \tag{19.24}$$

In the worst case, the supply bounce and the ground bounce add in-phase, yielding a total noise of roughly 250 mV, greater than 10% of the nominal supply voltage. To further suppress the noise, an external 1-μF MOS capacitor is placed on top of the chip and another 160 supply and ground bond wire pairs are connected from the chip to the capacitor [5].

In some applications, high transient currents drawn from the supply make it difficult to maintain a small bounce on the supply and ground individually. In such cases, a large on-chip capacitor may be used to stabilize the *difference* between V_{DD} and ground. Illustrated in Fig. 19.67, the idea is that if C_1 is sufficiently large, then V_{DD1} and GND$_1$ bounce in unison. As mentioned earlier, the residual noise on GND$_1$ may be negligible if the input signals are differential.

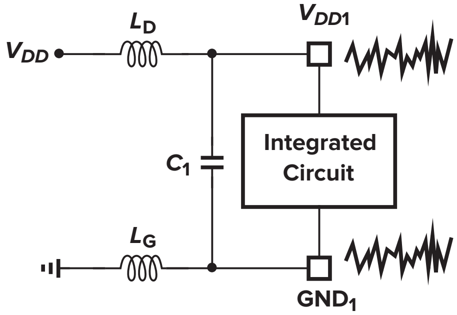


Figure 19.67 On-chip capacitor used to lower supply-ground noise voltage.

This remedy nonetheless involves several issues. First, the value of the capacitor must be chosen carefully because it may otherwise *resonate* with the package inductance at the operating frequency of the chip (e.g., the clock frequency or its harmonics or subharmonics), thereby *amplifying* the supply and ground noise. For this reason, some resistance is added in series with the capacitor (or a MOS capacitor is sized such that its channel resistance dampens the resonance) [5]. Even in the absence of exact resonance, an insufficient value of the decoupling capacitor may simply give rise to slower ringing on the power lines. Second, since the capacitor is usually formed by a very large MOS transistor (actually, as explained in Sec. 18.7.2, a large number of MOSFETs in parallel), the yield of the circuit may suffer. This is because, for the capacitor to be effective, its total area is typically comparable with the total gate area of all of the transistors in the circuit; e.g., it is as if the number of transistors on the chip were doubled.

Self-inductance also manifests itself in the connection to the substrate. As mentioned in Sec. 19.3, with the large transient currents injected by the devices into the substrate, a low-impedance connection is necessary to minimize the substrate bounce. As shown in Fig. 19.68, some modern packages contain a metal ground plane to which the die can be attached by conductive epoxy. The plane ends in several package pins that are tied to the board ground. Avoiding bond wires and long, narrow traces in the substrate connection, such packages substantially reduce the substrate noise with no additional assembly cost. In more expensive packages, the ground plane is exposed on the bottom and can be directly attached

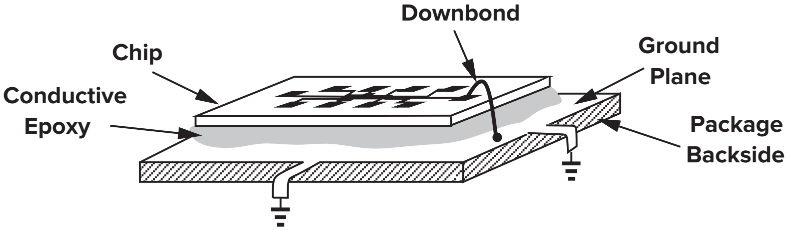


Figure 19.68 Package using a ground plane for substrate connection.

to the board ground, thus avoiding the inductance of the package pins. Also, the ground pads of the circuit can be "downbonded" to the underlying plane to minimize their inductance (while increasing the cost).

The effect of self-inductance must also be considered for input signals. The inductance, along with the pad capacitance and the circuit's input capacitance, forms a low-pass filter, attenuating high-frequency components and/or creating severe ringing in transient waveforms. For example, in the precision multiply-by-two circuit described in Sec. 13.3.3, when the two capacitors are switched to the input, the package inductance may limit the settling speed.

Some ICs require constant voltages that must be provided externally. Such voltages may serve as an accurate reference, e.g., in A/D or D/A converters, or to define some bias points on the chip. The package inductance degrades the settling behavior if the circuit injects significant switching noise into the reference.

▶ Example 19.8

Differential pairs are often used as "current switches." As shown in Fig. 19.69, the circuit routes its tail current to either of the outputs according to the large swings controlling the gates of M_1 and M_2. Explain what happens at node X during switching. If the tail currents of a large number of differential pairs feed from node X, should this voltage be provided externally?

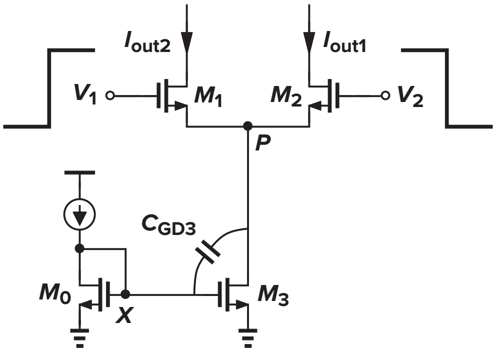


Figure 19.69 Differential pair operating as a current switch.

Solution

Recall from Chapter 4 that for the differential pair to experience complete switching, the differential swing $|V_2 - V_1|$ must exceed $\sqrt{2}(V_{GS} - V_{TH})_{eq}$, where $(V_{GS} - V_{TH})_{eq}$ is the overdrive of M_1 and M_2 in equilibrium, i.e., if $I_{D1} = I_{D2}$. We denote the voltage at node P when the pair is completely switched by V_{P1}, and in equilibrium by V_{P2}. Thus,

$$V_{P1} = V_2 - \sqrt{2}(V_{GS} - V_{TH})_{eq} \tag{19.25}$$

In equilibrium,

$$V_{P2} = \frac{V_1 + V_2}{2} - (V_{GS} - V_{TH})_{eq} \tag{19.26}$$

Assuming that $V_2 - V_1 = \sqrt{2}(V_{GS} - V_{TH})_{eq}$, and hence $V_1 = V_2 - \sqrt{2}(V_{GS} - V_{TH})_{eq}$, we have

$$V_{P2} = V_2 - \left(1 + \frac{\sqrt{2}}{2}\right)(V_{GS} - V_{TH})_{eq} \tag{19.27}$$

Thus, V_{P2} is *lower* than V_{P1} by $(1 - \sqrt{2}/2)(V_{GS} - V_{TH})_{eq}$, indicating that during switching, V_P drops by this amount. This voltage change is coupled to node X through the gate-drain overlap capacitance of M_3, disturbing I_{D3} and hence I_{out1} or I_{out2}.

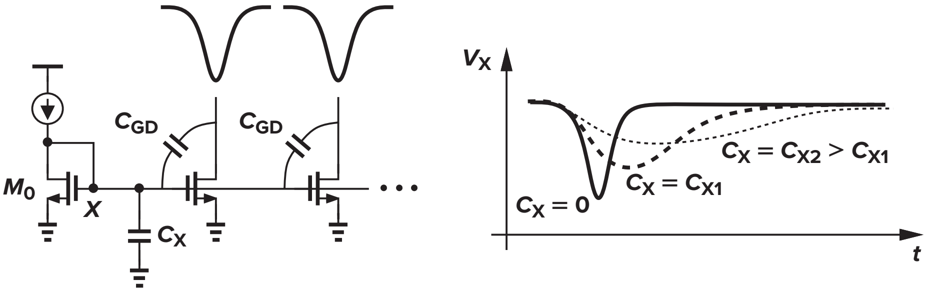


Figure 19.70 Addition of on-chip bypass capacitor to suppress noise at node X.

With a large number of current switches connected to node X, the disturbance may be quite significant, demanding that a decoupling capacitor be connected from node X to ground (Fig. 19.70). However, such a capacitor along with the small-signal resistance of M_0 introduces a long settling time at node X, possibly degrading the overall speed. To avoid this effect, C_X may need to be 100 to 1,000 times the *total* gate-drain overlap capacitance that injects noise into X. If such a large capacitor is placed off-chip, it actually appears in series with the package inductance (Fig. 19.71). In general, careful simulations are necessary to determine the preferable choice here. In many cases, leaving node X *agile* yields the fastest settling.

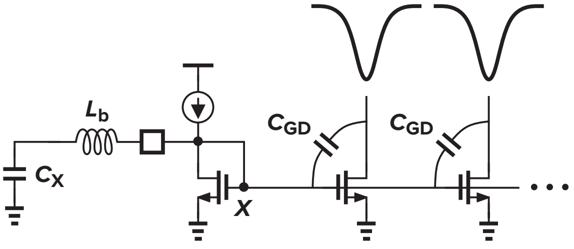


Figure 19.71 Addition of bypass capacitor externally.

The self-inductance of package connections also affects the performance of digital output buffers. In high-speed systems, these drivers must deliver tens of milliamps of current to the load with fast transitions. With many such buffers operating in a mixed-signal circuit, the resulting voltage drops on the power lines may become very large, increasing the rise time and fall time of the digital outputs and corrupting their timing.

Mutual Inductance While dedicating separate power lines to analog and digital sections reduces the noise on the analog supply, some noise may still couple to sensitive signals through the mutual inductance of bond wires and package traces. As illustrated in Fig. 19.72, both analog supplies and analog inputs are susceptible to noise or transitions on digital supplies, clock lines, or output buffers. With an arbitrary pad configuration, even differential signaling cannot eliminate this effect because the noisy lines may not surround the sensitive lines symmetrically. Thus, the design of the pad frame and the position of the pads play a critical role in the performance that can be achieved.

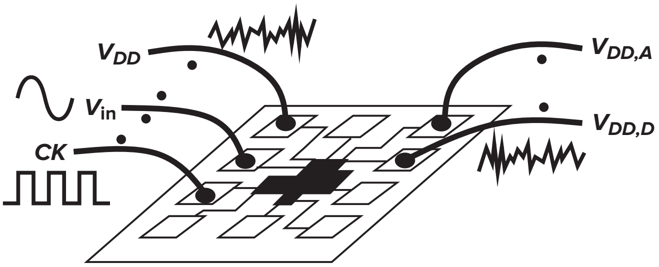


Figure 19.72 Coupling due to mutual inductance between wires.

Mutual inductance also manifests itself in parallel bond wires used to lower the overall self-inductance of a connection (Fig. 19.73). For two such wires, the equivalent inductance is equal to $(L_S + M)/2$, where M denotes the mutual inductance, rather than $L_S/2$.

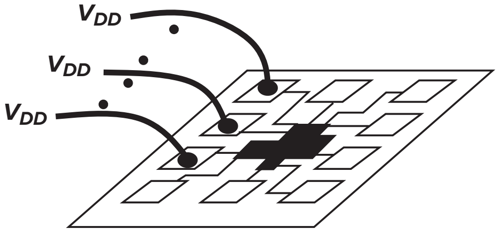


Figure 19.73 Multiple supply bond wires with mutual coupling.

Two methods can reduce the mutual coupling between inductors. First, the wires can be connected such that they are perpendicular to each other, i.e., they terminate on perpendicular sides of the chip [Fig. 19.74(a)]. Second, (quiet) ground or supply lines can be interposed between critical bond wires [Fig. 19.74(b)]. As shown in Fig. 19.74(c), even if several parallel lines are surrounded by ground wires, the effect of mutual inductance drops to negligible values.

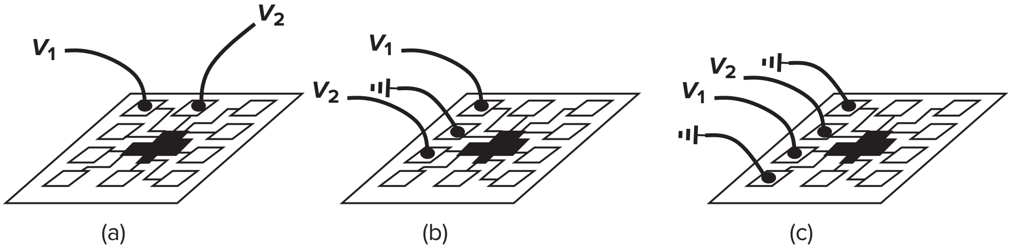


Figure 19.74 Reduction of mutual coupling by (a) perpendicular lines, (b) additional ground lines, and (c) occasional ground lines.

It is also interesting to note that mutual inductance *reduces* the self-inductance of two wires if they carry currents in opposite directions. If, as shown in Fig. 19.75, the supply and ground lines of a circuit are in parallel, then the total inductance equals $2L_S - M$ rather than $2L_S$. This observation proves useful in designing the pad frame and determining the package connections.

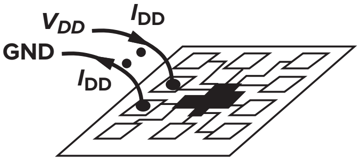


Figure 19.75 Reduction of mutual inductance between two wires carrying equal and opposite currents.

Self- and Mutual Capacitance The capacitance seen from each trace of the package to ground may limit the input bandwidth of the circuit or load the preceding stage. More important, this capacitance and the total inductance of the bond wire and the package trace yield a finite resonance frequency that may be stimulated by various transient currents drawn by the circuit. Since the wires and traces exhibit a small series resistance, a high quality factor (Q) results, giving rise to a sharp resonance and amplifying the noise considerably. The capacitance between the traces leads to additional coupling between lines and must be included in simulations.

References

[1] N. C. C. Lu et al., "Modeling and Optimization of Monolithic Polycrystalline Silicon Resistors," *IEEE Trans. Electron Devices*, vol. ED-28, pp. 818–830, July 1981.

[2] D. Su et al., "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Cicuits," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 420–430, April 1993.

[3] T. Blalack and B. A. Wooley, "The Effects of Switching Noise on an Oversampling A/D Converter," *ISSCC Dig. of Tech. Papers*, pp. 200–201, February 1995.

[4] B. Razavi, *Principles of Data Conversion System Design* (New York: IEEE Press, 1995).

[5] D. W. Dobberpuhl, "Circuits and Technology for Digital's StrongARM and ALPHA Microprocessors," *Proc. of 17th Conference on Advanced Research in VLSI*, pp. 2–11, September 1997.

[6] N. K. Verghese, T. J. Schmerbeck, and D. J. Allstot, *Simulation Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits* (Boston: Kluwer Academic Publishers, 1995).

Problems

Unless otherwise stated, in the following problems, use the device data shown in Table 2.1 and assume that $V_{DD} = 3$ V where necessary. Also, assume that all transistors are in saturation.

19.1. In Fig. 19.3, polysilicon has a sheet resistance of 30 $\Omega/\square$ (before silicidation) and metal 1 a sheet resistance of 80 m$\Omega/\square$. What is the ratio of the resistivities of the two materials?

19.2. A MOSFET with $W/L = 100\ \mu\text{m}/0.5\ \mu\text{m}$ undergoes ideal scaling by a factor of two. What happens to the sheet resistivity and the total resistance of the gate?

19.3. A cascode structure uses $W/L = 100\ \mu\text{m}/0.5\ \mu\text{m}$ for both the input device and the cascode device. If the sheet resistance of polysilicon is 5 $\Omega/\square$ and the maximum tolerable gate resistance 10 Ω, draw the layout of the structure while minimizing the drain junction capacitances.

19.4. In Fig. 19.7, explain what happens to the differential amplifier if each of the design rules A_1–A_8 is violated.

19.5. The input differential pair of an amplifier is to be laid out as in Fig. 19.19, but with each half device (e.g., $1/2M_1$) using four gate fingers. What is the minimum number of interconnect layers required here?

19.6. Large integrated circuits may suffer from significant temperature gradients. Compare the performance of the circuits shown in Figs. 19.23 and 19.24 in such an environment.

19.7. Suppose polysilicon with silicide block has a sheet resistance of 60 $\Omega/\square$ and a parallel-plate capacitance of 100 aF/μm^2 to the substrate. Also, assume that these parameters are respectively equal to 2 k$\Omega/\square$ and 1,000 aF/μm^2 for the n-well. Determine which material should be used to construct a 500-Ω resistor if matching considerations require a minimum poly width of 3 μm and a minimum n-well length of 6 μm. Neglect fringe capacitances.

19.8. Using the data in Table 18.1, calculate C and C_P for each structure in Fig. 19.35 and identify the one with minimum C_P/C. Neglect fringe capacitances.

19.9. A metal 4 wire with a length of 1,000 μm and width of 1 μm is driven by a source impedance of 500 Ω. Using the data in Table 18.1 and assuming a sheet resistance of 40 m$\Omega/\square$, calculate the delay through the wire and compare the result with the lumped time constant obtained by multiplying the source impedance by the total wire capacitance.

19.10. Repeat Problem 19.9 if the width of the wire is increased to 2 μm.

19.11. An interconnect having a length of 1,000 μm is required in a circuit. Using the data in Table 15.1 and assuming that the sheet resistance of metals 1–3 is 80 m$\Omega/\square$ and that of metal 4 is 40 m$\Omega/\square$, determine which metal layer must be used to obtain the minimum delay.

19.12. Some new technologies use copper for interconnects because its resistivity is about half that of aluminum. Repeat Problem 19.11 with copper interconnects.

19.13. In the circuit of Fig. 19.53(a), $(W/L)_1 = 100/0.5$ and $I_{D1} = 1$ mA. If the substrate noise, V_{sub}, has a peak-to-peak amplitude of 50 mV, what is the effect referred to the gate of M_1?

19.14. Suppose two bond wires are placed 5 mm above ground with a center-to-center spacing of 1 mm.

(a) What is the total mutual inductance if each wire is 4 mm long?

(b) If one wire carries a 100-MHz sinusoidal current with a peak amplitude of 1 mA, what is the voltage induced across the other wire?

19.15. In Problem 14, what center-to-center spacing is required to decrease the induced voltage by a factor of four?

19.16. In order to reduce the total bond wire inductance, a package uses 4 supply pads and 4 ground pads. Suppose the self-inductance of each wire is 4 nH and the mutual inductance between adjacent lines is 2 nH. Neglecting mutual inductance between nonadjacent lines, calculate the equivalent inductance of the supply and ground connections if (a) all of the supply wires are placed next to each other and so are the ground wires, and (b) every supply wire is placed next to a ground wire.

19.17. The input bandwidth of high-speed circuits may be limited by the bond wire inductance and the pad capacitance. Consider two cases: (a) the bond wire diameter is 50 μm and the pad size 100 μm × 100 μm; (b) the bond wire diameter is 25 μm and the pad size 50 μm × 50 μm. If all other dimensions are constant, which case is preferable?

Index