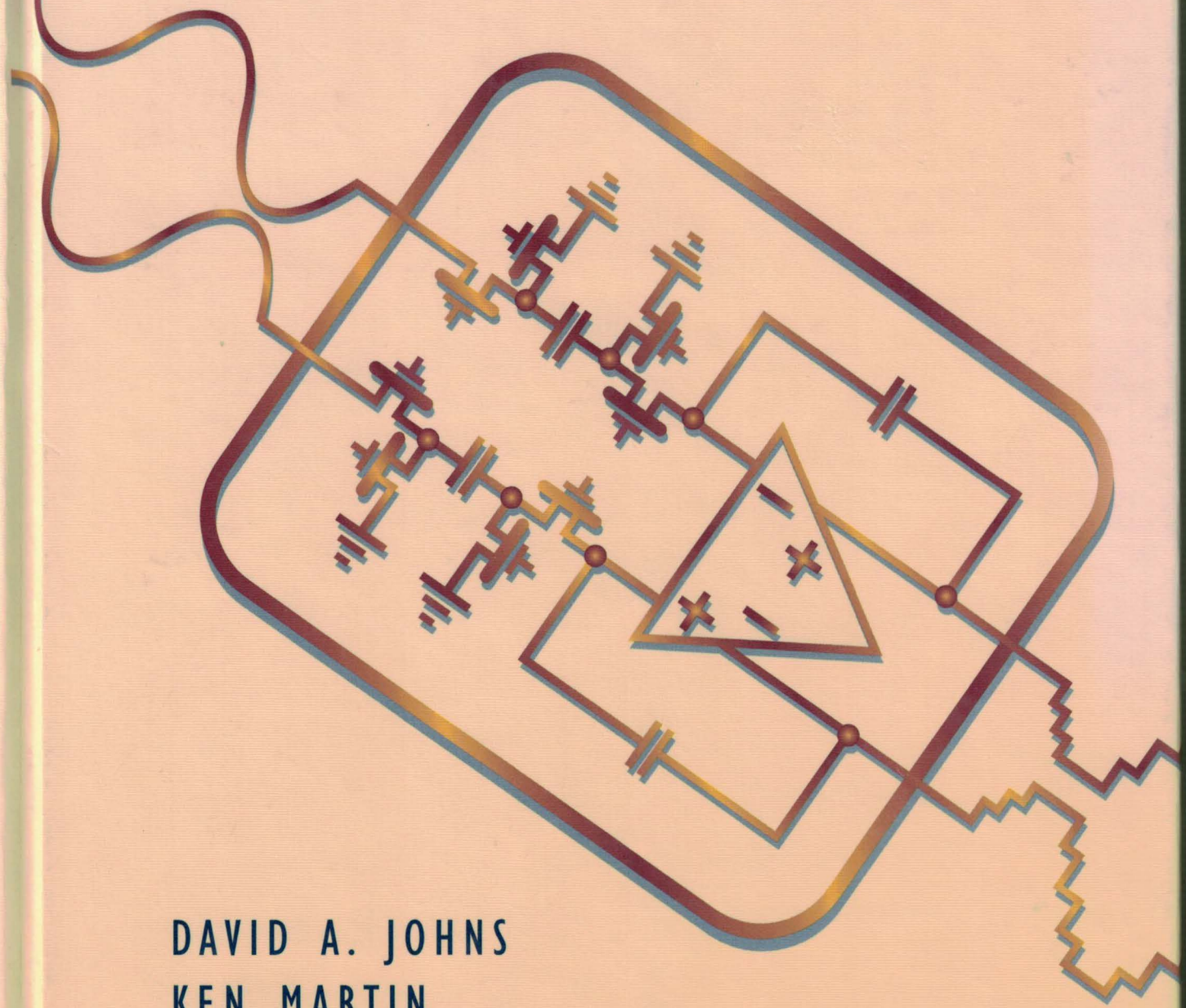
ANALOG INTEGRATED CIRCUIT DESIGN



DAVID A. JOHNS
KEN MARTIN

IPR2023-00697
Petitioners Lenovo (United States) Inc.
and Motorola Mobility LLC - Ex.1016

ANALOG INTEGRATED CIRCUIT DESIGN

David Johns
Ken Martin
University of Toronto

John Wiley & Sons, Inc.
New York • Chichester • Brisbane
Toronto • Singapore • Weinheim

Acquisitions Editor	Charity Robey
Marketing Manager	Jay Kirsch
Production Manager	Lucille Buonocore
Senior Production Editor	Tracey Kuehn
Designer	Kevin Murphy
Manufacturing Manager	Mark Cirillo
Illustration Editor	Sigmund Malinowski

This book was set in Times Roman by Publication Services and printed and bound by R.R. Donnelley/Crawfordsville. The cover was printed by Lehigh Press.

Recognizing the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the number of trees cut each year does not exceed the amount of new growth.

Copyright © 1997, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 and 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging-in-Publication Data:

Johns, David, 1958–
Analog integrated circuit design / David Johns, Ken Martin.
p. cm.
Includes bibliographical references.
ISBN 0-471-14448-7 (cloth : alk. paper)
1. Linear integrated circuits—Design and construction.
I. Martin, Kenneth W. (Kenneth W.) 1952– . II. Title.
TK7874.J65 1996
621.3815—dc20 96-34365
CIP

ISBN 0-471-14448-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

For the past twenty years, numerous people have predicted there soon would be little need for analog circuitry because the world would rely on digital circuits. However, although many applications have indeed replaced much analog circuitry with their digital counterparts (such as digital audio), the need for good analog circuit design remains strong. For example, when digitizing physical signals, analog-to-digital and digital-to-analog converters are always needed, together with their associated anti-aliasing and reconstruction filters. In addition, new applications continue to appear in which speed and power-consumption requirements often demand the use of high-speed analog front ends, such as digital communications over copper wires or wireless communication channels. Also as integrated circuits become larger due to system integration, it is much more likely that at least some portion of a modern integrated circuit will include analog circuitry required to interface to the real world. Often this analog circuitry, although it constitutes only a small portion of the total chip area, may be the limiting factor on overall system performance and the most difficult part of the IC to design. As a result, a strong industrial need for analog circuit designers continues. The purpose of this book is to help develop excellent analog circuit designers by presenting a concise treatment of the wide array of knowledge required by an integrated circuit designer.

Many refer to the designing and testing of high-performance analog circuits as a "mystical art". In other words, whereas digital design is relatively systematic, analog design appears to be much more confusing and based on gut feelings. In addition, analog testing may sometimes seem to depend more on the time of day and phase of the moon rather than on concrete electrical properties. These thoughts about analog circuits usually occur when one is not familiar with the many fundamentals required to create high-performance analog circuits. A major goal of this book is to help take the mystery out of analog integrated circuit design. The authors believe that most experienced electrical engineers are capable of good design if they are familiar with the most important design principles. We have attempted to highlight these principles throughout this text. Although many circuits and techniques are described, we have emphasized the most important and fundamental principles involved in realizing state-of-the-art analog circuits. Throughout this book, we give physical and intuitive explanations, and, although mathematical quantitative analyses of many circuits have necessarily been presented, we have attempted not to miss seeing the forest because of the trees. In other words, this book attempts to present the critical underlying concepts without becoming entangled in tedious and over-complicated circuit analyses.

INTENDED AUDIENCE

This book is primarily intended for use as a graduate-level textbook and as a reference for practicing engineers, although portions of this text are also useful for senior-level undergraduate courses. To appreciate the material in this book, it is expected that the reader has had at least one basic introductory course in analog circuits. Specifically, the reader should be familiar with the concept of small-signal analysis and have been exposed to basic transistor amplifier circuits. In addition, the reader should be comfortable working in the frequency domain (i.e., should be familiar with the Laplace transform) with possibly some knowledge of discrete-time signals.

In a graduate course, this book might be used in many ways. The chapters of this book have intentionally been made mostly independent so that some chapters can be covered while others are skipped. Also, it has been found to be very easy to change the order of presentation. For example, if readers have a good modelling background, they might skip Chapter 1, and if their discrete-time knowledge is good, Chapter 9 might be assigned only as review. We believe that such flexibility is essential in presenting textbooks for the later years of study.

Here at the University of Toronto, we present the material in this book in a few courses. A senior-level undergraduate course in analog circuits assigns Chapters 1, 3, 4, and 5, with the fundamentals and selected topics briefly covered from Chapters 6 and 7, as well. Another senior-level undergraduate course on analog filters uses Chapter 9 and the filter portion of Chapter 10 after presenting filter basics. In a first-level graduate course, we presently assign Chapters 1–3 and 5–7, while a second-level graduate course (often taken concurrently with the first course) assigns Chapters 4, 8–10, and 16. Chapters 9, 10, and 15 are presently covered in a separate graduate course with other material on advanced analog filters, whereas Chapters 11–14 (together with journal publications) are assigned in a data converters course. Furthermore, we have often modified this selection and order from year-to-year with little difficulty. Note that some chapters are used in both undergraduate and graduate courses since many of our graduate students do not come from our undergraduate program.

A secondary audience for this book includes recently graduated electrical engineers who wish to rapidly increase their knowledge of modern analog circuit-design techniques. In fact, much of the material covered in this text was originally taught and refined, over many years, in popular short courses offered to working engineers who realized the importance to their careers of upgrading their knowledge in analog circuit design. For this audience, we have put much effort into highlighting the most important considerations when designing the various circuits, and we have also tried to include modern, well-designed examples and references to sources for further study.

TEXT OUTLINE

In Chapter 1, the basic physical behavior and modelling for three integrated devices are presented—diodes, MOSFET transistors, and bipolar transistors. Here, many of

the modelling equations are derived to give the reader some appreciation of model parameters and how they are affected as processes evolve. Models for the three devices are summarized in a table format for quick reference.

In Chapter 2, issues associated with the manufacturing of a microcircuit are discussed. Emphasis is placed on CMOS fabrication, where details are given for a generic process. Layout and design rules are presented with respect to analog circuits where matching and noise considerations are important issues. This chapter concludes with a description of a destructive phenomenon known as latch-up.

Fundamental building blocks of analog integrated circuits are discussed in Chapter 3. Specifically, MOS current mirrors are presented as well as single-stage amplifiers. A point to note here is that only active-load amplifiers are considered since these types of amplifiers occur more often than other types in integrated circuits. The last section in Chapter 3 presents SPICE simulation results for selected examples within the chapter.

Noise analysis and modelling are discussed in Chapter 4. Here, we assume the reader has not previously been exposed to random-signal analysis, and thus basic concepts in analyzing random signals are first presented. Noise models are then presented for basic circuit elements. Finally, a variety of circuits are analyzed from a noise perspective to give the reader some experience in noise analysis.

In Chapter 5, the fundamental principles of basic opamp design are presented. To illustrate many of these principles, the design of a classic two-stage CMOS opamp is thoroughly discussed. A systematic compensation procedure, which also applies to many other opamp designs, is described. Also discussed here are methods to make the compensation and transconductance values quite stable. Finally, the chapter concludes with SPICE simulation results for selected examples.

Chapter 6 begins with a look at two advanced current-mirror approaches useful in low-power supply operation and in short-channel processes. Next, two modern high-speed opamps are discussed followed by a presentation of fully differential design techniques and common-mode feedback circuits. These fully differential opamps are used in many modern industrial applications where high speed and noise are important considerations. Next, current-feedback opamps are discussed. Their main advantage is that they maintain relatively constant bandwidth while the closed-loop gain is changed. Once again, SPICE simulation examples are found at the end of this chapter.

In Chapter 7, comparator design is discussed. Comparators are perhaps the second most common analog building block after opamps. Here, the practical limitations of comparators are described as well as circuit techniques to improve performance. In addition, examples of modern high-speed comparators are presented.

In Chapter 8, some additional analog building blocks are covered. Specifically, sample-and-hold circuits are first covered, followed by bandgap voltage references, and the chapter concludes with bipolar translinear-gain and multiplier circuits. At the end of this chapter, all the main analog building blocks have been covered and the remaining material in the text deals with more system-level analog blocks.

The basics of discrete-time signals and filters are presented in Chapter 9. This material is essential for understanding the operation of many modern analog circuits

such as switched-capacitor filters and oversampling converters. The approach taken here is to show the close relationship between the Z-transform and the Laplace transform, thereby building on the reader's experience in the continuous-time domain.

In Chapter 10, the basics of switched-capacitor circuits are described. Switched-capacitor techniques are the most common approach for realizing integrated filters due to their high degree of accuracy and linearity. The chapter concludes with a description of other switched-capacitor circuits, such as gain stages, modulators, and voltage-controlled oscillators.

In Chapter 11, the fundamentals of data converters are presented. Ideal converters are discussed first, and it is shown that quantization noise does not occur in D/A converters but is a fundamental limit in A/D converters. Signed codes are then presented, and the chapter concludes with a discussion of performance limitations.

Popular Nyquist-rate D/A architectures are discussed in Chapter 12, with various approaches for realizing Nyquist-rate A/D converters described in Chapter 13. The importance of data conversion cannot be overemphasized in today's largely digital world, and these two chapters discuss the main advantages and design issues of many modern approaches.

Oversampling converters are presented separately in Chapter 14 due to the large amount of signal-processing concepts needed to properly describe these converters. Here, digital issues (such as decimation filters) are also presented since good overall system knowledge is needed to properly design these types of converters. In addition, practical issues and recent new approaches (such as the use of multibit conversion) are also discussed. This chapter concludes with a third-order A/D converter example.

Continuous-time filters are presently being applied to many new products, particularly in high-speed data-communication applications, and they are the focus of Chapter 15. Bipolar, CMOS, and BiCMOS approaches are described, followed by some tuning approaches. Finally, dynamic range measurement aspects of continuous-time filters are discussed, such as THD, spurious-free dynamic range, and the third-order intercept point.

Finally, the text concludes with phase-locked loops in Chapter 16. Phase-locked loops are essential components for clock recovery in many digital and data communication circuits. Here, basic concepts and a simulation technique are presented for such systems.

ACKNOWLEDGMENTS

The authors would like to acknowledge many of their colleagues who participated in short courses in which much of the material for this text was originally taught and refined. In particular, Gabor C. Temes is acknowledged as well as instructors Jim McCreary and Bill Black. In addition, the authors acknowledge that much of the material and concepts originated with many practicing engineers they have worked with over the years as well as in the publications cited in the references section at the

end of each chapter. As much as possible, appropriate references for original concepts are cited in the text, but the authors have been working in the area of analog circuits for so many years that often the original sources of popular and important concepts have been forgotten. For any reference omissions, they sincerely apologize.

Contents

CHAPTER 1 INTEGRATED-CIRCUIT DEVICES AND MODELLING 1

1.1 Semiconductors and pn Junctions 1
1.2 MOS Transistors 16
1.3 Advanced MOS Modelling 39
1.4 Bipolar-Junction Transistors 42
1.5 Device Model Summary 56
1.6 SPICE-Modelling Parameters 61
1.7 Appendix 65
1.8 References 78
1.9 Problems 78

CHAPTER 2 PROCESSING AND LAYOUT 82

2.1 CMOS Processing 82
2.2 Bipolar Processing 95
2.3 CMOS Layout and Design Rules 96
2.4 Analog Layout Considerations 105
2.5 Latch-Up 118
2.6 References 121
2.7 Problems 121

CHAPTER 3 BASIC CURRENT MIRRORS AND SINGLE-STAGE AMPLIFIERS 125

3.1 Simple CMOS Current Mirror 125
3.2 Common-Source Amplifier 128
3.3 Source-Follower or Common-Drain Amplifier 129
3.4 Common-Gate Amplifier 132
3.5 Source-Degenerated Current Mirrors 135
3.6 High-Output-Impedance Current Mirrors 137
3.7 Cascode Gain Stage 140
3.8 MOS Differential Pair and Gain Stage 142
3.9 Bipolar Current Mirrors 146
3.10 Bipolar Gain Stages 149

3.11 Frequency Response 154
3.12 SPICE Simulation Examples 169
3.13 References 176
3.14 Problems 176

CHAPTER 4 NOISE ANALYSIS AND MODELLING 181

4.1 Time-Domain Analysis 181
4.2 Frequency-Domain Analysis 186
4.3 Noise Models for Circuit Elements 196
4.4 Noise Analysis Examples 204
4.5 References 216
4.6 Problems 217

CHAPTER 5 BASIC OPAMP DESIGN AND COMPENSATION 221

5.1 Two-Stage CMOS Opamp 221
5.2 Feedback and Opamp Compensation 232
5.3 SPICE Simulation Examples 251
5.4 References 252
5.5 Problems 253

CHAPTER 6 ADVANCED CURRENT MIRRORS AND OPAMPS 256

6.1 Advanced Current Mirrors 256
6.2 Folded-Cascode Opamp 266
6.3 Current-Mirror Opamp 273
6.4 Linear Settling Time Revisited 278
6.5 Fully Differential Opamps 280
6.6 Common-Mode Feedback Circuits 287
6.7 Current-Feedback Opamps 291
6.8 SPICE Simulation Examples 295
6.9 References 299
6.10 Problems 300

CHAPTER 7 COMPARATORS 304

7.1 Using an Opamp for a Comparator 304
7.2 Charge-Injection Errors 308
7.3 Latched Comparators 317
7.4 Examples of CMOS and BiCMOS Comparators 321
7.5 Examples of Bipolar Comparators 328
7.6 References 330
7.7 Problems 331

CHAPTER 8 SAMPLE AND HOLDS, VOLTAGE REFERENCES, AND TRANSLINEAR CIRCUITS 334

8.1 Performance of Sample-and-Hold Circuits 334
8.2 MOS Sample-and-Hold Basics 336
8.3 Examples of CMOS S/H Circuits 343
8.4 Bipolar and BiCMOS Sample and Holds 349
8.5 Bandgap Voltage Reference Basics 353
8.6 Circuits for Bandgap References 357
8.7 Translinear Gain Cell 364
8.8 Translinear Multiplier 366
8.9 References 368
8.10 Problems 370

CHAPTER 9 DISCRETE-TIME SIGNALS 373

9.1 Overview of Some Signal Spectra 373
9.2 Laplace Transforms of Discrete-Time Signals 374
9.3 z-Transform 377
9.4 Downsampling and Upsampling 379
9.5 Discrete-Time Filters 382
9.6 Sample-and-Hold Response 389
9.7 References 391
9.8 Problems 391

CHAPTER 10 SWITCHED-CAPACITOR CIRCUITS 394

10.1 Basic Building Blocks 394
10.2 Basic Operation and Analysis 398
10.3 First-Order Filters 409
10.4 Biquad Filters 415
10.5 Charge Injection 423
10.6 Switched-Capacitor Gain Circuits 427
10.7 Correlated Double-Sampling Techniques 433
10.8 Other Switched-Capacitor Circuits 434
10.9 References 441
10.10 Problems 443

CHAPTER 11 DATA CONVERTER FUNDAMENTALS 445

11.1 Ideal D/A Converter 445
11.2 Ideal A/D Converter 447
11.3 Quantization Noise 448
11.4 Signed Codes 452

11.5 Performance Limitations 454
11.6 References 461
11.7 Problems 461

CHAPTER 12 NYQUIST-RATE D/A CONVERTERS 463

12.1 Decoder-Based Converters 463
12.2 Binary-Scaled Converters 469
12.3 Thermometer-Code Converters 475
12.4 Hybrid Converters 481
12.5 References 484
12.6 Problems 484

CHAPTER 13 NYQUIST-RATE A/D CONVERTERS 487

13.1 Integrating Converters 487
13.2 Successive-Approximation Converters 492
13.3 Algorithmic (or Cyclic) A/D Converter 504
13.4 Flash (or Parallel) Converters 507
13.5 Two-Step A/D Converters 513
13.6 Interpolating A/D Converters 516
13.7 Folding A/D Converters 519
13.8 Pipelined A/D Converters 523
13.9 Time-Interleaved A/D Converters 526
13.10 References 527
13.11 Problems 528

CHAPTER 14 OVERSAMPLING CONVERTERS 531

14.1 Oversampling without Noise Shaping 531
14.2 Oversampling with Noise Shaping 538
14.3 System Architectures 547
14.4 Digital Decimation Filters 551
14.5 Higher-Order Modulators 555
14.6 Bandpass Oversampling Converters 557
14.7 Practical Considerations 559
14.8 Multi-Bit Oversampling Converters 565
14.9 Third-Order A/D Design Example 568
14.10 References 571
14.11 Problems 572

CHAPTER 15 CONTINUOUS-TIME FILTERS 574

15.1 Introduction to G_m-C Filters 575
15.2 Bipolar Transconductors 584

15.3 CMOS Transconductors Using Triode Transistors 597
15.4 CMOS Transconductors Using Active Transistors 607
15.5 BiCMOS Transconductors 616
15.6 MOSFET-C Filters 620
15.7 Tuning Circuitry 626
15.8 Dynamic Range Performance 635
15.9 References 643
15.10 Problems 645

CHAPTER 16 PHASE-LOCKED LOOPS **648**

16.1 Basic Loop Architecture 648
16.2 PLLs with Charge-Pump Phase Comparators 663
16.3 Voltage-Controlled Oscillators 670
16.4 Computer Simulation of PLLs 680
16.5 Appendix 689
16.6 References 692
16.7 Problems 693

INDEX **696**

CHAPTER

1 Integrated-Circuit Devices and Modelling

In this chapter, the operation and modelling of semiconductor devices are described. Although it is possible to do simple integrated-circuit design with a basic knowledge of semiconductor device modelling, for high-speed state-of-the-art design, an in-depth understanding of the second-order effects of device operation and their modelling is considered critical.

It is assumed that most readers have been introduced to transistors and their basic modelling in a previous course. Thus, fundamental semiconductor concepts are only briefly reviewed. Section 1.1 describes pn junctions (or diodes). This section is important in understanding the parasitic capacitances in many device models, such as junction capacitances. Section 1.2 covers MOS transistors and modelling. It should be noted that this section relies to some degree on the material previously presented in Section 1.1, in which depletion capacitance is covered. Section 1.4 covers bipolar-junction transistors and modelling. A summary of device models and important equations is presented in Section 1.5. This summary is particularly useful for a reader who already has a good background in transistor modelling, in which case the summary can be used to follow the notation used throughout the remainder of this book. In addition, a brief description is given of the most important process-related parameters used in SPICE modelling. Finally, this chapter concludes with an Appendix containing derivations of the more physically based device equations.

1.1 SEMICONDUCTORS AND pn JUNCTIONS

A semiconductor is a crystal lattice structure that can have free electrons (which are negative carriers) and/or free holes (which are an absence of electrons and are equivalent to positive carriers). The type of semiconductor typically used is silicon (commonly called *sand*). This material has a valence of four, implying that each atom has four free electrons to share with neighboring atoms when forming the covalent bonds of the crystal lattice. *Intrinsic* silicon (i.e., undoped silicon) is a very pure crystal structure having equal numbers of free electrons and holes. These free carriers are those electrons or holes that have gained enough energy due to thermal agitation to escape their bonds. At room temperature, there are approximately 1.5×10^{10} carriers of each type per cm^3, or equivalently 1.5×10^{16} carriers/m^3. The number of carriers approximately doubles for every 11 °C increase in temperature.

If one dopes silicon with a pentavalent impurity (i.e., atoms of an element having a valence of five, or equivalently five electrons in the outer shell, available when bonding with neighboring atoms), there will be almost one extra free electron for every impurity atom.[1] These free electrons can be used to conduct current. A pentavalent impurity is said to *donate* free electrons to the silicon crystal, and thus the impurity is known as a *donor*. Examples of donor elements are phosphorus, P, and arsenic, As. These impurities are also called n-type dopants since the free carriers resulting from their use have negative charge. When an n-type impurity is used, the total number of negative carriers or electrons is almost the same as the doping concentration, and is much greater than the number of free electrons in intrinsic silicon. In other words,

$$n_n = N_D \tag{1.1}$$

where n_n denotes the free-electron concentration in n-type material and N_D is the doping concentration (with the subscript D denoting donor). On the other hand, the number of free holes in n-doped material will be much less than the number of holes in intrinsic silicon and can be shown [Sze, 1981] to be given by

$$p_n = \frac{n_i^2}{N_D} \tag{1.2}$$

Here, n_i is the carrier concentration in intrinsic silicon.

Similarly, if one dopes silicon with atoms having a valence of three, for example, boron (B), the concentration of positive carriers or holes will be approximately equal to the *acceptor* concentration, N_A,

$$p_p = N_A \tag{1.3}$$

and the number of negative carriers in the p-type silicon, n_p, is given by

$$n_p = \frac{n_i^2}{N_A} \tag{1.4}$$

EXAMPLE 1.1

Intrinsic silicon is doped with boron at a concentration of 10^{26} atoms/m^3. At room temperature, what are the concentrations of holes and electrons in the resulting doped silicon? Assume that $n_i = 1.5 \times 10^{16}$ carriers/m^3.

Solution

The hole concentration, p_p, will approximately equal the doping concentration ($p_p = N_A = 10^{26}$ holes/m^3). The electron concentration is found from (1.4) to be

1. In fact, there will be slightly fewer mobile carriers than the number of impurity atoms since some of the free electrons from the dopants have recombined with holes. However, since the number of holes of intrinsic silicon is much less than typical doping concentrations, this inaccuracy is small.

$$n_p = \frac{(1.5 \times 10^{16})^2}{10^{26}} = 2.3 \times 10^6 \text{ electrons/m}^3 \tag{1.5}$$

Such doped silicon is referred to as p type since it has many more free holes than free electrons.

Diodes

To realize a diode, or, equivalently, a pn junction, one part of a semiconductor is doped n type, and a closely adjacent part is doped p type, as shown in Fig. 1.1. Here the diode, or junction, is formed between the p^+ region and the n region. It should be noted that the superscripts indicate the relative doping levels. For example, the p^- bulk region might have an impurity concentration of 5×10^{21} carriers/m^3, whereas the p^+ and n^+ regions would be doped more heavily to a value around 10^{25} to 10^{27} carriers/m^3. Also, note that the metal contacts to the diode (in this case, aluminum) are connected to a heavily doped region as opposed to a lightly doped region; otherwise a *Schottky diode* would occur. (Schottky diodes are discussed on page 15.) Thus, in order not to make a Schottky diode, the connection to the n region is actually made via the n^+ region.

In the p^+ side, a large number of free positive carriers are available, whereas in the n side, many free negative carriers are available. The holes in the p^+ side will tend to disperse or diffuse into the n side, whereas the free electrons in the n side will tend to diffuse to the p^+ side. This process is very similar to two gases randomly diffusing together. This diffusion lowers the concentration of free carriers in the region between the two sides. As the two types of carriers diffuse together, they recombine. Every electron that diffuses from the n side to the p side leaves behind a *bound* positive charge close to the transition region. Similarly, every hole that diffuses from the p side leaves behind a bound electron near the transition region. The end result is shown in Fig. 1.2. This diffusion of free carriers creates a *depletion region* at the junction of the two sides where no free carriers exist, and which has a net negative charge on the p^+ side and a net positive charge on the n side. The total amount of exposed or bound

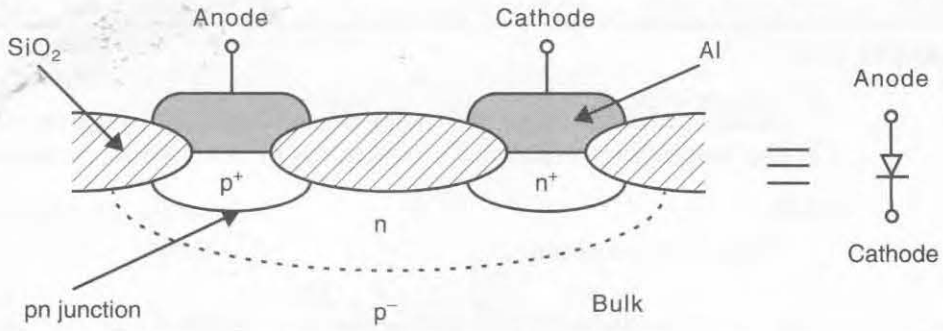


Fig. 1.1 A cross section of a pn diode.

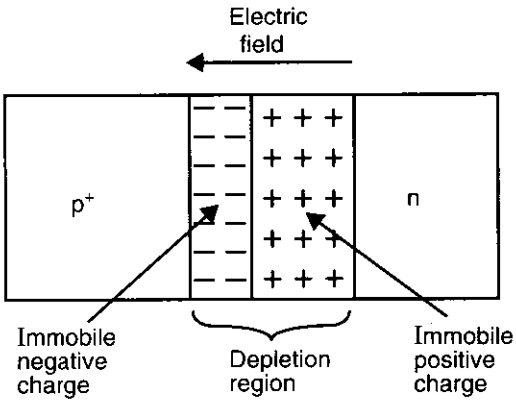


Fig. 1.2 A simplified model of a diode. Note that a depletion region exists at the junction due to diffusion and extends farther into the more lightly doped side.

charge on the two sides of the junction must be equal for charge neutrality. This requirement causes the depletion region to extend farther into the more lightly doped n side than into the p^+ side.

As these bound charges are exposed, an electric field develops going from the n side to the p side. This electric field is often called the built-in potential of the junction. It opposes the diffusion of free carriers until there is no net movement of charge under open-circuit and steady-state conditions. The built-in voltage of an open-circuit pn junction is given by Sze [1981] as

$$\Phi_0 = V_T \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{1.6}$$

where

$$V_T = \frac{kT}{q} \tag{1.7}$$

with T being the temperature in degrees Kelvin ($\cong$ 300 °K at room temperature), k being Boltzmann's constant (1.38×10^{-23} JK^{-1}), and q being the charge of an electron (1.602×10^{-19} C). At room temperature, V_T is approximately 26 mV.

EXAMPLE 1.2

A pn junction has $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3. What is the built-in junction potential? Assume that $n_i = 1.5 \times 10^{16}$ carriers/m^3.

Solution

Using (1.6), we obtain

$$\Phi_0 = 0.026 \times \ln\left(\frac{10^{25} \times 10^{22}}{(1.5 \times 10^{16})^2}\right) = 0.88 \text{ V} \tag{1.8}$$

This is a typical value for the built-in potential of a junction with one side heavily doped. As an approximation, we will normally use $\Phi_0 \cong 0.9$ V for the built-in potential of a junction having one side heavily doped.

Reverse-Biased Diodes

A silicon diode having an anode-to-cathode (i.e., p side to n side) voltage of 0.4 V or less will not be conducting appreciable current. In this case, it is said to be *reverse biased.* If a diode is reverse biased, current flow is primarily due to thermally generated carriers in the depletion region, and it is extremely small. Although this reverse-biased current is only weakly dependent on the applied voltage, *the reverse-biased current is directly proportional to the area of the diode junction.* However, an effect that should not be ignored, particularly at high frequencies, is the junction capacitance of a diode. In reverse-biased diodes, this junction capacitance is due to varying charge storage in the depletion regions and is modelled as a *depletion capacitance.*

To determine the depletion capacitance, we first state the relationship between the depletion widths and the applied reverse voltage, V_R [Sze, 1981].

$$x_n = \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{q}\frac{N_A}{N_D(N_A + N_D)}\right]^{1/2} \tag{1.9}$$

$$x_p = \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{q}\frac{N_D}{N_A(N_A + N_D)}\right]^{1/2} \tag{1.10}$$

Here, ε_0 is the permittivity of free space (equal to 8.854×10^{-12} F/m), V_R is the reverse-bias voltage of the diode, and K_s is the relative permittivity of silicon (equal to 11.8). It should be noted that these equations assume that the doping changes abruptly from the n to the p side.

From the above equations, we see that if one side of the junction is more heavily doped than the other, the depletion region will extend mostly on the lightly doped side. For example, if $N_A >> N_D$ (i.e., if the p region is more heavily doped), we can approximate (1.9) and (1.10) as

$$x_n \cong \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)}{qN_D}\right]^{1/2} \qquad x_p \cong \left[\frac{2K_s\varepsilon_0(\Phi_0 + V_R)N_D}{qN_A^2}\right]^{1/2} \tag{1.11}$$

Indeed, for this case

$$\frac{x_n}{x_p} \cong \frac{N_A}{N_D} \tag{1.12}$$

This special case is called a *single-sided diode.*

EXAMPLE 1.3

For a pn junction having $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3, what are the depletion-layer depths for a 5-V reverse-bias voltage?

Solution

Since $N_A >> N_D$ and we already have found in Example 1.2 that $\Phi_0 = 0.9$ V, we can use (1.11) to find

$$x_n = \left[\frac{2 \times 11.8 \times 8.854 \times 10^{-12} \times 5.9}{1.6 \times 10^{-19} \times 10^{22}}\right]^{1/2} = 0.88\ \mu\text{m} \tag{1.13}$$

$$x_p = \frac{x_n}{(N_A/N_D)} = 0.88\ \text{nm} \tag{1.14}$$

Note that the depletion width in the lightly doped n region is 1,000 times greater than that in the more heavily doped p region.

The charge stored in the depletion region, per unit cross-sectional area, is found by multiplying the depletion-region width by the concentration of the immobile charge (which is approximately equal to q times the impurity doping density). For example, on the n side, we find the charge in the depletion region to be given by multiplying (1.9) by qN_D, resulting in

$$Q^+ = \left[2qK_s\varepsilon_0(\Phi_0 + V_R)\frac{N_A N_D}{N_A + N_D}\right]^{1/2} \tag{1.15}$$

This amount of charge must also equal Q^- on the p side since there is charge equality. In the case of a single-sided diode when $N_A >> N_D$, we have

$$Q^- = Q^+ \cong [2qK_s\varepsilon_0(\Phi_0 + V_R)N_D]^{1/2} \tag{1.16}$$

Note that this result is independent of the impurity concentration on the heavily doped side. Thus, we see from the above relation that the charge stored in the depletion region is dependent on the applied reverse-bias voltage. *It is this charge-voltage relationship that is modelled by a nonlinear depletion capacitance.*

For small changes in the reverse-biased junction voltage, about a bias voltage, we can find an equivalent *small-signal* capacitance, C_j, by differentiating (1.15) with respect to V_R. Such a differentiation results in

$$C_j = \frac{dQ^+}{dV_R} = \left[\frac{qK_s\varepsilon_0}{2(\Phi_0 + V_R)}\frac{N_A N_D}{N_A + N_D}\right]^{1/2} = \frac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}} \tag{1.17}$$

where C_{j0} is the depletion capacitance per unit area at $V_R = 0$ and is given by

$$C_{j0} = \sqrt{\frac{qK_s\varepsilon_0}{2\Phi_0}\frac{N_A N_D}{N_A + N_D}} \tag{1.18}$$

In the case of a one-sided diode with $N_A >> N_D$, we have

$$C_j = \left[\frac{qK_s\varepsilon_0 N_D}{2(\Phi_0 + V_R)}\right]^{1/2} = \frac{C_{j0}}{\sqrt{1 + \frac{V_R}{\Phi_0}}} \tag{1.19}$$

where now

$$C_{j0} = \sqrt{\frac{qK_s\varepsilon_0 N_D}{2\Phi_0}} \tag{1.20}$$

It should be noted that many of the junctions encountered in integrated circuits are one-sided junctions with the lightly doped side being the substrate or sometimes what is called the *well*. The more heavily doped side is often used to form a contact to interconnecting metal. From (1.20), we see that, for these one-sided junctions, the depletion capacitance is approximately independent of the doping concentration on the heavily doped side, and is proportional to the square root of the doping concentration of the more lightly doped side. Thus, smaller depletion capacitances are obtained for more lightly doped substrates—a strong incentive to strive for lightly doped substrates.

Finally, note that by combining (1.15) and (1.18), we can express the equation for the immobile charge on either side of a reverse-biased junction as

$$Q = 2C_{j0}\Phi_0\sqrt{1 + \frac{V_R}{\Phi_0}} \tag{1.21}$$

As seen in Example 1.6, this equation is useful when one is approximating the large-signal charging (or discharging) time for a reverse-biased diode.

EXAMPLE 1.4

For a pn junction having $N_A = 10^{25}$ holes/m^3 and $N_D = 10^{22}$ electrons/m^3, what is the total zero-bias depletion capacitance for a diode of area 10 μm × 10 μm? What is its depletion capacitance for a 3-V reverse-bias voltage?

Solution

Making use of (1.20), we have

$$C_{j0} = \sqrt{\frac{1.6 \times 10^{-19} \times 11.8 \times 8.854 \times 10^{-12} \times 10^{22}}{2 \times 0.9}} = 304.7\ \mu\text{F/m}^2 \tag{1.22}$$

Since the diode area is 100×10^{-12} m^2, the total zero-bias depletion capacitance is

$$C_{T\text{-}j0} = 100 \times 10^{-12} \times 304.7 \times 10^{-6} = 30.5 \text{ fF} \tag{1.23}$$

At a 3-V reverse-bias voltage, we have from (1.19)

$$C_{T\text{-}j} = \frac{30.5 \text{ fF}}{\sqrt{1 + \left(\frac{3}{0.9}\right)}} = 14.7 \text{ fF} \tag{1.24}$$

As expected, we see a decrease in junction capacitance as the width of the depletion region is increased.

Graded Junctions

All of the above equations assumed an abrupt junction where the doping concentration changes quickly from p to n over a small distance. Although this is a good approximation for many integrated circuits, it is not always true. For example, the collector-to-base junction of a bipolar transistor is most commonly realized as a *graded* junction. In the case of graded junctions, the exponent 1/2 in Eq. (1.15) is inaccurate, and a better value to use is an exponent closer to unity, perhaps 0.6 to 0.7. Thus, for graded junctions, (1.15) is typically written as

$$Q = \left[2qK_s\varepsilon_0(\Phi_0 + V_R)\frac{N_A N_D}{N_A + N_D}\right]^{1-m} \tag{1.25}$$

where m is a constant typically around 1/3.

Differentiating (1.25) to find the depletion capacitance, we have

$$C_j = (1-m)\left[2qK_s\varepsilon_0\frac{N_A N_D}{N_A + N_D}\right]^{1-m}\frac{1}{(\Phi_0 + V_R)^m} \tag{1.26}$$

This depletion capacitance can also be written as

$$C_j = \frac{C_{j0}}{\left(1 + \frac{V_R}{\Phi_0}\right)^m} \tag{1.27}$$

where

$$C_{j0} = (1-m)\left[2qK_s\varepsilon_0\frac{N_A N_D}{N_A + N_D}\right]^{1-m}\frac{1}{\Phi_0^m} \tag{1.28}$$

From (1.27), we see that a graded junction results in a depletion capacitance that is less dependent on V_R than the equivalent capacitance in an abrupt junction. In other words, since m is less than 0.5, the depletion capacitance for a graded junction is

more linear than that for an abrupt junction. Correspondingly, increasing the reverse-bias voltage for a graded junction is not as effective in reducing the depletion capacitance as it is for an abrupt junction.

Finally, as in the case of an abrupt junction, the depletion charge on either side of the junction can also be written as

$$Q = \frac{C_{j0}}{1-m}\Phi_0\left(1+\frac{V_R}{\Phi_0}\right)^{1-m} \tag{1.29}$$

EXAMPLE 1.5

Repeat Example 1.4 for a graded junction with $m = 0.4$.

Solution

Noting once again that $N_A >> N_D$, we approximate (1.28) as

$$C_{j0} = (1-m)[2qK_s\varepsilon_0 N_D]^{1-m}\frac{1}{\Phi_0^m} \tag{1.30}$$

resulting in

$$C_{j0} = 81.5\ \mu F/m^2 \tag{1.31}$$

which, when multiplied by the diode's area of $10\ \mu m \times 10\ \mu m$, results in

$$C_{T\text{-}j0} = 8.1\ fF \tag{1.32}$$

For a 3-V reverse-bias voltage, we have

$$C_{T\text{-}j} = \frac{8.1 fF}{(1+3/0.9)^{0.4}} = 4.5\ fF \tag{1.33}$$

Large-Signal Junction Capacitance

The equations for the junction capacitance given above are only valid for small changes in the reverse-bias voltage. This limitation is due to the fact that C_j depends on the size of the reverse-bias voltage instead of being a constant. As a result, it is extremely difficult and time consuming to accurately take this nonlinear capacitance into account when calculating the time to charge or discharge a junction over a large voltage change. A commonly used approximation when analyzing the transient response for large voltage changes is to use an *average size* for the junction capacitance by calculating the junction capacitance at the two extremes of the reverse-bias voltage. Unfortunately, a problem with this approach is that when the diode is forward biased with $V_R \cong -\Phi_0$, Eq. (1.17) "blows up" (i.e., is equal to infinity). To circumvent this

problem, one can instead calculate the charge stored in the junction for the two extreme values of applied voltage (through the use of (1.21)), and then through the use of $Q = CV$, calculate the average capacitance according to

$$C_{j\text{-av}} = \frac{Q(V_2) - Q(V_1)}{V_2 - V_1} \tag{1.34}$$

where V_1 and V_2 are the two voltage extremes [Hodges, 1988].

From (1.21), for an abrupt junction with reverse-bias voltage V_i, we have

$$Q(V_i) = 2C_{j0}\Phi_0\sqrt{1 + \frac{V_i}{\Phi_0}} \tag{1.35}$$

Therefore,

$$C_{j\text{-av}} = 2C_{j0}\Phi_0\frac{\left(\sqrt{1 + \frac{V_2}{\Phi_0}} - \sqrt{1 + \frac{V_1}{\Phi_0}}\right)}{V_2 - V_1} \tag{1.36}$$

One special case often encountered is charging a junction from 0 V to 5 V. For this special case, and using $\Phi_0 = 0.9$ V, we find that

$$C_{j\text{-av}} = 0.56C_{j0} \tag{1.37}$$

Thus, as a rough approximation to quickly estimate the charging time of a junction capacitance from 0 V to 5 V (or vice versa), one can use

$$C_{j\text{-av}} = \frac{C_{j0}}{2} \tag{1.38}$$

It will be seen in the following example that (1.37) compares well with a SPICE simulation.

EXAMPLE 1.6

For the circuit shown in Fig. 1.3, where a reverse-biased diode is being charged from 0 V to 5 V, through a 10-kΩ resistor, calculate the time required to charge the diode from 0 V to 3.5 V. Assume that $C_{j0} = 0.2$ fF/(μm)^2 and that the diode has an area of 20 μm × 5 μm. Compare your answer to that obtained using SPICE. Repeat the question for the case of the diode being discharged from 5 V to 1.5 V.

Solution

The total small-signal capacitance of the junction at 0-V bias voltage is obtained by multiplying 0.2 fF/(μm)^2 by the junction area to obtain

$$C_{j0} = 0.2 \times 10^{-15} \times 20 \times 5 = 0.02 \text{ pF} \tag{1.39}$$

Using (1.37), we have

$$C_{j\text{-av}} = 0.56 \times 0.02 = 0.011 \text{ pF} \tag{1.40}$$

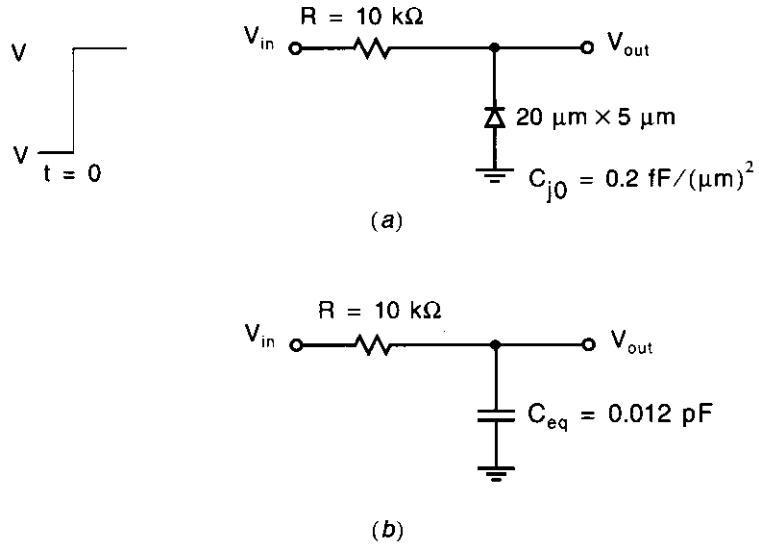


Fig. 1.3 (*a*) The circuit used in Example 1.6; (*b*) its RC approximate equivalent.

resulting in a time constant of

$$\tau = RC_{j\text{-}av} = 0.11 \text{ ns} \tag{1.41}$$

It is not difficult to show that the time it takes for a first-order circuit to rise (or fall) 70 percent of its final value is equal to 1.2τ. Thus, in this case,

$$t_{70\%} = 1.2\tau = 0.13 \text{ ns} \tag{1.42}$$

As a check, the circuit of Fig. 1.3(*a*) was analyzed using SPICE. The input data file was as follows:

```
R 1 2 10k
D 0 2 DMOD
*
VIN 1 0 dc 2.5 PULSE (0 5 0 10p 10p 0.49n 1.0n)
*
.MODEL DMOD D(CJO=0.02E-12)
*
.OPTIONS NUMDGT=5 ITL1=500
.WIDTH OUT=80
.TRAN 0.01n 1.0n
.PRINT TRAN V(2)
.END
```

The SPICE simulation gave a 0-V to 3.5-V rise time of 0.14 ns and a 5-V to 1.5-V fall time of 0.12 ns. These times compare favorably with the 0.13 ns predicted. The reason for the different values of the rise and fall times is the nonlinearity of the junction capacitance. For smaller bias voltages it is larger than that

predicted by (1.37), whereas for larger bias voltages it is smaller. If we use the more accurate approximation of (1.36) for the rise time with $V_2 = 3.5$ and $V_1 = 0$ V, we find

$$C_{j\text{-av}} = 2 \times 0.02 \times \frac{0.9}{3.5}\left(\sqrt{1 + \frac{3.5}{0.9}} - 1\right) = 0.012 \text{ pF} \tag{1.43}$$

Also, for the fall time, we find that

$$C_{j\text{-av}} = 2 \times 0.02 \times \frac{0.9}{1.5 - 5}\left(\sqrt{1 + \frac{1.5}{0.9}} - \sqrt{1 + \frac{5}{0.9}}\right) = 0.010 \text{ pF} \tag{1.44}$$

These more accurate approximations result in

$$t_{+70\%} = 0.144 \text{ ns} \tag{1.45}$$

and

$$t_{-70\%} = 0.114 \text{ ns} \tag{1.46}$$

in closer agreement with SPICE. Normally, the extra accuracy that results from using (1.36) instead of (1.37) is not worth the extra complication because one seldom knows the area of C_{j0} to better than 20 percent accuracy.

Forward-Biased Junctions

A positive voltage applied from the p side to the n side of a diode reduces the electric field opposing the diffusion of the free carriers across the depletion region. It also reduces the width of the depletion region. If this forward-bias voltage is large enough, the carriers will start to diffuse across the junction, resulting in a current flow from the anode to the cathode. For silicon, appreciable diode current starts to occur for a forward-bias voltage around 0.5 V. For germanium and gallium arsenide semiconductor materials, current conduction starts to occur around 0.3 V and 0.9 V, respectively.

When the junction potential is sufficiently lowered for conduction to occur, the carriers diffuse across the junction due to the large gradient in the mobile carrier concentrations. Note that there are more carriers diffusing from the heavily doped side to the lightly doped side than from the lightly doped side to the heavily doped side.

After the carriers cross the depletion region, they greatly increase the *minority charge* at the edge of the depletion region. These minority carriers will diffuse away from the junction toward the bulk. As they diffuse, they recombine with the majority carriers, thereby decreasing their concentration. This concentration gradient of the minority charge (which decreases the farther one gets from the junction) is responsible for the current flow near the junction.

The majority carriers that recombine with the diffusing minority carriers come from the metal contacts at the junctions because of the forward-bias voltage. These majority carriers flow across the bulk, from the contacts to the junction, due to an electric field applied across the bulk. This current flow is called *drift*. It results in

small potential drops across the bulk, especially in the lightly doped side. Typical values of this voltage drop might be 50 mV to 0.1 V, depending primarily on the doping concentration of the lightly doped side, the distance from the contacts to the junction, and the cross-sectional area of the junction.

In the forward-bias region, the current-voltage relationship is exponential and can be shown (see Appendix) to be

$$I_D = I_S e^{V_D/V_T} \tag{1.47}$$

where V_D is the voltage applied across the diode and

$$I_S \propto A_D\left(\frac{1}{N_A} + \frac{1}{N_D}\right) \tag{1.48}$$

I_S is known as the *scale current* and is seen to be proportional to the area of the diode junction, A_D, and inversely proportional to the doping concentrations.

Junction Capacitance of Forward-Biased Diode

When a junction changes from reverse biased (with little current through it) to forward biased (with significant current flow across it), the charge being stored near and across the junction changes. Part of the change in charge is due to the change in the width of the depletion region and therefore the amount of immobile charge stored in it. This change in charge is modelled by the depletion capacitance, C_j, similar to when the junction is reverse biased. An additional change in charge storage is necessary to account for the change of the minority carrier concentration close to the junction required for the diffusion current to exist. For example, if a forward-biased diode current is to double, then the slopes of the minority charge storage at the diode junction edges must double, and this, in turn, implies that the minority charge storage must double. This component is modelled by another capacitance, called the *diffusion capacitance,* and denoted C_d.

The diffusion capacitance can be shown (see Appendix) to be

$$C_d = \tau_T \frac{I_D}{V_T} \tag{1.49}$$

where τ_T is the transit time of the diode. Normally τ_T is specified for a given technology, so that one can calculate the diffusion capacitance. *Note that the diffusion capacitance of a forward-biased junction is proportional to the diode current.*

The total capacitance of the forward-biased junction is the sum of the diffusion capacitance, C_d, and the depletion capacitance, C_j. Thus, the total junction capacitance is given by

$$C_T = C_d + C_j \tag{1.50}$$

For a forward-biased junction, the depletion capacitance, C_j, can be roughly approximated by $2C_{j0}$. The accuracy of this approximation is not critical since the diffusion capacitance is typically much larger than the depletion capacitance.

Finally, it should be mentioned that as a diode is turned off for a short period of time a current will flow in the negative direction until the minority charge is removed. This behavior does not occur in Schottky diodes since they do not have minority charge storage.

Small-Signal Model of a Forward-Biased Diode

A small-signal equivalent model for a forward-biased diode is shown in Fig. 1.4. A resistor, r_d, models the change in the diode voltage, V_D, that occurs when I_D changes. Using (1.47), we have

$$\frac{1}{r_d} = \frac{dI_D}{dV_D} = I_S \frac{e^{V_D/V_T}}{V_T} = \frac{I_D}{V_T} \tag{1.51}$$

This resistance is called the incremental resistance of the diode. For very accurate modelling, it is sometimes necessary to add the series resistance due to the bulk and also the resistance associated with the contacts. Typical values for the contact resistance (caused by the work-function[2] difference between metal and silicon) might be 20 Ω to 40 Ω.

By combining (1.49) and (1.51), we see that an alternative equation for the diffusion capacitance, C_d, is

$$C_d = \frac{\tau_T}{r_d} \tag{1.52}$$

Since for moderate forward-bias currents, $C_d >> C_j$, the total small-signal capacitance is $C_T \cong C_d$, and

$$r_d C_T \cong \tau_T \tag{1.53}$$

Thus, for charging or discharging a forward-biased junction with a current source having an impedance much larger than r_d, the time constant of the charging is approximately equal to the transit time of the diode and is independent of the diode current. For smaller diode currents, where C_j becomes important, the charging or discharging time constant of the circuit becomes larger than τ_T.

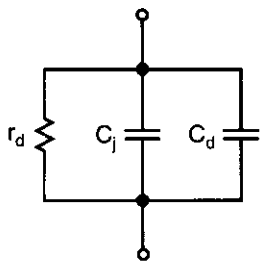


Fig. 1.4 The small-signal model for a forward-biased junction.

2. The work-function of a material is defined as the minimum energy required to remove an electron at the Fermi level to the outside vacuum region.

EXAMPLE 1.7

A given diode has a transit time of 100 ps and is biased at 1 mA. What are the values of its small-signal resistance and diffusion capacitance? Assume room temperature, so that $V_T = kT/q = 26$ mV.

Solution

We have

$$r_d = \frac{V_T}{I_D} = \frac{26 \text{ mV}}{1 \text{ mA}} = 26\ \Omega$$

and

$$C_d = \frac{\tau_T}{r_d} = \frac{100 \text{ ps}}{26\ \Omega} = 3.8 \text{ pF}$$

Note that this diffusion capacitance is over 100 times larger than the total depletion capacitance found in Examples 1.4 and 1.5.

Schottky Diodes

A different type of diode, one often used in microcircuit design, is realized by contacting metal to a lightly doped semiconductor region (rather than a heavily doped region) as shown in Fig. 1.5. Notice that the aluminum anode is in direct contact with a relatively lightly doped n^- region. Because the n^- region is relatively lightly doped, the work-function difference between the aluminum contact and the n^- silicon is larger than would be the case for aluminum contacting to an n^+ region, as occurs at the cathode. This causes a depletion region and, correspondingly, a diode to occur at the interface between the aluminum anode and the n^- silicon region. This diode has different characteristics than a normal pn junction diode. First, its voltage drop when forward

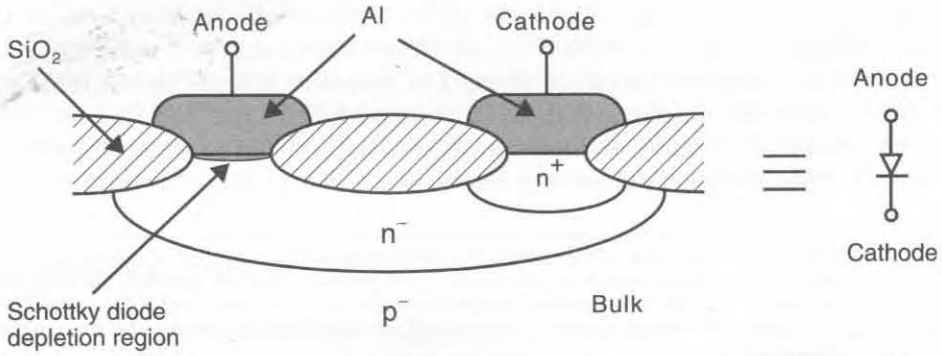


Fig. 1.5 A cross section of a Schottky diode.

biased is smaller. This voltage drop is dependent on the metal used; for aluminum it might be around 0.5 V. More importantly, when the diode is forward biased, there is no minority-charge storage in the lightly doped n^- region. Thus, the small-signal model of a forward-biased Schottky diode has $C_d = 0$ (with reference to Fig. 1.4). The absence of this diffusion capacitance makes the diode much faster. It is particularly faster when turning off, because it is not necessary to remove the minority charge first. Rather, it is only necessary to discharge the depletion capacitance through about 0.2 V.

Schottky diodes have been used extensively in bipolar logic circuits. They are also used in a number of high-speed analog circuits, particularly those realized in gallium arsenide (GaAs) technologies, rather than silicon technologies.

1.2 MOS TRANSISTORS

Presently, the most popular technology for realizing microcircuits makes use of MOS transistors. Unlike most bipolar junction transistor (BJT) technologies, which make dominant use of only one type of transistor (npn transistors in the case of BJT processes[3]), MOS circuits normally use two complementary types of transistors—n-channel and p-channel. While n-channel devices conduct with a positive gate voltage, p-channel devices conduct with a negative gate voltage. Moreover, electrons are used to conduct current in n-channel transistors, while holes are used in p-channel transistors. Microcircuits containing both n-channel and p-channel transistors are called CMOS circuits, for *complementary MOS*. The acronym MOS stands for *metal-oxide semiconductor,* which historically denoted the gate, insulator, and channel region materials, respectively. However, most present CMOS technologies utilize polysilicon gates rather than metal gates.

Before CMOS technology became widely available, most MOS processes made use of only n-channel transistors (NMOS). However, often two different types of n-channel transistors could be realized. One type, enhancement n-channel transistors, is similar to the n-channel transistors realized in CMOS technologies. Enhancement transistors require a positive gate-to-source voltage to conduct current. The other type, depletion transistors, conduct current with a gate-source voltage of 0 V. Depletion transistors were used to create high-impedance loads in NMOS logic gates.

A typical cross section of an n-channel enhancement-type MOS transistor is shown in Fig. 1.6. With no voltage applied to the gate, the n^+ source and drain regions are separated by the p^- substrate. The separation between the drain and the source is called the channel length, L. In present MOS technologies, the minimum channel length is typically between 0.3 μm and 1.0 μm. It should be noted that there is no physical difference between the drain and the source.[4] *The source terminal of an*

3. Most BJT technologies can also realize low-speed lateral pnp transistors. Normally these would only be used to realize current sources as they have low gains and poor frequency responses. Recently, bipolar technologies utilizing high-speed vertical pnp transistors, as well as high-speed npn transistors, have become available and are growing in popularity. These technologies are called complementary bipolar technologies.

4. Large MOS transistors used for power applications might not be realized with symmetric drain and source junctions.

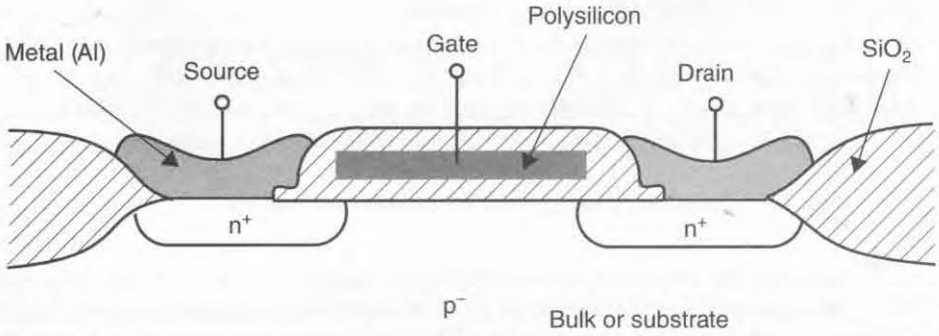


Fig. 1.6 A cross section of a typical n-channel transistor.

n-channel transistor is defined as whichever of the two terminals has a lower voltage. For a p-channel transistor, the source would be the terminal with the higher voltage. When a transistor is turned on, current flows from the drain to the source in an n-channel transistor and from the source to the drain in a p-channel transistor. In both cases, the true carriers travel from the source to drain, but the current directions are different because n-channel carriers (electrons) are negative, whereas p-channel carriers (holes) are positive.

The gate is normally realized using polysilicon, which is heavily doped noncrystalline (or amorphous) silicon. Polysilicon gates are used nowadays (instead of metal) because polysilicon allows the dimensions of the transistor to be realized much more accurately during the patterning of the transistor, which involves what is called a self-aligned process. This higher geometric accuracy results in smaller, faster transistors.

The gate is physically separated from the surface of the silicon by a thin insulator made of silicon dioxide (SiO_2). Thus, the gate is electrically isolated from the channel and affects the channel (and hence, the transistor current) only through electrostatic coupling, similar to capacitive coupling. A typical thickness for the SiO_2 insulator between the gate and the channel is presently from 0.01 μm to 0.03 μm. Since the gate is electrically isolated from the channel, it never conducts dc current. Indeed, the excellent isolation results in leakage currents being almost undetectable. However, because of the inherent capacitances in MOS transistors, transient gate currents do exist when gate voltages are quickly changing.

Normally the p^- substrate (or bulk) is connected to the most negative voltage in a microcircuit. In analog circuits, this might be the negative power supply, but in digital circuits it is normally ground or 0 V. This connection results in all transistors placed in the substrate being surrounded by reverse-biased junctions, which electrically isolate the transistors and thereby prevent conduction through the substrate between transistors (unless, of course, they are connected together through some other means).

Symbols for MOS Transistors

Many symbols have been used to represent MOS transistors. Figure 1.7 shows some of the symbols that have been used to represent n-channel MOS transistors. The symbol in Fig. 1.7(*a*) is often used; note that there is nothing in the symbol to specify

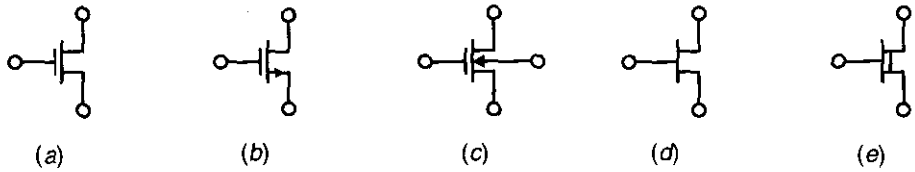


Fig. 1.7 Commonly used symbols for n-channel transistors.

whether the transistor is n-channel or p-channel. A common rule is to assume, when in doubt, that the transistor is an n-channel enhancement transistor. Figure 1.7(*b*) is the most commonly used symbol for an n-channel enhancement transistor and is used throughout this text. The arrow pointing outward on the source indicates that the transistor is n-channel, similar to the convention used for npn transistors, and indicates the direction of hole current.

MOS transistors are actually four-terminal devices, with the substrate being the fourth terminal. In n-channel devices, the p^- substrate is normally connected to the most negative voltage in the microcircuit, whereas for p-channel devices, the n^- substrate is normally connected to the most positive voltage. In these cases the substrate connection is normally not shown in the symbol. However, for CMOS technologies, at least one of the two types of transistors will be formed in a *well* substrate that need not be connected to one of the power supply nodes. For example, an n-well process would form n-channel transistors in a p^- substrate encompassing the entire microcircuit, while the p-channel transistors would be formed in many n-well substrates. In this case, most of the n-well substrates would be connected to the most positive power supply, while some might be connected to other nodes in the circuit (often the well is connected to the source of a transistor that is not connected to the power supply). In these cases, the symbol shown in Fig. 1.7(*c*) can be used to show the substrate connection explicitly. It should be noted that this case is not encountered often in digital circuits and is more common in analog circuits. Sometimes, in the interest of simplicity, the isolation of the gate is not explicitly shown, as is the case of the symbol of Fig. 1.7(*d*). This simple notation is more common for digital circuits in which a large number of transistors are present. Since this symbol is also used for JFET transistors, it will never be used to represent MOS transistors in this text. The last symbol, shown in Fig. 1.7(*e*), denotes an n-channel depletion transistor. The extra line is used to indicate that a physical channel exists for a 0-V gate-source voltage. Depletion transistors were used in older NMOS technologies but are not typically available in CMOS processes.

Figure 1.8 shows some commonly used symbols for p-channel transistors. In this text, the symbol of Fig. 1.8(*a*) will be most often used. The symbol in Fig. 1.8(*c*) is

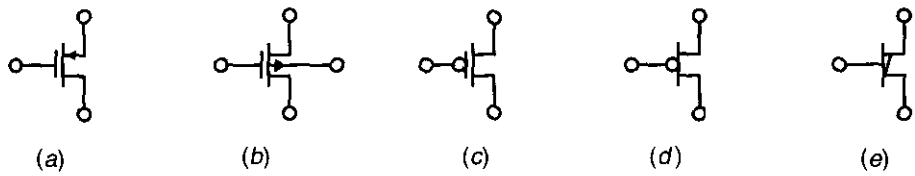


Fig. 1.8 Commonly used symbols for p-channel transistors.

sometimes used in digital circuits, where the circle indicates that a low voltage on the gate turns the transistor on, as opposed to a high voltage for an n-channel transistor (Fig. 1.7(*a*)). The symbols of Fig. 1.8(*d*) or Fig. 1.8(*e*) might be used in larger circuits where many transistors are present, to simplify the drawing somewhat. They will not be used in this text.

Basic Operation

The basic operation of MOS transistors will be described with respect to an n-channel transistor. First, consider the simplified cross sections shown in Fig. 1.9, where the source, drain, and substrate are all connected to ground. In this case, the MOS transistor operates similarly to a capacitor. The gate acts as one plate of the capacitor, and the surface of the silicon, just under the thin insulating SiO_2, acts as the other plate.

If the gate voltage is very negative, as shown in Fig. 1.9(*a*), positive charge will be attracted to the channel region. Since the substrate was originally doped p^-, this negative gate voltage has the effect of simply increasing the channel doping to p^+,

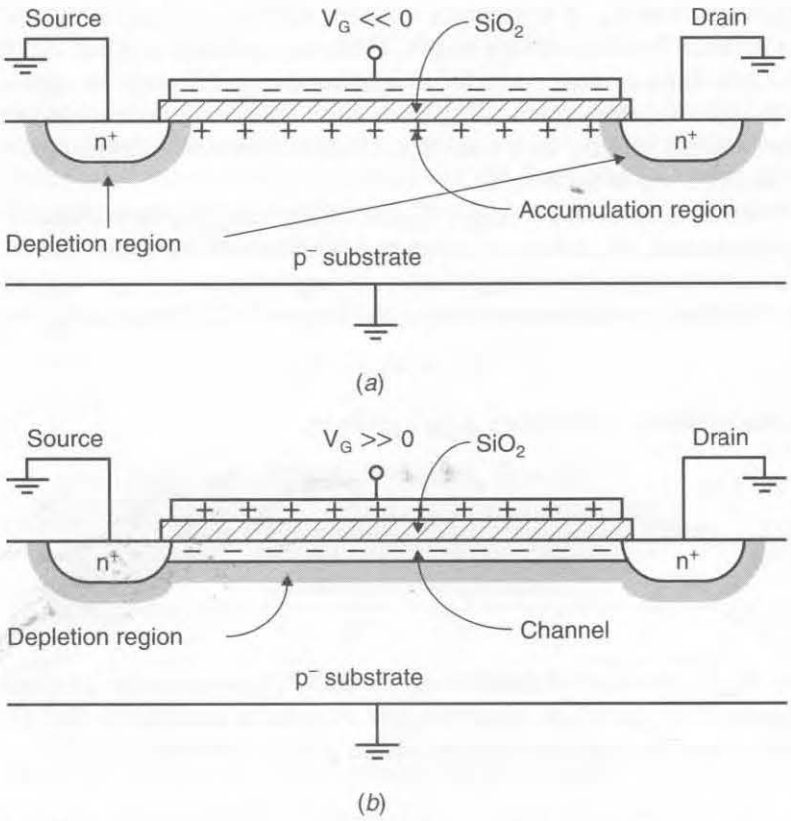


Fig. 1.9 An n-channel MOS transistor. (*a*) $V_G << 0$, resulting in an accumulated channel (no current flow); (*b*) $V_G >> 0$, and the channel is present (current flow possible from drain to source).

resulting in what is called an *accumulated channel.* The n^+ source and drain regions are separated from the p^+-channel region by depletion regions, resulting in the equivalent circuit of two back-to-back diodes. Thus, only leakage current will flow even if one of the source or drain voltages becomes large (unless the drain voltage becomes so large as to cause the transistor to break down).

In the case of a positive voltage being applied to the gate, the opposite situation occurs, as shown in Fig. 1.9(*b*). For small positive gate voltages, the positive carriers in the channel under the gate are initially repulsed and the channel changes from a p^- doping level to a depletion region. As a more positive gate voltage is applied, the gate attracts negative charge from the source and drain regions, and the channel becomes an n region with mobile electrons connecting the drain and source regions.[5] In short, a sufficiently large positive gate-source voltage changes the channel beneath the gate to an n region, and the channel is said to be *inverted.*

The gate-source voltage, for which the concentration of electrons under the gate is equal to the concentration of holes in the p^- substrate far from the gate, is commonly referred to as the *transistor threshold voltage* and denoted V_{tn} (for n-channel transistors). For gate-source voltages larger than V_{tn}, there is an n-type channel present, and conduction between the drain and the source can occur. For gate-source voltages less than V_{tn}, it is normally assumed that the transistor is off and no current flows between the drain and the source. However, it should be noted that this assumption of zero drain-source current for a transistor that is off is only an approximation. In fact, for gate voltages around V_{tn}, there is no abrupt current change, and for gate-source voltages slightly less than V_{tn}, small amounts of *subthreshold current* can flow, as discussed in Section 1.3.

When the gate-source voltage, V_{GS}, is larger than V_{tn}, the channel is present. As V_{GS} is increased, the density of electrons in the channel increases. Indeed, the carrier density, and therefore the charge density, is proportional to $V_{GS} - V_{tn}$, which is often called the *effective gate-source voltage* and denoted V_{eff}. Specifically, define

$$V_{eff} \equiv V_{GS} - V_{tn} \tag{1.54}$$

The charge density of electrons is then given by

$$Q_n = C_{ox}(V_{GS} - V_{tn}) = C_{ox}V_{eff} \tag{1.55}$$

Here, C_{ox} is the gate capacitance per unit area and is given by

$$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}} \tag{1.56}$$

where K_{ox} is the relative permittivity of SiO_2 (approximately 3.9) and t_{ox} is the thickness of the thin oxide under the gate. A point to note here is that (1.55) is only accurate when both the drain and the source voltages are zero.

5. The drain and source regions are sometimes called diffusion regions or junctions for historical reasons. This use of the word *junction* is not synonymous with our previous use, in which it designated a pn interface of a diode.

To obtain the total gate capacitance, (1.56) should be multiplied by the effective gate area, WL, where W is the gate width and L is the effective gate length. These dimensions are shown in Fig. 1.10. Thus the total gate capacitance, C_{gs}, is given by

$$C_{gs} = WLC_{ox} \tag{1.57}$$

and the total charge of the channel, $Q_{T\text{-}n}$, is given by

$$Q_{T\text{-}n} = WLC_{ox}(V_{GS} - V_{tn}) = WLC_{ox}V_{eff} \tag{1.58}$$

The gate capacitance, C_{gs}, is one of the major load capacitances that circuits must be capable of driving. Gate capacitances are also important when one is calculating *charge injection,* which occurs when a MOS transistor is being turned off because the channel charge, $Q_{T\text{-}n}$, must flow from under the gate out through the terminals to other places in the circuit.

Next, if the drain voltage is increased above 0 V, a drain-source potential difference exists. This difference results in current flowing from the drain to the source.[6] The relationship between V_{DS} and the drain-source current, I_D, is the same as for a resistor, assuming V_{DS} is small. This relationship is given [Sze, 1981] by

$$I_D = \mu_n Q_n \frac{W}{L} V_{DS} \tag{1.59}$$

where $\mu_n \cong 0.06\ m^2/Vs$ is the mobility of electrons near the silicon surface, and Q_n is the charge concentration of the channel per unit area (looking from the top down). Note that as the channel length increases, the drain-source current decreases, whereas this current increases as either the charge density or the transistor width increases. Using (1.58) and (1.59) results in

$$I_D = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{tn})V_{DS} = \mu_n C_{ox} \frac{W}{L} V_{eff} V_{DS} \tag{1.60}$$

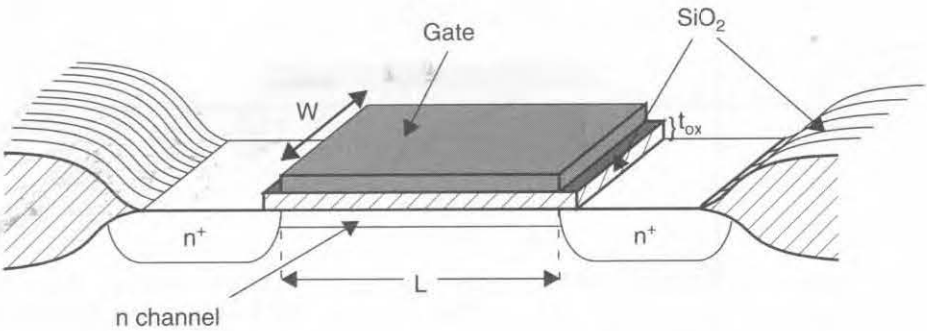


Fig. 1.10 The important dimensions of a MOS transistor.

6. The current is actually conducted by negative carriers (electrons) flowing from the source to the drain. Negative carriers flowing from source to drain results in a positive current from drain to source, I_{DS}.

where it should be emphasized that this relationship is only valid for drain-source voltages near zero (i.e., V_{DS} much smaller than V_{eff}).

As the drain-source voltage increases, the channel charge concentration decreases at the drain end. This decrease is due to the smaller gate-to-channel voltage difference across the thin gate oxide as one moves closer to the drain. In other words, since the drain voltage is assumed to be at a higher voltage than the source, there is an increasing voltage gradient from the source to the drain, resulting in a smaller gate-to-channel voltage near the drain. Since the charge density at a distance x from the source end of the channel is proportional to $V_G - V_{ch}(x) - V_{tn}$, as $V_G - V_{ch}(x)$ decreases, the charge density also decreases.[7] This effect is illustrated in Fig. 1.11.

Note that at the drain end of the channel, we have

$$V_G - V_{ch}(L) = V_{GD} \tag{1.61}$$

For small V_{DS}, we saw from (1.60) that I_D was linearly related to V_{DS}. However, as V_{DS} increases, and the charge density decreases near the drain, the relationship becomes nonlinear. In fact, the linear relationship for I_D versus V_{DS} flattens for larger V_{DS}, as shown in Fig. 1.12.

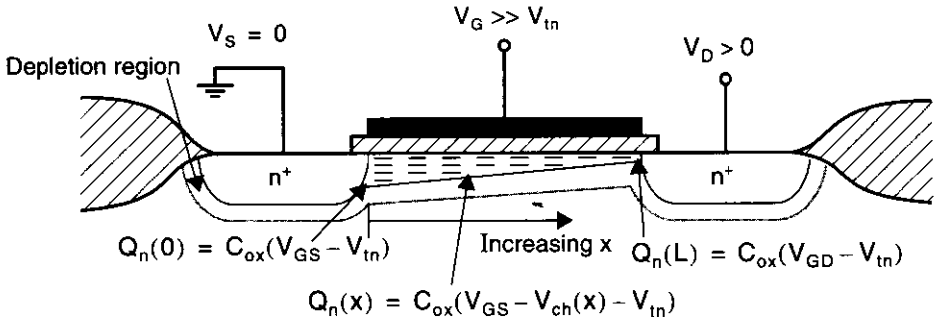


Fig. 1.11 The channel charge density for $V_{DS} > 0$.

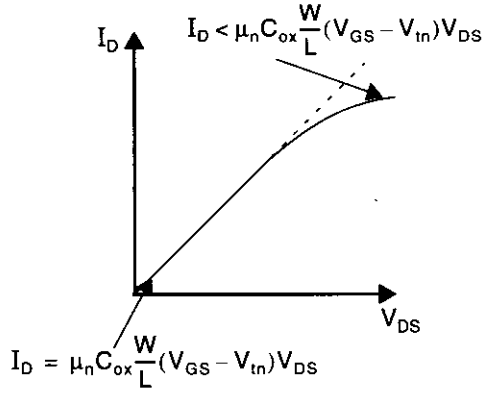


Fig. 1.12 For V_{DS} not close to zero, the I_D versus V_{DS} relationship is no longer linear.

7. $V_G - V_{CH}(x)$ is the gate-to-channel voltage drop at distance x from the source end, with V_G being the same everywhere in the gate, since the gate material is highly conductive.

As the drain voltage is increased, at some point the gate-to-channel voltage at the drain end will decrease to the threshold value V_{tn} — the minimum gate-to-channel voltage needed for n carriers in the channel to exist. Thus, at the drain end, the channel becomes *pinched off*, as shown in Fig. 1.13. This pinch-off occurs at $V_{GD} = V_{tn}$, since the channel voltage at the drain end is simply equal to V_D. Thus, pinch-off occurs for

$$V_{DG} > -V_{tn} \tag{1.62}$$

Denoting $V_{DS\text{-}sat}$ as the drain-source voltage when the channel becomes pinched off, we can substitute $V_{DG} = V_{DS} - V_{GS}$ into (1.62) and find an equivalent pinch-off expression

$$V_{DS} > V_{DS\text{-}sat} \tag{1.63}$$

where $V_{DS\text{-}sat}$ is given[8] by

$$V_{DS\text{-}sat} = V_{GS} - V_{tn} = V_{eff} \tag{1.64}$$

The electron carriers travelling through the pinched-off drain region are velocity saturated, similar to a gas under pressure travelling through a very small tube. If the drain-gate voltage rises above this critical pinch-off voltage of $-V_{tn}$, the charge concentration in the channel remains constant (to a first-order approximation) and the drain current no longer increases with increasing V_{DS}. The result is the current-voltage relationship shown in Fig. 1.14 for a given gate-source voltage. In the region of operation where $V_{DS} > V_{DS\text{-}sat}$, the drain current is independent of V_{DS} and is called the *active region*.[9] The region where I_D changes with V_{DS} is called the *triode region*. When MOS transistors are used in analog amplifiers, they almost always are biased in the active region. When they are used in digital logic gates, they often operate in both regions.

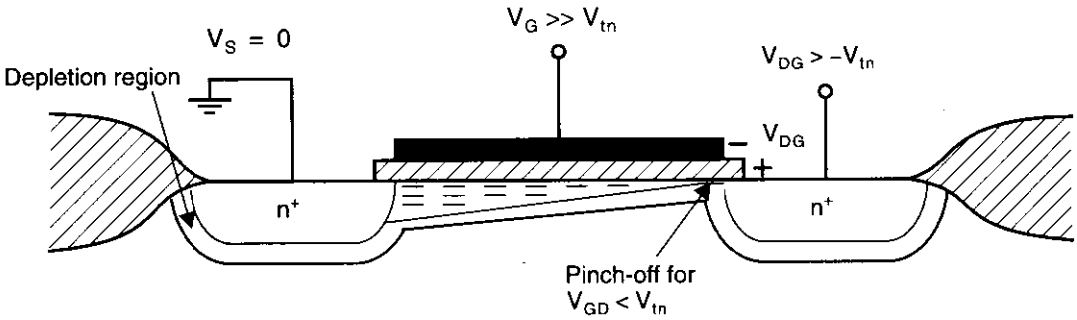


Fig. 1.13 When V_{DS} is increased so that $V_{GD} < V_{tn}$, the channel becomes pinched off at the drain end.

8. Because of the body effect, the threshold voltage at the drain end of the transistor is increased, resulting in the true value of $V_{DS\text{-}sat}$ being slightly lower than V_{eff}.
9. Historically, the active region was called the saturation region, but this led to confusion because in the case of bipolar transistors, the saturation region occurs for small V_{CE}, whereas for MOS transistors it occurs for large V_{DS}. The renaming of the saturation region to the active region is becoming widely accepted.

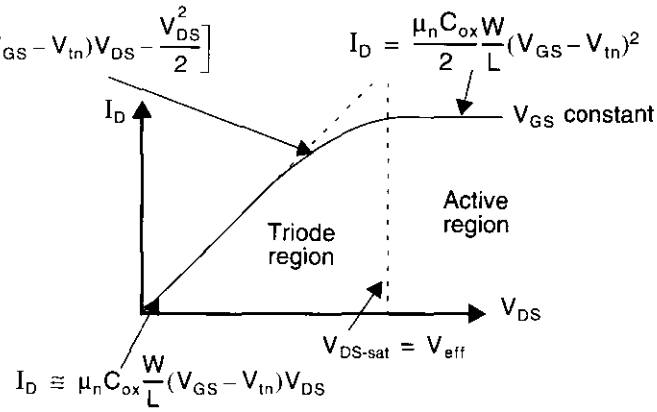


Fig. 1.14 The I_D versus V_{DS} curve for an ideal MOS transistor. For $V_{DS} > V_{DS\text{-}sat}$, I_D is approximately constant.

Before proceeding, it is worth discussing the terms *weak, moderate,* and *strong inversion.* As just discussed, a gate-source voltage greater than V_{tn} results in an inverted channel, and drain-source current can flow. However, as the gate-source voltage is increased, the channel does not become inverted (i.e., n-region) suddenly, but rather gradually. Thus, it is useful to define three regions of channel inversion with respect to the gate-source voltage. In most circuit applications, noncutoff MOSFET transistors are operated in strong inversion, with $V_{eff} > 100$ mV (many prudent circuit designers use a minimum value of 200 mV). As the name suggests, strong inversion occurs when the channel is strongly inverted. It should be noted that all the equation models in this section assume strong inversion operation. Weak inversion occurs when V_{GS} is approximately 100 mV or more below V_{tn} and is discussed as subthreshold operation in Section 1.3. Finally, moderate inversion is the region between weak and strong inversion.

Large-Signal Modelling

The *triode region equation* for a MOS transistor relates the drain current to the gate-source and drain-source voltages. It can be shown (see Appendix) that this relationship is given by

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS}-V_{tn})V_{DS}-\frac{V_{DS}^2}{2}\right] \tag{1.65}$$

As V_{DS} increases, I_D increases until the drain end of the channel becomes pinched off, and then I_D no longer increases. This pinch-off occurs for $V_{DG} = -V_{tn}$, or approximately,

$$V_{DS} = V_{GS} - V_{tn} = V_{eff} \tag{1.66}$$

Right at the edge of pinch-off, the drain current resulting from (1.65) and the drain current in the active region (which, to a first-order approximation, is constant with

respect to V_{DS}) must have the same value. Therefore, the *active region equation* can be found by substituting (1.66) into (1.65), resulting in

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 \tag{1.67}$$

For $V_{DS} > V_{eff}$, the current stays constant at the value given by (1.67), ignoring second-order effects such as the finite output impedance of the transistor. This equation is perhaps the most important one that describes the large-signal operation of a MOS transistor. It should be noted here that (1.67) represents a squared current-voltage relationship for a MOS transistor in the active region. In the case of a BJT transistor, an exponential current-voltage relationship exists in the active region.

As just mentioned, (1.67) implies that the drain current, I_D, is independent of the drain-source voltage. This independence is only true to a first-order approximation. The major source of error is due to the channel length shrinking as V_{DS} increases. To see this effect, consider Fig. 1.15, which shows a cross section of a transistor in the active region. A pinched-off region with very little charge exists between the drain and the channel. The voltage at the end of the channel closest to the drain is fixed at $V_{GS} - V_{tn} = V_{eff}$. The voltage difference between the drain and the near end of the channel lies across a short depletion region often called the *pinch-off region*. As V_{DS} becomes larger than V_{eff}, this depletion region surrounding the drain junction increases its width in a square-root relationship with respect to V_{DS}. This increase in the width of the depletion region surrounding the drain junction decreases the effective channel length. In turn, this decrease in effective channel length increases the drain current, resulting in what is commonly referred to as *channel-length modulation*.

To derive an equation to account for channel-length modulation, we first make use of (1.11) and denote the width of the depletion region by x_d, resulting in

$$\begin{aligned} x_d &\cong k_{ds}\sqrt{V_{D\text{-}ch} + \Phi_0} \\ &= k_{ds}\sqrt{V_{DG} + V_{tn} + \Phi_0} \end{aligned} \tag{1.68}$$

where

$$k_{ds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}} \tag{1.69}$$

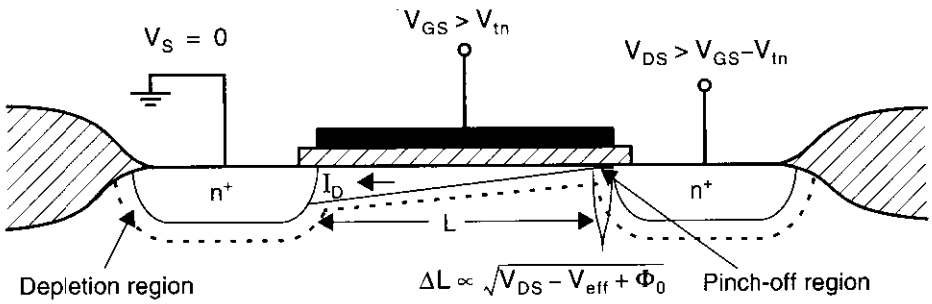


Fig. 1.15 Channel length shortening for $V_{DS} > V_{eff}$.

and has units of $m/\sqrt{V}$. Note that N_A is used here since the n-type drain region is more heavily doped than the p-type channel (i.e., $N_D >> N_A$). By writing a Taylor approximation for I_D around its operating value of $V_{DS} = V_{GS} - V_{tn} = V_{eff}$, we find I_D to be given by

$$I_D = I_{D\text{-sat}} + \left(\frac{\partial I_D}{\partial L}\right)\left(\frac{\partial L}{\partial V_{DS}}\right)\Delta V_{DS} \cong I_{D\text{-sat}}\left(1 + \frac{k_{ds}(V_{DS} - V_{eff})}{2L\sqrt{V_{DG} + V_{tn} + \Phi_0}}\right) \tag{1.70}$$

where $I_{D\text{-sat}}$ is the drain current when $V_{DS} = V_{eff}$, or equivalently, the drain current when the channel-length modulation is ignored. Note that in deriving the final equation of (1.70), we have used the relationship $\partial L/\partial V_{DS} = -\partial x_d/\partial V_{DS}$. Usually, (1.70) is written as

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2[1 + \lambda(V_{DS} - V_{eff})] \tag{1.71}$$

where λ is the output impedance constant (in units of V^{-1}) given by

$$\lambda = \frac{k_{ds}}{2L\sqrt{V_{DG} + V_{tn} + \Phi_0}} = \frac{k_{ds}}{2L\sqrt{V_{DS} - V_{eff} + \Phi_0}} \tag{1.72}$$

Equation (1.71) is accurate until V_{DS} is large enough to cause second-order effects, often called *short-channel effects*. For example, (1.71) assumes that current flow down the channel is not *velocity-saturated* (i.e., increasing the electric field no longer increases the carrier speed). Velocity saturation commonly occurs in new technologies that have very short channel lengths and therefore large electric fields. If V_{DS} becomes large enough so short-channel effects occur, I_D increases more than is predicted by (1.71). Of course, for quite large values of V_{DS}, the transistor will eventually break down.

A plot of I_D versus V_{DS} for different values of V_{GS} is shown in Fig. 1.16. Note that in the active region, the small (but nonzero) slope indicates the small dependence of I_D on V_{DS}.

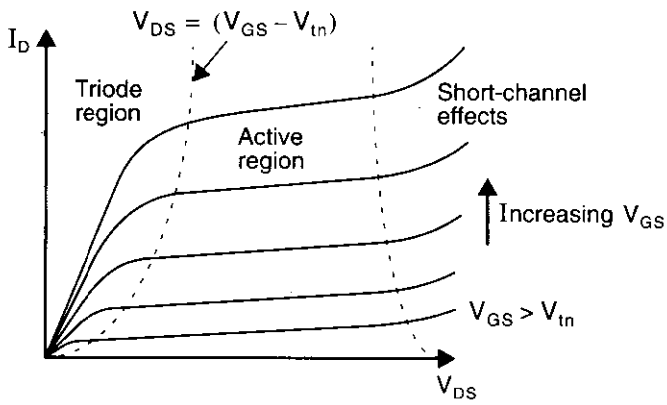


Fig. 1.16 I_D versus V_{DS} for different values of V_{GS}.

EXAMPLE 1.8

Find I_D for an n-channel transistor that has doping concentrations of $N_D = 10^{25}$, $N_A = 10^{22}$, $\mu_n C_{ox} = 92\ \mu A/V^2$, $W/L = 20\ \mu m/2\ \mu m$, $V_{GS} = 1.2$ V, $V_{tn} = 0.8$ V, and $V_{DS} = V_{eff}$. Assuming λ remains constant, estimate the new value of I_D if V_{DS} is increased by 0.5 V.

Solution

From (1.69), we have

$$k_{ds} = \sqrt{\frac{2 \times 11.8 \times 8.854 \times 10^{-12}}{1.6 \times 10^{-19} \times 10^{22}}} = 362 \times 10^{-9}\ m/\sqrt{V}$$

which is used in (1.72) to find λ as

$$\lambda = \frac{362 \times 10^{-9}}{2 \times 2 \times 10^{-6} \times \sqrt{0.9}} = 95.3 \times 10^{-3}\ V^{-1}$$

Using (1.71), we find for $V_{DS} = V_{eff} = 0.4$ V,

$$I_{D1} = \left(\frac{92 \times 10^{-6}}{2}\right)\left(\frac{20}{2}\right)(0.4)^2(1) = 73.6\ \mu A$$

In the case where $V_{DS} = V_{eff} + 0.5\ V = 0.9\ V$, we have

$$I_{D2} = 73.6\ \mu A \times (1 + \lambda \times 0.5) = 77.1\ \mu A$$

Note that this example shows almost a 5 percent increase in drain current for a 0.5 V increase in drain-source voltage.

Body Effect

The large-signal equations in the preceding section were based on the assumption that the source voltage was the same as the substrate (i.e., bulk) voltage. However, often the source and substrate can be at different voltage potentials. In these situations, a second-order effect exists that is modelled as an increase in the threshold voltage, V_{tn}, as the source-to-substrate reverse-bias voltage increases. This effect, typically called the *body effect*, is more important for transistors in a well of a CMOS process where the substrate doping is higher. It should be noted that the body effect is often important in analog circuit designs and should not be ignored without consideration.

To account for the body effect, it can be shown (see Appendix at the end of this chapter) that the threshold voltage of an n-channel transistor is now given by

$$V_{tn} = V_{tn0} + \gamma(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}) \tag{1.73}$$

where V_{tn0} is the threshold voltage with zero V_{SB} (i.e., source-to-substrate voltage),

and

$$\gamma = \frac{\sqrt{2qN_A K_s \varepsilon_0}}{C_{ox}} \tag{1.74}$$

The factor γ is often called the *body-effect constant* and has units of $\sqrt{V}$. Notice that γ is proportional to $\sqrt{N_A}$,[10] so the body effect is larger for transistors in a well where typically the doping is higher than the substrate of the microcircuit.

p-Channel Transistors

All of the preceding equations have been presented for n-channel enhancement transistors. *In the case of p-channel transistors, these equations can also be used if a negative sign is placed in front of every voltage variable.* Thus, V_{GS} becomes V_{SG}, V_{DS} becomes V_{SD}, V_{tn} becomes $-V_{tp}$, and so on. The condition required for conduction is now $V_{SG} > V_{tp}$, where V_{tp} is now a negative quantity for an enhancement p-channel transistor.[11] The requirement on the source-drain voltage for a p-channel transistor to be in the active region is $V_{SD} > V_{SG} + V_{tp}$. The equations for I_D, in both regions, remain unchanged, because all voltage variables are squared, resulting in positive hole current flow from the source to the drain in p-channel transistors. For n-channel depletion transistors, the only difference is that $V_{td} < 0$ V. A typical value might be $V_{td} = -2$ V.

Small-Signal Modelling in the Active Region

The most commonly used small-signal model for a MOS transistor operating in the active region is shown in Fig. 1.17. We first consider the dc parameters in which all the capacitors are ignored (i.e., replaced by open circuits). This leads to the low-frequency, small-signal model shown in Fig. 1.18. The voltage-controlled current source, $g_m v_{gs}$, is the most important component of the model, with the transistor transconductance g_m defined as

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{1.75}$$

In the active region, we use (1.67), which is repeated here for convenience,

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 \tag{1.76}$$

10. For an n-channel transistor. For a p-channel transistor, γ is proportional to N_D.
11. It is possible to realize depletion p-channel transistors, but these are of little value and seldom worth the extra processing involved. Depletion n-channel transistors are also seldom encountered in CMOS microcircuits, although they might be worth the extra processing involved in some applications, especially if they were in a well.

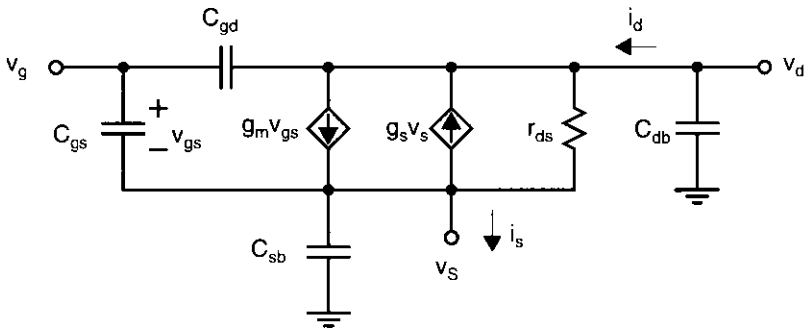


Fig. 1.17 The small-signal model for a MOS transistor in the active region.

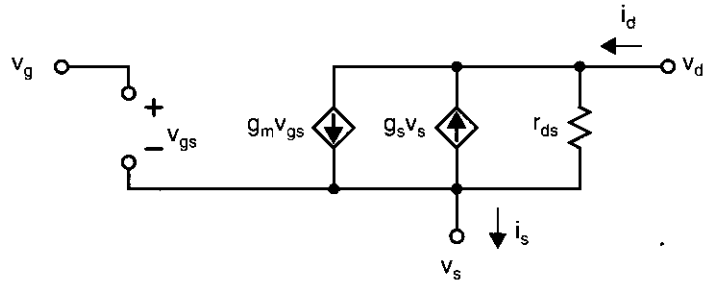


Fig. 1.18 The low-frequency, small-signal model for an active MOS transistor.

and we apply the derivative shown in (1.75) to obtain

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{tn}) = \mu_n C_{ox}\frac{W}{L}V_{eff} \tag{1.77}$$

or equivalently,

$$g_m = \mu_n C_{ox}\frac{W}{L}V_{eff} \tag{1.78}$$

where the effective gate-source voltage, V_{eff}, is defined as $V_{eff} \equiv V_{GS} - V_{tn}$. Thus, we see that the transconductance of a MOS transistor is directly proportional to V_{eff}.

Sometimes it is desirable to express g_m in terms of I_D rather than V_{GS}. From (1.76), we have

$$V_{GS} = V_{tn} + \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{1.79}$$

The second term in (1.79) is the effective gate-source voltage, V_{eff}, where

$$V_{eff} = V_{GS} - V_{tn} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{1.80}$$

Substituting (1.80) in (1.78) results in an alternate expression for g_m.

$$g_m = \sqrt{2\mu_n C_{ox}\frac{W}{L}I_D} \tag{1.81}$$

Thus, the transistor transconductance is proportional to $\sqrt{I_D}$ for a MOS transistor, whereas it is proportional to I_C for a BJT.

A third expression for g_m is found by rearranging (1.81) and then using (1.80) to obtain

$$g_m = \frac{2I_D}{V_{eff}} \tag{1.82}$$

Note that this expression is independent of $\mu_n C_{ox}$ and W/L, and it relates the transconductance to the ratio of drain current to effective gate-source voltage. This simple relationship can be quite useful during an initial circuit design.

The second voltage-controlled current-source in Fig. 1.18, shown as $g_s v_s$, models the body effect on the small-signal drain current, i_d. When the source is connected to small-signal ground, or when its voltage does not change appreciably, then this current source can be ignored. When the body effect cannot be ignored, we have

$$g_s = \frac{\partial I_D}{\partial V_{SB}} = \frac{\partial I_D}{\partial V_{tn}}\frac{\partial V_{tn}}{\partial V_{SB}} \tag{1.83}$$

From (1.76) we have

$$\frac{\partial I_D}{\partial V_{tn}} = -\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn}) = -g_m \tag{1.84}$$

Using (1.73), which gives V_{tn} as

$$V_{tn} = V_{tn0} + \gamma(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}) \tag{1.85}$$

we have

$$\frac{\partial V_{tn}}{\partial V_{SB}} = \frac{\gamma}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{1.86}$$

The negative sign of (1.84) is eliminated by subtracting the current $g_s v_s$ from the major component of the drain current, $g_m v_{gs}$, as shown in Fig. 1.18. Thus, using (1.84) and (1.86), we have

$$g_s = \frac{\gamma g_m}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{1.87}$$

Note that although g_s is nonzero for $V_{SB} = 0$, if the source is connected to the bulk, ΔV_{SB} is zero, and so the effect of g_s does not need to be taken into account. However, if the source happens to be biased at the same potential as the bulk but is not

directly connected to it, then the effect of g_s should be taken into account since ΔV_{SB} is not necessarily zero.

The resistor, r_{ds}, shown in Fig. 1.18, accounts for the finite output impedance (i.e., it models the channel-length modulation and its effect on the drain current due to changes in V_{DS}). Using (1.71), repeated here for convenience,

$$I_D = \frac{\mu_n C_{ox}}{2}\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2[1 + \lambda(V_{DS} - V_{eff})] \tag{1.88}$$

we have

$$\frac{1}{r_{ds}} = g_{ds} = \frac{\partial I_D}{\partial V_{DS}} = \lambda\left(\frac{\mu_n C_{ox}}{2}\right)\left(\frac{W}{L}\right)(V_{GS} - V_{tn})^2 = \lambda I_{D\text{-sat}} \cong \lambda I_D \tag{1.89}$$

where the approximation assumes λ is small, such that we can approximate the drain bias current as being the same as $I_{D\text{-sat}}$. Thus,

$$r_{ds} \cong \frac{1}{\lambda I_D} \tag{1.90}$$

where

$$\lambda = \frac{k_{ds}}{2L\sqrt{V_{DS} + (-V_{eff}) + \Phi_0}} \tag{1.91}$$

and

$$k_{ds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}} \tag{1.92}$$

It should be noted here that (1.90) is often empirically adjusted to take into account second-order effects.

EXAMPLE 1.9

Derive the low-frequency model parameters for an n-channel transistor that has doping concentrations of $N_D = 10^{25}$, $N_A = 10^{22}$, $\mu_n C_{ox} = 92\ \mu A/V^2$, $W/L = 20\ \mu m/2\ \mu m$, $V_{GS} = 1.2$ V, $V_{tn} = 0.8$ V, and $V_{DS} = V_{eff}$. Assume $\gamma = 0.5\ \sqrt{V}$ and $V_{SB} = 0.5$ V. What is the new value of r_{ds} if the drain-source voltage is increased by 0.5 V?

Solution

Since these parameters are the same as in Example 1.8, we have

$$g_m = \frac{2I_D}{V_{eff}} = \frac{2 \times 73.6\ \mu A}{0.4\ V} = 0.368\ mA/V$$

and from (1.87), we have

$$g_s = \frac{0.5 \times 0.368 \times 10^{-3}}{2\sqrt{0.5 + 1.8}} = 0.061 \text{ mA/V}$$

Note that this source-bulk transconductance value is about 1/6 that of the gate-source transconductance.

For r_{ds}, we use (1.90) to find

$$r_{ds} = \frac{1}{95.3 \times 10^{-3} \times 73.6 \times 10^{-6}} = 143 \text{ k}\Omega$$

At this point, it is interesting to calculate the gain $g_m r_{ds} = 52.6$, which is the largest voltage gain this single transistor can achieve for these operating bias conditions. As we will see, this gain of 52.6 is much smaller than the corresponding single-transistor gain in a bipolar transistor.

Recalling that $V_{eff} = 0.4$ V, if V_{DS} is increased to 0.9 V, the new value for λ is

$$\lambda = \frac{362 \times 10^{-9}}{2(2 \times 10^{-6})\sqrt{1.4}} = 76.4 \times 10^{-3} \text{ V}^{-1}$$

resulting in a new value of r_{ds} given by

$$r_{ds} = \frac{1}{\lambda I_{D2}} = \frac{1}{76.4 \times 10^{-3} \times 77.1 \ \mu\text{A}} = 170 \text{ k}\Omega$$

An alternate low-frequency model, known as a T model, is shown in Fig. 1.19. This *T model* can often result in simpler equations and is most often used by experienced designers for a quick analysis. At first glance, it might appear that this model allows for nonzero gate current, but a quick check confirms that the drain current must always equal the source current, and, therefore, the gate current must always be zero. For this reason, when using the T model, one assumes from the beginning that the gate current is zero.

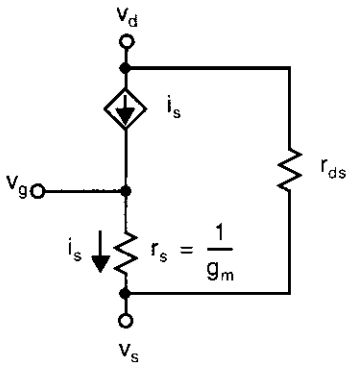


Fig. 1.19 The small-signal, low-frequency T model for an active MOS transistor (the body effect is not modelled).

EXAMPLE 1.10

Find the T model parameter, r_s, for the transistor in Example 1.9.

Solution

The value of r_s is simply the inverse of g_m, resulting in

$$r_s = \frac{1}{g_m} = \frac{1}{0.368 \times 10^{-3}} = 2.72\ \text{k}\Omega$$

The value of r_{ds} remains the same, either 143 kΩ or 170 kΩ, depending on the drain-source voltage.

Most of the capacitors in the small-signal model are related to the physical transistor. Shown in Fig. 1.20 is a cross section of a MOS transistor, where the parasitic capacitances are shown at the appropriate locations. The largest capacitor in Fig. 1.20 is C_{gs}. This capacitance is primarily due to the change in channel charge as a result of a change in V_{GS}. It can be shown [Tsividis, 1987] that C_{gs} is approximately given by

$$C_{gs} \cong \frac{2}{3} WLC_{ox} \tag{1.93}$$

When accuracy is important, an additional term should be added to (1.93) to take into account the overlap between the gate and source junction, which should include the *fringing capacitance* (fringing capacitance is due to boundary effects). This additional component is given by

$$C_{ov} = WL_{ov}C_{ox} \tag{1.94}$$

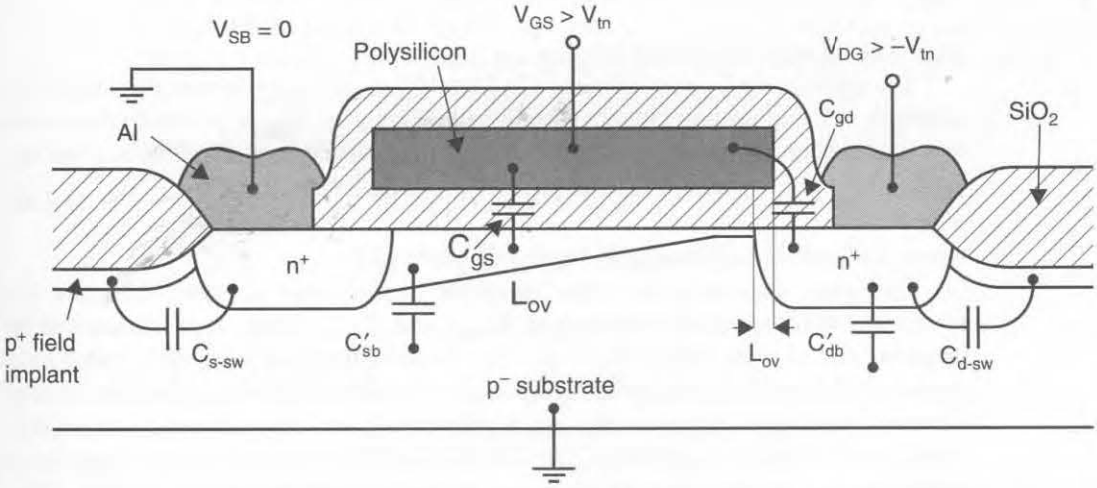


Fig. 1.20 A cross section of an n-channel MOS transistor showing the small-signal capacitances.

where L_{ov} is the overlap distance and is usually empirically derived. Thus,

$$C_{gs} = WC_{ox}\left(\frac{2}{3}L + L_{ov}\right) \tag{1.95}$$

when higher accuracy is needed.

The next largest capacitor in Fig. 1.20 is C'_{sb}, the capacitor between the source and the substrate. This capacitor is due to the depletion capacitance of the reverse-biased source junction, and it includes the channel-to-bulk capacitance (assuming the transistor is on). Its size is given by

$$C'_{sb} = (A_s + A_{ch})C_{js} \tag{1.96}$$

where A_s is the area of the source junction, A_{ch} is the area of the channel (i.e., WL) and C_{js} is the depletion capacitance of the source junction, given by

$$C_{js} = \frac{C_{j0}}{\sqrt{1 + \frac{V_{SB}}{\Phi_0}}} \tag{1.97}$$

Note that the total area of the effective source includes the original area of the junction (when no channel is present) plus the effective area of the channel.

The depletion capacitance of the drain is smaller because it does not include the channel area. Here, we have

$$C'_{db} = A_d C_{jd} \tag{1.98}$$

where

$$C_{jd} = \frac{C_{j0}}{\sqrt{1 + \frac{V_{DB}}{\Phi_0}}} \tag{1.99}$$

and A_d is the area of the drain junction.

The capacitance C_{gd}, sometimes called the *Miller-capacitor,* is important when the transistor is being used in circuits with large voltage gain. C_{gd} is primarily due to the overlap between the gate and the drain and fringing capacitance. Its value is given by

$$C_{gd} = C_{ox}WL_{ov} \tag{1.100}$$

where, once again, L_{ov} is usually empirically derived.

Two other capacitors are often important in integrated circuits. These are the source and drain *sidewall capacitances,* $C_{s\text{-}sw}$ and $C_{d\text{-}sw}$. These capacitances can be large because of some highly doped p^+ regions under the thick field oxide called *field implants.* The major reason these regions exist is to ensure there is no leakage current between transistors. Because they are highly doped and they lie beside the highly doped source and drain junctions, the sidewall capacitances can result in large additional capacitances that must be taken into account in determining C_{sb} and C_{db}. The sidewall capacitances are especially important in modern technologies as dimensions

shrink. For the source, the sidewall capacitance is given by

$$C_{s\text{-}sw} = P_s C_{j\text{-}sw} \tag{1.101}$$

where P_s is the length of the perimeter of the source junction, excluding the side adjacent to the channel, and

$$C_{j\text{-}sw} = \frac{C_{j\text{-}sw0}}{\sqrt{1 + \frac{V_{SB}}{\Phi_0}}} \tag{1.102}$$

It should be noted that $C_{j\text{-}sw0}$, the sidewall capacitance per unit length at 0-V bias voltage, can be quite large because the field implants are heavily doped.

The situation is similar for the drain sidewall capacitance, $C_{d\text{-}sw}$,

$$C_{d\text{-}sw} = P_d C_{j\text{-}sw} \tag{1.103}$$

where P_d is the drain perimeter excluding the portion adjacent to the gate.

Finally, the source-bulk capacitance, C_{sb}, is given by

$$C_{sb} = C'_{sb} + C_{s\text{-}sw} \tag{1.104}$$

with the drain-bulk capacitance, C_{db}, given by

$$C_{db} = C'_{db} + C_{d\text{-}sw} \tag{1.105}$$

EXAMPLE 1.11

An n-channel transistor is modelled as having the following capacitance parameters: $C_j = 2.4 \times 10^{-4}$ pF/(μm)^2, $C_{j\text{-}sw} = 2.0 \times 10^{-4}$ pF/μm, $C_{ox} = 1.9 \times 10^{-3}$ pF/(μm)^2, $C_{gs\text{-}ov} = C_{gd\text{-}ov} = 2.0 \times 10^{-4}$ pF/μm. Find the capacitances C_{gs}, C_{gd}, C_{db}, and C_{sb} for a transistor having $W = 100$ μm and $L = 2$ μm. Assume the source and drain junctions extend 4 μm beyond the gate, so that the source and drain areas are $A_s = A_d = 400$ (μm)^2 and the perimeter of each is $P_s = P_d = 108$ μm.

Solution

We calculate the various capacitances as follows:

$$C_{gs} = \left(\frac{2}{3}\right) WLC_{ox} + C_{gs\text{-}ov} \times W = 0.27 \text{ pF}$$

$$C_{gd} = C_{gd\text{-}ov} \times W = 0.02 \text{ pF}$$

$$C_{sb} = C_j(A_s + WL) + (C_{j\text{-}sw} \times P_s) = 0.17 \text{ pF}$$

$$C_{db} = (C_j \times A_d) + (C_{j\text{-}sw} \times P_d) = 0.12 \text{ pF}$$

Note that the source-bulk and drain-bulk capacitances are significant compared to the gate-source capacitance. Thus, for high-speed circuits, it is important to

keep the areas and perimeters of drain and source junctions as small as possible (possibly by sharing junctions between transistors, as seen in the next chapter).

Small-Signal Modelling in the Triode and Cutoff Regions

The low-frequency, small-signal model of a MOS transistor in the triode region (which is sometimes referred to as the linear region) is a resistor. Using (1.65), the large-signal equation for I_D in the triode region,

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{1.106}$$

results in

$$\frac{1}{r_{ds}} = g_{ds} = \frac{dI_D}{dV_{DS}} = \mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn} - V_{DS}) \tag{1.107}$$

where r_{ds} is the small-signal drain-source resistance (and g_{ds} is the conductance). For the common case of V_{DS} near zero, we have

$$g_{ds} = \frac{1}{r_{ds}} = \mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{tn}) = \mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff} \tag{1.108}$$

which is similar to the I_D-versus-V_{DS} relationship given earlier in (1.60).

EXAMPLE 1.12

For the transistor of Example 1.9, find the triode model parameters when V_{DS} is near zero.

Solution

From (1.108), we have

$$g_{ds} = 92 \times 10^{-6} \times \left(\frac{20}{2}\right) \times 0.4 = 0.368 \text{ mA/V}$$

Note that this conductance value is the same as the transconductance of the transistor, g_m, in the active region. The resistance, r_{ds}, is simply $1/g_{ds}$, resulting in $r_{ds} = 2.72$ kΩ.

The accurate modelling of the high-frequency operation of a transistor in the triode region is nontrivial (even with the use of a computer simulation). A moderately accurate model is shown in Fig. 1.21, where the gate-to-channel capacitance

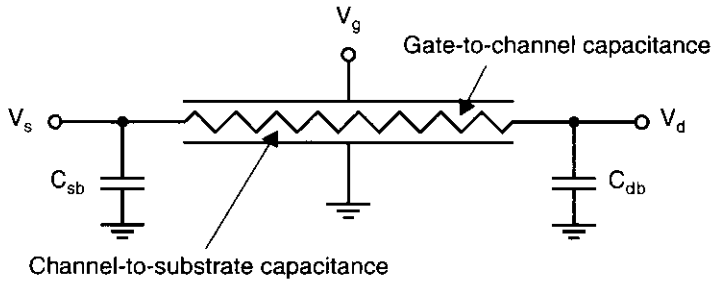


Fig. 1.21 A distributed RC model for a transistor in the active region.

and the channel-to-substrate capacitance are modelled as distributed elements. However, the I-V relationships of the distributed RC elements are highly nonlinear because the junction capacitances of the source and drain are nonlinear depletion capacitances, as is the channel-to-substrate capacitance. Also, if V_{DS} is not small, then the channel resistance per unit length should increase as one moves closer to the drain. This model is much too complicated for use in hand analysis.

A simplified model often used for small V_{DS} is shown in Fig. 1.22, where the resistance, r_{ds}, is given by (1.108). Here, the gate-to-channel capacitance has been evenly divided between the source and drain nodes,

$$C_{gs} = C_{gd} = \frac{A_{ch}C_{ox}}{2} = \frac{WLC_{ox}}{2} \tag{1.109}$$

Note that this equation ignores the gate-to-junction overlap capacitances, as given by (1.94), which should be taken into account when accuracy is very important. The channel-to-substrate capacitance has also been divided in half and shared between the source and drain junctions. Each of these capacitors should be added to the junction-to-substrate capacitance and the junction-sidewall capacitance at the appropriate

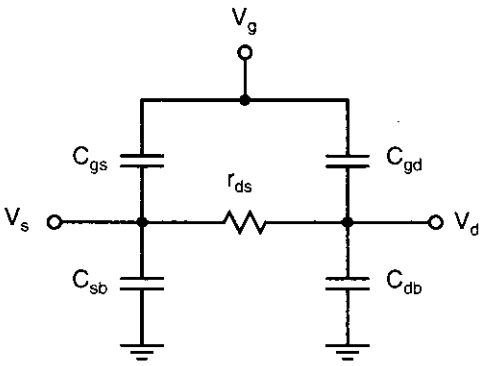


Fig. 1.22 A simplified triode-region model valid for small V_{DS}.

node. Thus, we have

$$C_{sb\text{-}0} = C_{j0}\left(A_s + \frac{A_{ch}}{2}\right) + C_{j\text{-}sw0}P_s \tag{1.110}$$

and

$$C_{db\text{-}0} = C_{j0}\left(A_d + \frac{A_{ch}}{2}\right) + C_{j\text{-}sw0}P_d \tag{1.111}$$

Also,

$$C_{sb} = \frac{C_{sb\text{-}0}}{\sqrt{1 + \frac{V_{sb}}{\Phi_0}}} \tag{1.112}$$

and

$$C_{db} = \frac{C_{db\text{-}0}}{\sqrt{1 + \frac{V_{sb}}{\Phi_0}}} \tag{1.113}$$

It might be noted that C_{sb} is often comparable in size to C_{gs} due to its larger area and the sidewall capacitance.

When the transistor turns off, the model changes considerably. A reasonable model is shown in Fig. 1.23. Perhaps the biggest difference is that r_{ds} is now infinite. Another major difference is that C_{gs} and C_{gd} are now much smaller. Since the channel has disappeared, these capacitors are now due to only overlap and fringing capacitance. Thus, we have

$$C_{gs} = C_{gd} = WL_{ov}C_{ox} \tag{1.114}$$

However, the reduction of C_{gs} and C_{gd} does not mean that the total gate capacitance is necessarily smaller. We now have a "new" capacitor, C_{gb}, which is the gate-

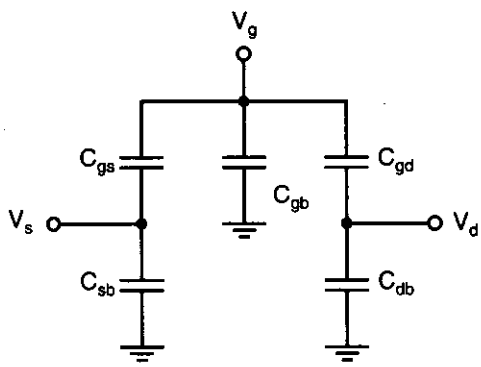


Fig. 1.23 A small-signal model for a MOSFET that is turned off.

to-substrate capacitance. This capacitor is highly nonlinear and dependent on the gate voltage. If the gate voltage has been very negative for some time and the gate is accumulated, then we have

$$C_{gb} = A_{ch}C_{ox} = WLC_{ox} \tag{1.115}$$

If the gate-to-source voltage is around 0 V, then C_{gb} is equal to C_{ox} in series with the channel-to-bulk depletion capacitance and is considerably smaller, especially when the substrate is lightly doped. Another case where C_{gb} is small is just after a transistor has been turned off, before the channel has had time to accumulate. Because of the complicated nature of correctly modelling C_{gb} when the transistor is turned off, equation (1.115) is usually used for hand analysis as a worst-case estimate.

The capacitors C_{sb} and C_{db} are also smaller when the channel is not present. We now have

$$C_{sb\text{-}0} = A_sC_{j0} \tag{1.116}$$

and

$$C_{db\text{-}0} = A_dC_{j0} \tag{1.117}$$

1.3 ADVANCED MOS MODELLING

In this section, we look at three advanced modelling concepts that a microcircuit designer is likely to encounter—short-channel effects, subthreshold operation, and leakage currents.

Short-Channel Effects

A number of short-channel effects degrade the operation of MOS transistors as device dimensions are scaled down. These effects include mobility degradation, reduced output impedance, and hot-carrier effects (such as oxide trapping and substrate currents). These short-channel effects will be briefly described here. For more detailed modelling of short-channel effects, see [Wolf, 1995].

Transistors that have short channel lengths and large electric fields experience a degradation in the effective mobility of their carriers due to several factors. One of these factors is the large lateral electric field (which has a vector in a direction perpendicular from the gate into the silicon) caused by large gate voltages and short channel lengths. This large lateral field causes the effective channel depth to change and also causes more electron collisions, thereby lowering the effective mobility. Another factor causing this degradation is that, due to large electric fields, carrier velocity begins to saturate. A first-order approximation that models this carrier-velocity saturation for electrons is given by

$$\upsilon_d \cong \frac{\mu_n E}{1 + E/E_c} \tag{1.118}$$

where E is the electric field and E_c is the critical electrical field, which might be on the order of 1.5×10^6 V/m. Using this equation in the derivation of the I_D-V_{eff}

characteristics of a MOS transistor, it can be shown [Gray, 1993] that the drain current is now given by

$$I_D = \frac{\mu_n C_{ox}}{2[1 + \theta V_{eff}]} \frac{W}{L} V_{eff}^2 \tag{1.119}$$

where $\theta = 1/(LE_c)$ and, for a 0.8-μm technology, might have a typical value of 0.6 V^{-1}. It can be shown that this mobility degradation is equivalent to a finite series source resistance given by

$$R_{SX} = \frac{1}{E_C} \frac{1}{\mu_n C_{ox}} \frac{1}{W} \tag{1.120}$$

For $\mu_n C_{ox} = 90\ \mu A/V^2$, this resistance might be on the order of 6 kΩ per μm of width (again, for a 0.8-μm-long transistor). This equivalent series source resistance is typically larger than the physical source resistance. This saturation causes the square-law characteristic of the current-voltage relationship to be inaccurate, and the true relationship will be somewhere between linear and square. In many voltage-to-current conversion circuits that rely on the square-law characteristic, this inaccuracy can be a major source of error. Taking channel lengths larger than the minimum allowed helps to minimize this degradation.

Transistors with short channel lengths also experience a reduced output impedance because depletion region variations at the drain end (which affect the effective channel length) have an increased proportional effect on the drain current. In addition, a phenomenon known as drain-induced barrier lowering (DIBL) effectively lowers V_t as V_{DS} is increased, thereby further lowering the output impedance of a short-channel device. This lower output impedance is the main reason that cascode current mirrors are becoming increasingly popular.

Another important short-channel effect is due to *hot carriers*. These high-velocity carriers can cause harmful effects, such as the generation of electron-hole pairs by impact ionization and avalanching. These extra electron-hole pairs can cause currents to flow from the drain to the substrate, as shown in Fig. 1.24. This effect can be mod-

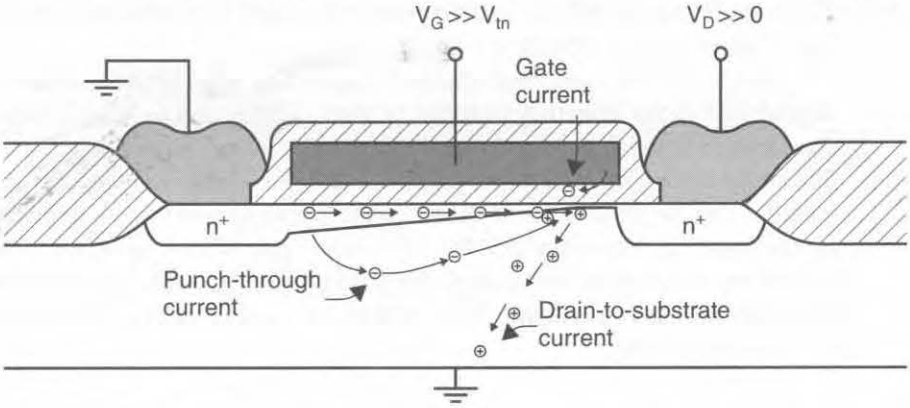


Fig. 1.24 Drain-to-substrate current caused by electron-hole pairs generated by impact ionization at drain end of channel.

elled by a finite drain-to-ground impedance. As a result, this effect is one of the major limitations on achieving very high output impedances of cascode current sources. In addition, this current flow can cause voltage drops across the substrate and possibly cause latch-up, as the next section describes.

Another hot-carrier effect occurs when electrons gain energies high enough so they can tunnel into and possibly through the thin gate oxide. Thus, this effect can cause dc gate currents. However, often more harmful is the fact that any charge trapped in the oxide will cause a shift in transistor threshold voltage. As a result, hot carriers are one of the major factors limiting the long-term reliability of MOS transistors.

A third hot-carrier effect occurs when electrons with enough energy *punch through* from the source to the drain. As a result, these high-energy electrons are no longer limited by the drift equations governing normal conduction along the channel. This mechanism is somewhat similar to punch-through in a bipolar transistor, where the collector depletion region extends right through the base region to the emitter. In a MOS transistor, the channel length becomes effectively zero, resulting in unlimited current flow (except for the series source and drain impedances, as well as external circuitry). This effect is an additional cause of lower output impedance and possibly transistor breakdown.

It should be noted that all of the hot-carrier effects just described are more pronounced for n-channel transistors than for their p-channel counterparts because electrons have larger velocities than holes.

Finally, it should be noted that short-channel transistors have much larger subthreshold currents than long-channel devices.

Subthreshold Operation

The device equations presented for MOS transistors in the preceding sections are all based on the assumption that V_{eff} (i.e., $V_{GS} - V_t$) is greater than about 100 mV and the device is in strong inversion. When this is not the case, the accuracy of the square-law equations is poor. If $V_{eff} < -100$ mV, the transistor is in weak inversion and is said to be operating in the *subthreshold region.* In this region, the transistor is more accurately modelled by an exponential relationship between its control voltage and current, somewhat similar to a bipolar transistor. In the subthreshold region, the drain current is approximately given by the exponential relationship [Geiger, 1990]

$$I_D \cong I_{D0}\left(\frac{W}{L}\right)e^{(qV_{GS}/nkT)} \tag{1.121}$$

where

$$n = \frac{C_{ox} + C_{depl}}{C_{ox}} \cong 1.5 \tag{1.122}$$

and it has been assumed that $V_S = 0$ and $V_{DS} > 75$ mV. The constant I_{D0} might be around 20 nA.

Although the transistors have an exponential relationship in this region, the transconductances are still small because of the small bias currents, and the transistors are slow because of small currents for charging and discharging capacitors. In addition, matching between transistors suffers because it now strongly depends on transistor-threshold-voltage matching. Normally, transistors are not operated in the subthreshold region, except in very low-frequency and low-power applications.

Leakage Currents

An important second-order device limitation in some applications is the leakage current of the junctions. For example, this leakage can be important in estimating the maximum time a sample-and-hold circuit or a dynamic memory cell can be left in hold mode. The leakage current of a reverse-biased junction (not close to breakdown) is approximately given by

$$I_{lk} \cong \frac{qA_j n_i}{2\tau_0} x_d \tag{1.123}$$

where A_j is the junction area, n_i is the intrinsic concentration of carriers in undoped silicon, τ_0 is the effective minority carrier lifetime, and x_d is the thickness of the depletion region. τ_0 is given by

$$\tau_0 \cong \frac{1}{2}(\tau_n + \tau_p) \tag{1.124}$$

where τ_n and τ_p are the electron and hole lifetimes. Also, x_d is given by

$$x_d = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}(\Phi_0 + V_r)} \tag{1.125}$$

and n_i is given by

$$n_i \cong \sqrt{N_C N_V}\ e^{(-E_g)/(kT)} \tag{1.126}$$

where N_C and N_V are the densities of states in the conduction and valence bands and E_g is the difference in energy between the two bands.

Since the intrinsic concentration, n_i, is a strong function of temperature (it approximately doubles for every temperature increase of 11 °C for silicon), the leakage current is also a strong function of temperature. Roughly speaking, the leakage current also doubles for every 11 °C rise in temperature. Thus, the leakage current at higher temperatures is much larger than at room temperature. This leakage current imposes a maximum time on how long a dynamically charged signal can be maintained in a high impedance state.

1.4 BIPOLAR-JUNCTION TRANSISTORS

In the early electronic years, the majority of microcircuits were realized using bipolar-junction transistors (BJTs). However, in the late 1970s, microcircuits that used MOS

transistors began to dominate the industry, with BJT microcircuits remaining popular for high-speed applications. More recently, bipolar CMOS (BiCMOS) technologies, where both bipolar and MOS transistors are realized in the same microcircuit, have grown in popularity. BiCMOS technologies are particularly attractive for mixed analog-digital applications, and thus it is important for an analog designer to become familiar with bipolar devices.

Modern bipolar transistors can have unity-gain frequencies as high as 15 to 45 GHz or more, compared to unity-gain frequencies of only 1 to 4 GHz for MOS transistors that use a technology with similar lithography resolution. Unfortunately, in bipolar transistors, the *base* control terminal has a nonzero input current when the transistor is conducting current (from the collector to the emitter for an npn transistor; from the emitter to the collector for a pnp transistor). Fortunately, at low frequencies, the base current is much smaller than the collector-to-emitter current—it may be only 1/100 of the collector current for an npn transistor. For lateral pnp transistors, the base current may be as large as 1/20 of the emitter-to-collector current.

A typical cross section of an npn bipolar-junction transistor is shown in Fig. 1.25. Although this structure looks quite complicated, it corresponds approximately to the equivalent structure shown in Fig. 1.26. In a good BJT transistor, the width of the base

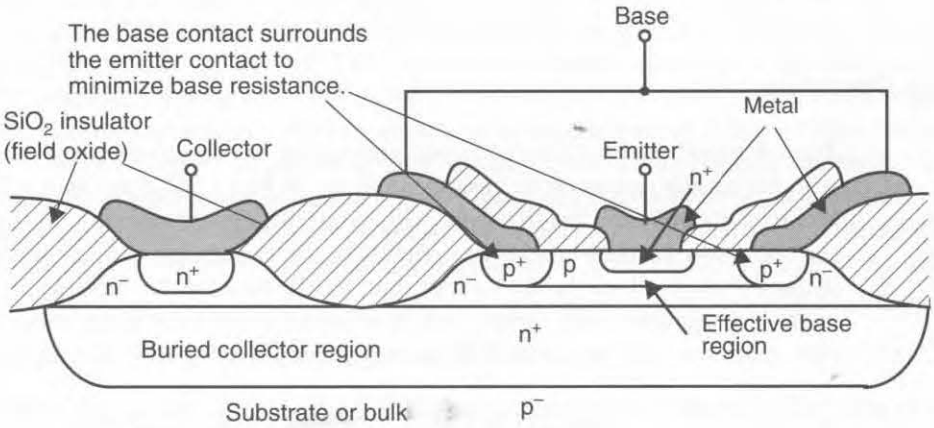


Fig. 1.25 A cross section of an npn bipolar-junction transistor.

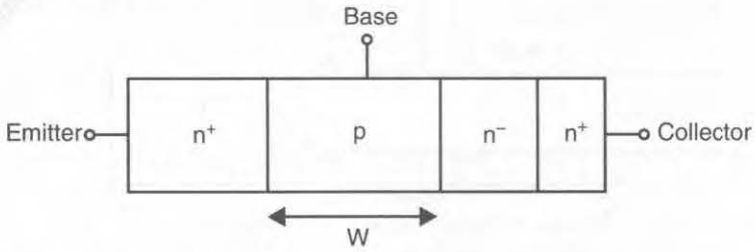


Fig. 1.26 A simplified structure of an npn transistor.

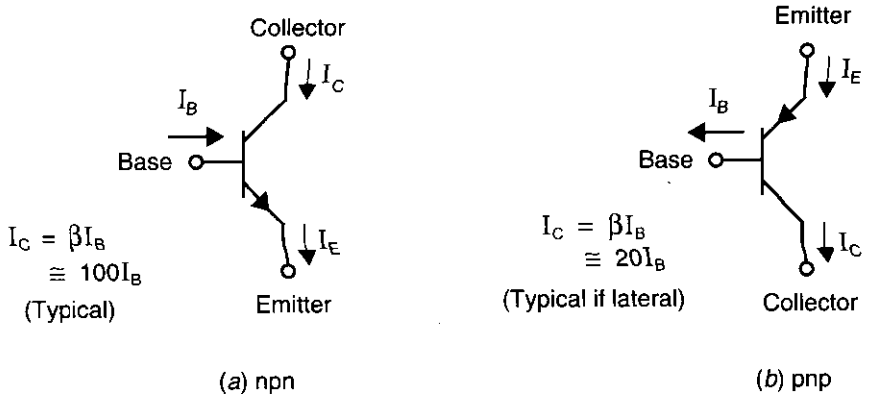


Fig. 1.27 The symbols representing (*a*) an npn bipolar-junction transistor and (*b*) a pnp bipolar-junction transistor.

region, W, is small (typically, less than 1 μm). Also, as we will see, the base must be more lightly doped than the emitter.

The circuit symbols used to represent npn and pnp transistors are shown in Fig. 1.27.

Basic Operation

To understand the operation of bipolar transistors, we consider here an npn transistor with the emitter connected to ground, as shown in Fig. 1.28. If the base voltage, V_B, is less than about 0.5 V, the transistor will be cut off, and no current will flow. We will see that when the base-emitter pn junction becomes forward biased, current will start to flow from the base to the emitter, but, partly because the base width is small, a much larger proportional current will flow from the collector to the emitter. Thus, the npn transistor can be considered a current amplifier at low frequencies. In other

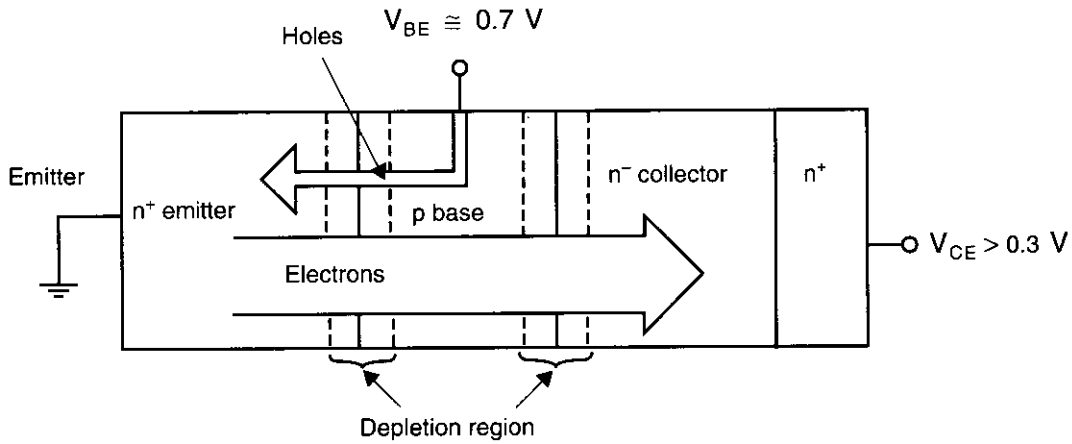


Fig. 1.28 Various components of the currents of an npn transistor.

words, if the transistor is not cut off and the collector-base junction is reverse biased, a small base current controls a much larger collector-emitter current.

A simplified overview of how an npn transistor operates follows: When the base-emitter junction becomes forward biased, it starts to conduct, similar to any forward-biased junction. The current consists of majority carriers from the base (in this case, holes) and majority carriers from the emitter (in this case, electrons) diffusing across the junction. Because the emitter is more heavily doped than the base, there are many more electrons injected from the emitter than there are holes injected from the base. Assuming the collector voltage is large enough so that the collector-base junction is reverse biased, no holes from the base will go to the collector. However, the electrons that travel from the emitter to the base, where they are now minority carriers, diffuse away from the base-emitter junction because of the minority-carrier concentration gradient in the base region. Any of these minority electrons that get close to the collector-base junction will immediately be "whisked" across the junction due to the large positive voltage on the collector, which attracts the negatively charged electrons. In a properly designed bipolar transistor, such as that shown in Fig. 1.25, the vertical base width, W, is small, and almost all of the electrons that diffuse from the emitter to the base reach the collector-base junction and are swept across the junction, thus contributing to current flow in the collector. The result is that the collector current very closely equals the electron current flowing from the emitter to the base. The much smaller base current very closely equals the current due to the holes that flow from the base to the emitter. The total emitter current is the sum of the electron collector current and the hole base current, but since the hole current is much smaller than the electron current, the emitter current is approximately equal to the collector current.

Since the collector current is approximately equal to the electron current flowing from the emitter to the base, and the amount of this electron current is determined by the base-emitter voltage, it can be shown (see Appendix at the end of this chapter) that the collector current is exponentially related to the base-emitter voltage by the relationship

$$I_C \cong I_{CS} e^{V_{BE}/V_T} \tag{1.127}$$

where I_{CS} is the *scale current*. This scale current is proportional to the area of the base-emitter junction. The base current, determined by the hole current flowing from the base to the emitter, is also exponentially related to the base-emitter voltage, resulting in the ratio of the collector current to the base current being a constant that, to a first-order approximation, is independent of voltage and current. This ratio, typically denoted as β, is defined to be

$$\beta \equiv \frac{I_C}{I_B} \tag{1.128}$$

where I_C and I_B are the collector and base currents. Typical values of β are between 50 and 200.

Note that (1.127) implies that the collector current is independent of the collector voltage. This independence ignores second-order effects such as the decrease in effective base width, W, due to the increase in the width of the collector-base depletion

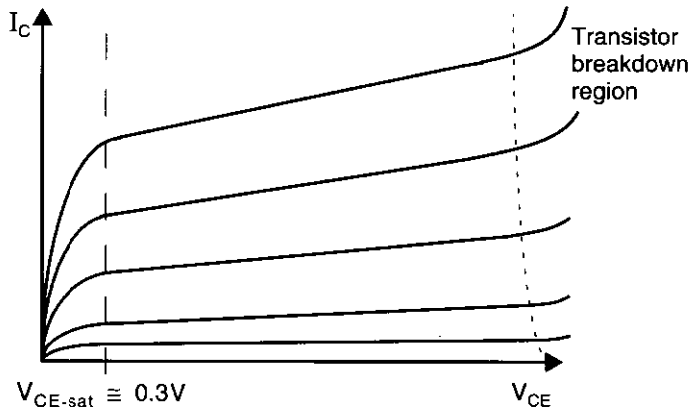


Fig. 1.29 Typical plot of I_C versus V_{CE} for a BJT.

region when the collector bias voltage is increased. To illustrate this point, a typical plot of the collector current, I_C, as a function of collector-to-emitter voltage, V_{CE}, for different values of I_B is shown in Fig. 1.29 for a practical transistor. The fact that the curves are not flat for $V_{CE} > V_{CE\text{-}sat}$ indicates the dependence of I_C on V_{CE}. Indeed, to a good approximation, the dependence is linear with a slope that intercepts the V_{CE} axis at $V_{CE} = -V_A$ for all values of I_B. The intercept voltage value, V_A, is called the *Early voltage* for bipolar transistors, with a typical value being from 50 V to 100 V. This dependency results in a finite output impedance (as in a MOS transistor) and can be modelled by modifying equation (1.127) [Sze, 1981] to be

$$I_C \cong I_{CS} e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right) \tag{1.129}$$

Large-Signal Modelling

A conducting BJT that has a V_{CE} greater than $V_{CE\text{-}sat}$ (which is approximately 0.3 V) is said to be operating in the active region. Such a collector-emitter voltage is required to ensure that none of the holes from the base go to the collector. A large-signal model of a BJT operating in the active region is shown in Fig. 1.30.

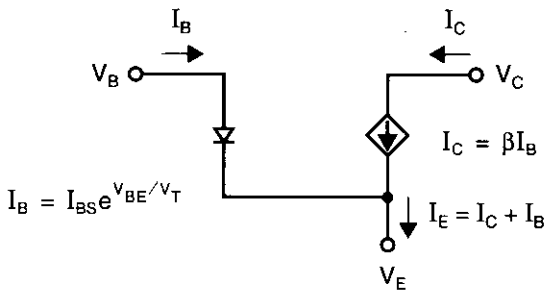


Fig. 1.30 A large-signal model for a BJT in the active region.

Since $I_B = I_C/\beta$, we have

$$I_B = \frac{I_{CS}}{\beta} e^{V_{BE}/V_T} = I_{BS} e^{V_{BE}/V_T} \tag{1.130}$$

which is similar to a diode equation, but with a multiplying constant of $I_{CS}/\beta = I_{BS}$. Since $I_E = I_B + I_C$, we have

$$I_E = I_{CS}\left(\frac{\beta+1}{\beta}\right) e^{V_{BE}/V_T} = I_{ES} e^{V_{BE}/V_T} \tag{1.131}$$

or equivalently

$$I_C = \alpha I_E \tag{1.132}$$

where α has been defined as

$$\alpha = \frac{\beta}{\beta+1} \tag{1.133}$$

and for large values of β, can be approximated as

$$\alpha \cong 1 - \frac{1}{\beta} \cong 1 \tag{1.134}$$

If the effect of V_{CE} on I_C is included in the model, the current-controlled source, βI_B, should be replaced by a current source given by

$$I_C = \beta I_B\left(1 + \frac{V_{CE}}{V_A}\right) \tag{1.135}$$

where V_A is the Early-voltage constant. This additional modelling of the finite output impedance is normally not done in large-signal analysis without the use of a computer due to its complexity.

As the collector-emitter voltage approaches $V_{CE\text{-}sat}$ (typically around 0.2 to 0.3 V), the base-collector junction becomes forward biased, and holes from the base will begin to diffuse to the collector. A common model for this case, when the transistor is *saturated* or in the *saturation region,* is shown in Fig. 1.31. It should be noted that the value of $V_{CE\text{-}sat}$ decreases for smaller values of collector current.

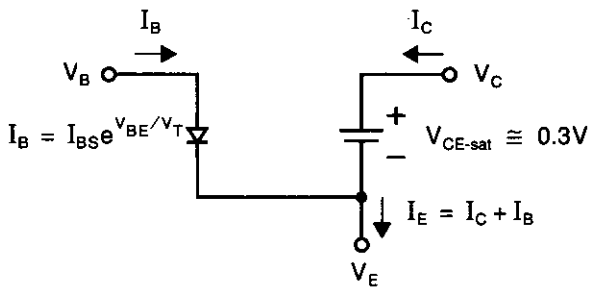


Fig. 1.31 A large-signal model for a BJT in the saturation region.

Base-Charge Storage in the Active Region

When a transistor is in the active region, many minority carriers are stored in the base region (electrons are stored in an npn transistor). Recall that this minority charge is responsible for I_C, so this charge must be removed (through the base contact) before a transistor can turn off. As in a forward-bias diode, this charge can be modelled as a diffusion capacitance, C_d, between the base and emitter given by (see Appendix at the end of this chapter)

$$C_d = \tau_b \frac{I_C}{V_T} \tag{1.136}$$

where τ_b is the base-transit-time constant. Thus, we see that the diffusion capacitance is proportional to I_C. The total base-emitter capacitance, C_{be}, will include the base-emitter depletion capacitance, C_j, in parallel with C_d. Normally, however, C_j is much less than C_d, unless the transistor current is small, and can often be ignored.

Base-Charge Storage of a Saturated Transistor

When a transistor becomes saturated, the minority-charge storage in the base and, even more so, in the lightly doped region of the collector, increases drastically. The major component of this charge storage is due to holes diffusing from the base, through the collector junction, and continuing on through the lightly doped n^- *epitaxial region* of the collector to the n^+ collector region. The n^- epitaxial region is so named because it is epitaxially grown on a p region. Most of the charge storage occurs in this region. Also, additional charge storage occurs because electrons that diffused from the collector are stored in the base, but this charge is normally smaller. The magnitude of the additional charge stored by a transistor that is saturated is given by

$$Q_s = \tau_s\left(I_B - \frac{I_C}{\beta}\right) \tag{1.137}$$

where the base overdrive current, defined to be $I_B - I_C/\beta$, is approximately equal to the hole current from the base to the collector. Normally, in saturation, $I_B >> I_C/\beta$, and (1.137) can be approximated by

$$Q_s \cong \tau_s I_B \tag{1.138}$$

The constant τ_s is approximately equal to the epitaxial-region transit time, τ_E (ignoring the storage of electrons in the base that diffused from the collector). Since the epitaxial region is much wider than the base, the constant τ_s is normally much larger than the base transit time, the constant τ_b, often by up to two orders of magnitude. The specific value of τ_s is usually found empirically for a given technology.

When a saturated transistor is being turned off, first the base current will reverse. *However, before the collector current will change, the saturation charge, Q_s, must be removed.* After Q_s is removed, the base minority charge, Q_b, will be removed. Dur-

ing this time, the collector current will decrease until the transistor shuts off. Typically, the time to remove Q_s greatly dominates the overall charge removal.

If the time required to remove the base saturation charge, t_s, is much shorter than the epitaxial-region transit time, τ_E, then one can derive a simple expression for the time required to remove the saturation charge. If the reverse base current (when the saturation charge is being removed), denoted by I_{BR}, remains constant while Q_s is being removed, then we have [Hodges, 1988]

$$t_s \cong \frac{Q_s}{I_{BR}} \cong \frac{\tau_s\left(I_B - \frac{I_C}{\beta}\right)}{I_{BR}} \cong \tau_s \frac{I_B}{I_{BR}} \tag{1.139}$$

where $\tau_s \cong \tau_E$.

Normally, the forward base current during saturation, I_B, will be much smaller than the reverse base current during saturation-charge removal, I_{BR}. If this were not the case, then our original assumption that $t_s << \tau_E \cong \tau_s$ would not be true. In this case, the turn-off time of the BJT would be so slow as to make the circuit unusable in most applications. Nevertheless, the turn-off time for this case, when t_s is not much less than τ_E, is given by [Hodges, 1988]

$$t_s = \tau_s \ln\left[\frac{I_{BR} + I_B}{I_{BR} + \frac{I_C}{\beta}}\right] \tag{1.140}$$

The reader should verify that for $I_{BR} >> I_B$ and $I_{BR} >> I_C/\beta$, the expression in (1.140) is approximately equivalent to the much simpler one in (1.139).

In both of the cases just described, the time required to remove the storage charge of a saturated transistor is much larger than the time required to turn off a transistor in the active region. *In high-speed microcircuit designs, one never allows bipolar transistors to saturate, to avoid the long turn-off time that would result.*

EXAMPLE 1.13

For $\tau_b = 0.2$ ns, $\tau_s = 100$ ns (a small value for τ_s), $I_B = 0.2$ mA, $I_C = 1$ mA, $\beta = 100$, and $I_{BR} = 1$ mA, calculate the time required to remove the base saturation charge using (1.139), and compare it to the time obtained using the more accurate expression of (1.140). Compare these results to the time required to remove the base minority charge for the same I_{BR}.

Solution

Using (1.139), we have

$$t_s = \frac{10^{-7}(2 \times 10^{-4})}{10^{-3}} = 20 \text{ ns} \tag{1.141}$$

Using (1.140), we have

$$t_s = 10^{-7}\ln\left[\frac{10^{-3} + 2\times 10^{-4}}{10^{-3} + \dfrac{10^{-3}}{100}}\right] = 17.2 \text{ ns} \tag{1.142}$$

which is fairly close to the first result.

The time required for an active transistor to remove the base minority charge, Q_b, is given by

$$t_A = \frac{Q_b}{I_{BR}} = \frac{\tau_b I_C}{I_{BR}} = 0.2 \text{ ns} \tag{1.143}$$

This is approximately 100 times shorter than the time for removing the base saturation charge!

Small-Signal Modelling

The most commonly used small-signal model is the hybrid-π model. This model is similar to the small-signal model used for MOS transistors, except it includes a finite base-emitter impedance, r_π, and it has no emitter-to-bulk capacitance. The hybrid-π model is shown in Fig. 1.32. As in the MOS case, we will first discuss the transconductance, g_m, and the small-signal resistances, and then we will discuss the parasitic capacitances.

The transistor transconductance, g_m, is perhaps the most important parameter of the small-signal model. The transconductance is the ratio of the small-signal collector current, i_c, to the small-signal base-emitter voltage, v_{be}. Thus, we have

$$g_m = \frac{i_c}{v_{be}} = \frac{\partial I_C}{\partial V_{BE}} \tag{1.144}$$

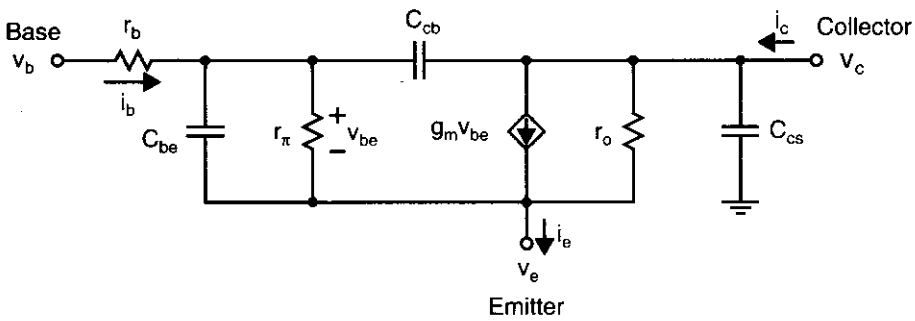


Fig. 1.32 The small-signal model of an active BJT.

Recall that in the active region

$$I_C = I_{CS}e^{V_{BE}/V_T} \tag{1.145}$$

Then

$$g_m = \frac{\partial I_C}{\partial V_{BE}} = \frac{I_{CS}}{V_T}e^{V_{BE}/V_T} \tag{1.146}$$

Using (1.145) again, we obtain

$$g_m = \frac{I_C}{V_T} \tag{1.147}$$

where V_T is given by

$$V_T = \frac{kT}{q} \tag{1.148}$$

and is approximately 26 mV at a room temperature of $T = 300\ °K$. Thus, the transconductance is proportional to the bias current of a BJT. In integrated-circuit design, it is important that the transconductance (and hence speed) remain temperature independent, so the bias currents are usually made proportional to absolute temperature (since V_T is proportional to absolute temperature).

The presence of the resistor r_π reflects the fact that the base current is nonzero. We have

$$r_\pi = \frac{\partial V_{BE}}{\partial I_B} \tag{1.149}$$

Because from (1.130) we have

$$I_B = \frac{I_C}{\beta} = \frac{I_{CS}}{\beta}e^{V_{BE}/V_T} \tag{1.150}$$

we therefore have

$$\frac{1}{r_\pi} = \frac{\partial I_B}{\partial V_{BE}} = \frac{I_{CS}}{\beta V_T}e^{V_{BE}/V_T} \tag{1.151}$$

Using (1.150) again, we have

$$r_\pi = \frac{V_T}{I_B} \tag{1.152}$$

or equivalently,

$$r_\pi = \beta\frac{V_T}{I_C} = \frac{\beta}{g_m} \tag{1.153}$$

Since

$$i_e = i_c + i_b \tag{1.154}$$

we also have

$$\begin{aligned}\frac{\partial I_E}{\partial V_{BE}} &= \frac{\partial I_c}{\partial V_{BE}} + \frac{\partial I_B}{\partial V_{BE}} \\ &= g_m + \frac{g_m}{\beta} \\ &= g_m\left(\frac{1+\beta}{\beta}\right) \\ &= \frac{g_m}{\alpha}\end{aligned} \tag{1.155}$$

Some alternative models, usually called T models (see page 55), use the emitter resistance, r_e, where

$$r_e = \frac{\partial V_{BE}}{\partial I_E} = \frac{\alpha}{g_m} \tag{1.156}$$

Continuing, we have

$$\frac{1}{r_o} = \frac{\partial I_C}{\partial V_{CE}} \tag{1.157}$$

The small-signal resistance, r_o, models the dependence of the collector current on the collector-emitter voltage. Repeating (1.129) here for convenience,

$$I_C = I_{CS}e^{V_{BE}/V_T}\left(1 + \frac{V_{CE}}{V_A}\right) \tag{1.158}$$

we have

$$\frac{1}{r_o} = \frac{\partial I_C}{\partial V_{CE}} = \frac{I_{CS}}{V_A}e^{V_{BE}/V_T} \tag{1.159}$$

Thus,

$$r_o = \frac{V_A}{I_C} \tag{1.160}$$

which is inversely proportional to the collector current. As an aside, note that $g_m r_o = V_A/V_T$ is a constant value independent of the transistor operating point. This constant is usually between 2,000 and 8,000 for an npn BJT and is an upper limit on the attainable voltage gain for a single-transistor amplifier.

The resistor r_b models the resistance of the semiconductor material between the base contact and the effective base region due to the moderately lightly doped base p material (see Fig. 1.25). This resistor, although small (typically a few hundred ohms), can be important in limiting the speed of very-high-frequency low-gain BJT circuits and is a major source of noise.

EXAMPLE 1.14

For $I_C = 1$ mA, $\beta = 100$, and $V_A = 100$ V, calculate g_m, r_π, r_e, r_o, and $g_m r_o$.

Solution

We have

$$g_m = \frac{I_C}{V_T} = \frac{10 \times 10^{-3}\ \text{A}}{0.026\ \text{V}} = 38.5\ \text{mA/V} \tag{1.161}$$

$$r_\pi = \frac{\beta}{g_m} = 2.6\ \text{k}\Omega \tag{1.162}$$

$$r_e = \frac{\alpha}{g_m} = \left(\frac{100}{101}\right)26 = 25.7\ \Omega \tag{1.163}$$

$$r_o = \frac{V_A}{I_C} = \frac{100}{10^{-3}} = 100\ \text{k}\Omega \tag{1.164}$$

and $g_m r_o$, the maximum possible gain with a single-transistor amplifier, is given by

$$g_m r_o = \frac{V_A}{V_T} = 3{,}846 \tag{1.165}$$

Note that this gain is much higher than the 52.6 that was found for a single MOS transistor in Example 1.9. Also note that this BJT maximum gain is independent of the bias current. For MOS transistors, it can be shown that the maximum gain decreases with larger bias currents in a square-root relationship. This is one of the reasons why it is possible to realize a single-transistor BJT amplifier with a much larger gain than would result if a MOS transistor were used, especially at high current levels (and therefore at high frequencies).

The high-frequency operation of a BJT is limited by the capacitances of the small-signal model. We have already encountered one of these capacitances, C_{be}, in Section 1.1. Recapping, we have

$$C_{be} = C_j + C_d \tag{1.166}$$

where C_j is the depletion capacitance of the base-emitter junction. For a forward-biased junction, a rough approximation for C_j is

$$C_j \cong 2A_E C_{je0} \tag{1.167}$$

The diffusion capacitance, C_d, is given in (1.136) as

$$C_d = \tau_b \frac{I_C}{V_T} = g_m \tau_b \tag{1.168}$$

The capacitor, C_{cb}, models the depletion capacitance of the collector-base junction. Since this is a graded junction, we can approximate C_{cb} by

$$C_{cb} = \frac{A_C C_{jc0}}{\left(1 + \frac{V_{CB}}{\Phi_{c0}}\right)^{1/3}} \tag{1.169}$$

where A_C is the effective area of the collector-base interface.

Due to the lower doping levels in the base and especially in the collector (perhaps 5×10^{22} acceptors/m^3 and 10^{21} donors/m^3, respectively), Φ_{c0}, the built-in potential for the collector-base junction, will be less than that of the base-emitter junction (perhaps 0.75 V as opposed to 0.9 V). It should be noted that the cross-sectional area of the collector-base junction, A_C, is typically much larger than the effective area of the base-emitter junction, A_E, which is shown in Fig. 1.25. This size differential results in $A_C\ C_{jc0}$ being larger than $A_E\ C_{je0}$, the base-emitter junction capacitance at 0 V bias, despite the lower doping levels.

Finally, another large capacitor is C_{cs}, the capacitance of the collector-to-substrate junction. Since this area is quite large, C_{cs}, which is the depletion capacitance that results from this area, will be much larger than either C_{cb} or the depletion capacitance component of C_{be}, that is, C_j. The value of C_{cs} can be calculated using

$$C_{cs} = \frac{A_T C_{js0}}{\left(1 + \frac{V_{CS}}{\Phi_{s0}}\right)^{1/2}} \tag{1.170}$$

where A_T is the effective transistor area and C_{js0} is the collector-to-substrate capacitance per unit area at 0-V bias voltage.

A common indicator for the *speed* of a BJT is the frequency at which the transistor's current gain drops to unity, when its collector is connected to a small-signal ground. This frequency is denoted f_t and is called the transistor *unity-gain frequency.* We can see how this frequency is related to the transistor model parameters by analyzing the small-signal circuit of Fig. 1.33. In the simplified model in part *b*, the resistor r_b is ignored because it has no effect on i_b since the circuit is being driven by a perfect current source. We have

$$V_{be} = i_b \left(r_\pi \left\| \frac{1}{sC_{be}} \right\| \frac{1}{sC_{cb}} \right) \tag{1.171}$$

and

$$i_c = g_m V_{be} \tag{1.172}$$

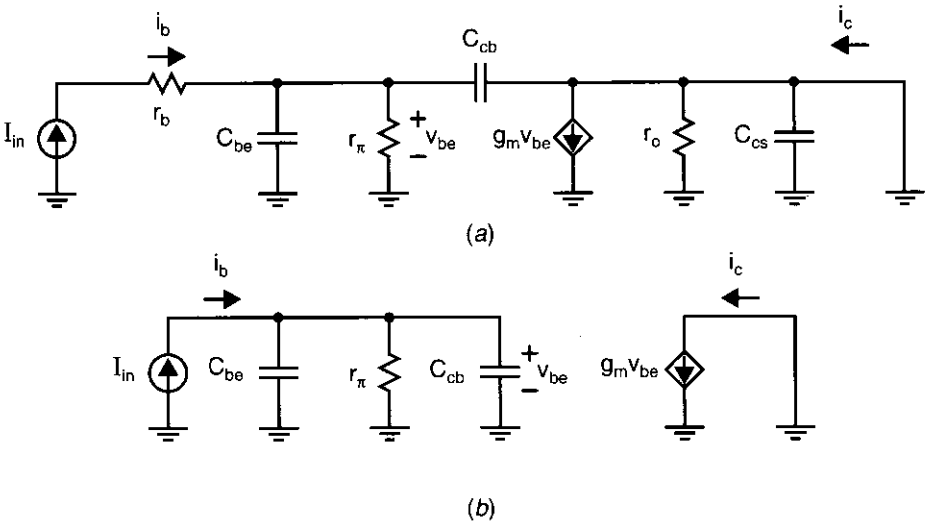


Fig. 1.33 (*a*) A small-signal model used to find f_t; (*b*) an equivalent simplified model.

Solving for i_c / i_b gives

$$\frac{i_c}{i_b} = \frac{g_m r_\pi}{1 + s(C_{be} + C_{cb})r_\pi} \tag{1.173}$$

At low frequencies, the current gain is $g_m r_\pi$, which equals the expected value of β (using (1.153)). At high frequencies, i_c / i_b is approximately given by

$$\left|\frac{i_c}{i_b}(\omega)\right| \cong \frac{g_m r_\pi}{\omega(C_{be} + C_{cb})r_\pi} = \frac{g_m}{\omega(C_{be} + C_{cb})} \tag{1.174}$$

To find the unity-gain frequency, we set $|(i_c/i_b)(\omega_t)| = 1$ and solve for ω_t, which results in

$$\omega_t = \frac{g_m}{C_{be} + C_{cb}} \tag{1.175}$$

or

$$f_t = \frac{g_m}{2\pi(C_{be} + C_{cb})} \tag{1.176}$$

Often, either f_t, ω_t, or $\tau_t = 1/\omega_t$ will be specified for a transistor at a particular bias current. These values indicate an upper limit on the maximum frequency at which the transistor can be effectively used.

The hybrid-π model is only one of a number of small-signal models that can be used. One common alternative is the low-frequency T model shown in Fig. 1.34. Use of this T model often results in a much simplified analysis, compared to use of the hybrid-π model, and thus it is useful for hand analysis.

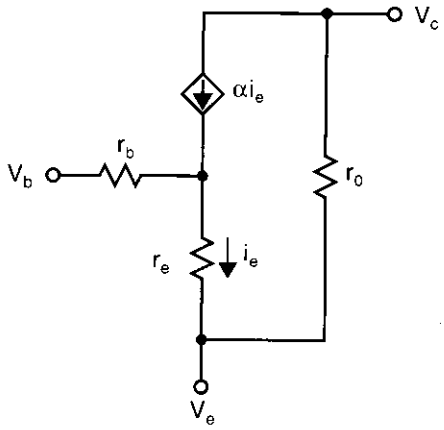


Fig. 1.34 A low-frequency, small-signal T model for an active BJT.

1.5 DEVICE MODEL SUMMARY

As a useful aid, all of the equations for the large-signal and small-signal modelling of diodes, MOS transistors, and bipolar transistors, along with values for the various constants, are listed in the next few pages.

Constants

$q = 1.602 \times 10^{-19}$ C	$k = 1.38 \times 10^{-23}$ JK^{-1}
$n_i = 1.1 \times 10^{16}$ carriers/m^3 at $T = 300$ °K	$\varepsilon_0 = 8.854 \times 10^{-12}$ F/m
$K_{ox} \cong 3.9$	$K_s \cong 11.8$
$\mu_n = 0.05$ m^2/V · s	$\mu_p = 0.02$ m^2/V · s

Diode Equations

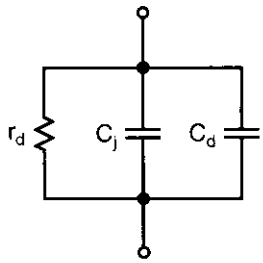
Reverse-Biased Diode (Abrupt Junction)

$C_j = \dfrac{C_{j0}}{\sqrt{1 + \dfrac{V_R}{\Phi_0}}}$	$Q = 2C_{j0}\Phi_0\sqrt{1 + \dfrac{V_R}{\Phi_0}}$
$C_{j0} = \sqrt{\dfrac{qK_s\varepsilon_0}{2\Phi_0}\dfrac{N_D N_A}{N_A + N_D}}$	$C_{j0} = \sqrt{\dfrac{qK_s\varepsilon_0 N_D}{2\Phi_0}}$ if $N_A >> N_D$
$\Phi_0 = \dfrac{kT}{q}\ln\left(\dfrac{N_A N_D}{n_i^2}\right)$	

Forward-Biased Diode

$I_D = I_S e^{V_D/V_T}$	$I_S = A_D q n_i^2 \left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right)$
$V_T = \frac{kT}{q} \cong 26 \text{ mV at } 300\,^\circ\text{K}$	

Small-Signal Model of Forward-Biased Diode



$r_d = \frac{V_T}{I_D}$	$C_T = C_d + C_j$
$C_d = \tau_T \frac{I_D}{V_T}$	$C_j \cong 2C_{j0}$
$\tau_T = \frac{L_n^2}{D_n}$	

MOS Transistor Equations

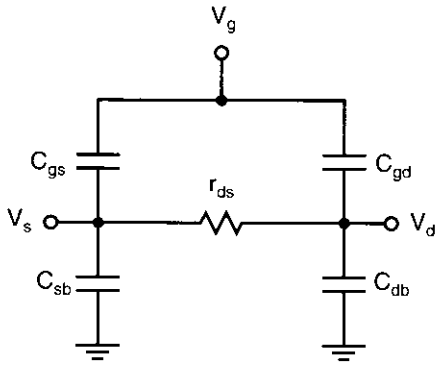
The following equations are for n-channel devices—for p-channel devices, put negative signs in front of all voltages. These equations do not account for short-channel effects (i.e., $L < 2L_{min}$).

Triode Region ($V_{GS} > V_{tn}$, $V_{DS} \le V_{eff}$)

$I_D = \mu_n C_{ox} \left(\frac{W}{L} \right) \left[(V_{GS} - V_{tn}) V_{DS} - \frac{V_{DS}^2}{2} \right]$	
$V_{eff} = V_{GS} - V_{tn}$	$V_{tn} = V_{tn\text{-}0} + \gamma(\sqrt{V_{SB} + 2\phi_F} - \sqrt{2\phi_F})$

$\phi_F = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right)$	$\gamma = \frac{\sqrt{2qK_{si}\varepsilon_0 N_A}}{C_{ox}}$
$C_{ox} = \frac{K_{ox}\varepsilon_0}{t_{ox}}$	

Small-Signal Model in Triode Region (for $V_{DS} << V_{eff}$)

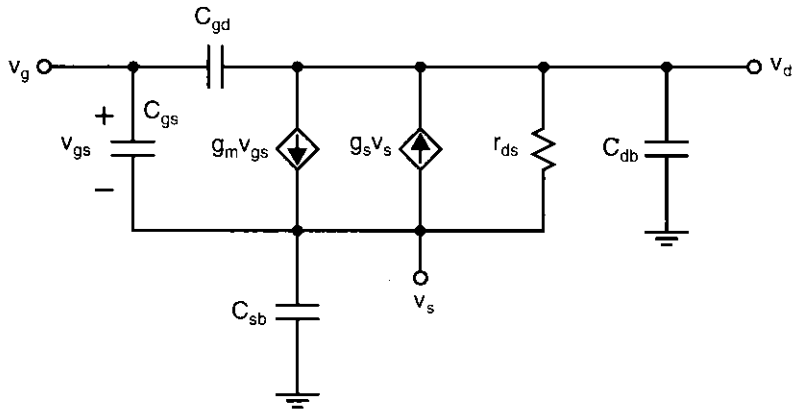


$r_{ds} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}}$	
$C_{gd} = C_{gs} \cong \frac{1}{2}WLC_{ox} + WL_{ov}C_{ox}$	$C_{sb} = C_{db} = \frac{C_{j0}(A_s + WL/2)}{\sqrt{1 + \frac{V_{sb}}{\Phi_0}}}$

Active (or Pinch-Off) Region ($V_{GS} > V_{tn}$, $V_{DS} \geq V_{eff}$)

$I_D = \frac{\mu_n C_{ox}}{2}\frac{W}{L}(V_{GS} - V_{tn})^2[1 + \lambda(V_{DS} - V_{eff})]$	
$\lambda \propto \frac{1}{L\sqrt{V_{DS} - V_{eff} + \Phi_0}}$	$V_{tn} = V_{tn\text{-}0} + \gamma(\sqrt{V_{SB} + 2\phi_F} - \sqrt{2\phi_F})$
$V_{eff} = V_{GS} - V_{tn} = \sqrt{\frac{2I_D}{\mu_n C_{ox} W/L}}$	

Small-Signal Model (Active Region)



$g_m = \mu_n C_{ox}\left(\frac{W}{L}\right)V_{eff}$	$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$
$g_m = \frac{2I_D}{V_{eff}}$	$g_s = \frac{\gamma g_m}{2\sqrt{V_{SB} + \lvert 2\phi_F \rvert}}$
$r_{ds} = \frac{1}{\lambda I_D}$	$g_s \cong 0.2 g_m$
$\lambda = \frac{k_{rds}}{2L\sqrt{V_{DS} - V_{eff} + \Phi_0}}$	$k_{rds} = \sqrt{\frac{2K_s\varepsilon_0}{qN_A}}$
$C_{gs} = \frac{2}{3}WLC_{ox} + WL_{ov}C_{ox}$	$C_{gd} = WL_{ov}C_{ox}$
$C_{sb} = (A_s + WL)C_{js} + P_s C_{j\text{-}sw}$	$C_{js} = \frac{C_{j0}}{\sqrt{1 + V_{SB}/\Phi_0}}$
$C_{db} = A_d C_{jd} + P_d C_{j\text{-}sw}$	$C_{jd} = \frac{C_{j0}}{\sqrt{1 + V_{DB}/\Phi_0}}$

Typical Values for a 0.8-μm Process

$V_{tn} = 0.8$ V	$V_{tp} = -0.9$ V
$\mu_n C_{ox} = 90\ \mu A/V^2$	$\mu_p C_{ox} = 30\ \mu A/V^2$

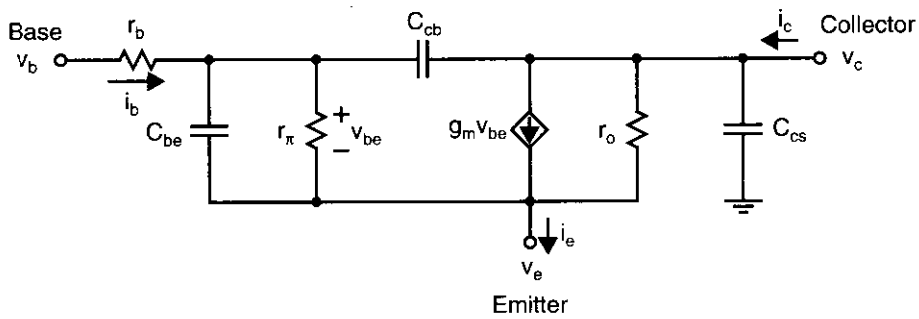
$C_{ox} = 1.9 \times 10^{-3}$ pF/(μm)^2	$C_j = 2.4 \times 10^{-4}$ pF/(μm)^2
$C_{j\text{-}sw} = 2.0 \times 10^{-4}$ pF/μm	$C_{gs(overlap)} = 2.0 \times 10^{-4}$ pF/μm
$\phi_F = 0.34$ V	$\Phi_0 = 0.9$ V
$\gamma = 0.5$ V$^{1/2}$	$t_{ox} = 0.02$ μm
$N_B = 6 \times 10^{21}$ impurities/m^3	

Bipolar-Junction Transistors

Active Transistor

$I_C = I_{CS} e^{V_{BE}/V_T}$	$V_T = \dfrac{kT}{q} \cong 26$ mV at 300 °K
For more accuracy, $I_C = I_{CS} e^{V_{BE}/V_T}\left(1 + \dfrac{V_{CE}}{V_A}\right)$	
$I_{CS} = \dfrac{A_E q D_n n_i^2}{W N_A}$	$I_B = \dfrac{I_C}{\beta}$
$I_E = \left(1 + \dfrac{1}{\beta}\right) I_C = \dfrac{1}{\alpha} I_C = (\beta + 1) I_B$	$\beta = \dfrac{I_C}{I_B} = \dfrac{D_n N_D L_p}{D_p N_A W} \cong 2.5 \dfrac{N_D L_p}{N_A W}$
$\alpha = \dfrac{\beta}{1 + \beta}$	

Small-Signal Model of an Active BJT



$g_m = \dfrac{I_C}{V_T}$	$r_\pi = \dfrac{V_T}{I_B} = \dfrac{\beta}{g_m}$
$r_e = \dfrac{\alpha}{g_m}$	$r_o = \dfrac{V_A}{I_C}$
$g_m r_o = \dfrac{V_A}{V_T}$	$C_d = \tau_b \dfrac{I_C}{V_T} = g_m \tau_b$
$C_{be} = C_j + C_d$	$C_{cs} = \dfrac{A_T C_{js0}}{\left(1 + \dfrac{V_{CS}}{\Phi_{s0}}\right)^{1/2}}$
$C_{cb} = \dfrac{A_C C_{jc0}}{\left(1 + \dfrac{V_{CB}}{\Phi_{c0}}\right)^{1/3}}$	

1.6 SPICE-MODELLING PARAMETERS

This section briefly describes some of the important model parameters for diodes, bipolar transistors, and MOS transistors used during a SPICE simulation. It should be noted here that not all SPICE model parameters are described. However, enough are described to enable the reader to understand the relationship between the relative parameters and the corresponding constants used when doing hand analysis.

Diode Model

There are a number of important dc parameters. The constant I_S is specified using either the parameter IS or JS in SPICE. These two parameters are synonyms, and only one should be specified. A typical value specified for I_S might be between 10^{-18} A and 10^{-15} A for small diodes in a microcircuit. Another important parameter is called the *emission coefficient,* n. This constant multiplies V_T in the exponential diode I-V relationship given by

$$I_D = I_S e^{V_{BE}/(nV_T)} \tag{1.177}$$

The SPICE parameter for n is N and is defaulted to 1 when not specified (1 is a reasonable value for junctions in a microcircuit). A third important dc characteristic is the *series resistance,* which is specified in SPICE using RS. It should be noted here that some SPICE programs allow the user to specify the area of the diode, whereas

others expect absolute parameters that already take into account the effective area. The manual for the program being used should be consulted.

The diode transit time is specified using the SPICE parameter TT. The most important capacitance parameter specified is CJ. CJO and CJ are synonyms—one should never specify both. This parameter specifies the capacitance at 0-V bias. Once again, it may be specified as absolute or as relative to the area (i.e., F/m^2), depending on the version of SPICE used. Also, the area junction grading coefficient, MJ, might be specified to determine the exponent used in the capacitance equation. Typical values are 0.5 for abrupt junctions and 0.33 for graded junctions. In some SPICE versions, it might also be possible to specify the sidewall capacitance at 0-V bias as well as its grading junction coefficient. Finally, the built-in potential of the junction, which is also used in calculating the capacitance, can be specified using PB. PHI, VJ, and PHA are all synonyms of PB.

Reasonably accurate diode simulations can usually be obtained by specifying only IS, CJ, MJ, and PB. However, most modern versions of SPICE have many more parameters that can be specified if one wants accurate temperature and noise simulations. Users should consult their manuals for more information.

Table 1.1 summarizes some of the more important diode parameters. This set of parameters constitutes a minimal set for reasonable simulation accuracy under ordinary conditions.

MOS Transistors

Modern MOS models are quite complicated, so only some of the more important MOS parameters used in SPICE simulations are described here. These parameters are used in what are called the Level 2 or Level 3 models. The model level can be chosen by setting the SPICE parameter LEVEL to either 2 or 3. The oxide thickness, t_{ox}, is specified using the SPICE parameter TOX. If it is specified, then it is not necessary to specify the thin gate-oxide capacitance (C_{ox}, specified by parameter COX). The mobility, μ_n, can be specified using UO. If UO is specified, the intrinsic transistor conductance ($\mu_n C_{ox}$) will be calculated automatically, unless this automatic calculation is overridden by specifying either KP (or its synonym, BETA). The transistor threshold voltage at $V_S = 0$ V, V_{tn}, is specified by VTO. The body-effect parameter, γ,

Table 1.1 Important SPICE parameters for modelling diodes

SPICE Parameter	Model Constant	Brief Description	Typical Value
IS	I_S	Transport saturation current	10^{-17} A
RS	R_d	Series resistance	30 Ω
TT	τ_T	Diode transit time	12 ps
CJ	C_{j0}	Capacitance at 0-V bias	0.01 pF
MJ	m_j	Diode grading coefficient exponent	0.5
PB	Φ_0	Built-in diode contact potential	0.9 V

can be specified using GAMMA, or it will be automatically calculated if the substrate doping, N_A, is specified using NSUB. Normally, one would not want SPICE to calculate γ because the effective substrate doping under the channel can differ significantly from the substrate doping in the bulk due to threshold-voltage adjust implants. The output impedance constant, λ, can be specified using LAMBDA. Normally, LAMBDA should not be specified since it takes precedence over internal calculations and does not change the output impedance as a function of different transistor lengths or bias voltages (which should be the case). Indeed, modelling the transistor output impedance is one of weakest points in SPICE. If LAMBDA is not specified, it is calculated automatically. The surface inversion potential, $|2\phi_F|$, can be specified using PHI, or it will be calculated automatically. Another parameter usually specified is the lateral diffusion of the junctions under the gate, L_D, which is specified by LD. For accurate simulations, one might also specify the resistances in series with the source and drain by specifying RS and RD (typically only the source resistance is important). Many other parameters exist to model such things as short-channel effects, subthreshold effects, and channel-width effects, but these parameters are outside the scope of this book.

The modelling of parasitic capacitances in SPICE is quite involved. Originally, this modelling was not very accurate since it did not include charge conservation for the gate charge. However, this modelling has greatly improved in recent commercial versions of SPICE. The capacitances under the junctions per unit area at 0-V bias, (i.e., C_{j0}) can be specified using CJ or can be calculated automatically from the specified substrate doping. The sidewall capacitances at 0 V, $C_{j\text{-}sw0}$, should normally be specified using CJSW because this parameter is used to calculate significant parasitic capacitances. The bulk grading coefficient specified by MJ can usually be defaulted to 0.5. Similarly, the sidewall grading coefficient specified by MJSW can usually be defaulted to 0.33 (SPICE assumes a graded junction). The built-in bulk-to-junction contact potential, Φ_0, can be specified using PB or defaulted to 0.8 V (note that 0.9 V would typically be more accurate, but the resulting simulation differences are small). Sometimes the gate-to-source or drain-overlap capacitances can be specified using CGSO or CGDO, but normally these would be left to be calculated automatically using COX and LD.

Some of the more important parameters that should result in reasonable simulations (except for modelling short-channel effects) are summarized in Table 1.2 for both n- and p-channel transistors. Table 1.2 lists reasonable parameters for a typical 0.8-μm technology.

Bipolar Junction Transistors

For historical reasons, most parameters for modelling bipolar transistors are specified absolutely. Also, rather than specifying the emitter area of a BJT in $(\mu m)^2$ on the line where the individual transistor connections are specified, most SPICE versions have multiplication factors. These multiplication factors can be used to automatically multiply parameters when a transistor is composed of several transistors connected in parallel. This multiplying parameter is normally called M.

Table 1.2 A reasonable set of MOS parameters for a typical 0.8-μm technology

SPICE Parameter	Model Constant	Brief Description	Typical Value
VTO	$V_{tn}:V_{tp}$	Transistor threshold voltage (in V)	0.7:–0.9
UO	$\mu_n:\mu_p$	Carrier mobility in bulk (in $cm^2/V\cdot s$)	500:175
TOX	t_{ox}	Thickness of gate oxide (in m)	1.8×10^{-8}
LD	L_D	Lateral diffusion of junction under gate (in m)	6×10^{-8}
GAMMA	γ	Body-effect parameter	0.5: 0.8
NSUB	$N_A:N_D$	The substrate doping (in cm^{-3})	$3 \times 10^{16}:7.5 \times 10^{16}$
PHI	$\|2\phi_F\|$	Surface inversion potential (in V)	0.7
PB	Φ_0	Built-in contact potential of junction to bulk (in V)	0.9
CJ	C_{j0}	Junction-depletion capacitance at 0-V bias (in F/m^2)	$2.5 \times 10^{-4}:4.0 \times 10^{-4}$
CJSW	$C_{j\text{-}sw0}$	Sidewall capacitance at 0-V bias (in F/m)	$2.0 \times 10^{-10}:2.8 \times 10^{-10}$
MJ	m_j	Bulk-to-junction exponent (grading coefficient)	0.5
MJSW	$m_{j\text{-}sw}$	Sidewall-to-junction exponent (grading coefficient)	0.3

The most important dc parameters are the transistor current gain, β, specified by the SPICE parameter BF; the transistor-transport saturation current, I_{CS}, specified using the parameter IS; and the Early-voltage constant, specified by the parameter VAF. Typical values for these might be 100, 10^{-17} A, and 50 V, respectively. If one wants to model the transistor in reverse mode (where the emitter voltage is higher than the collector voltage for an npn), then one might specify BR, ISS, and VAR, as well; these are the parameters that correspond to BIF, IS, and VAF in the reverse direction. Typically, this reverse-mode modelling is not important for most circuits. Some other important dc parameters for accurate simulations are the base, emitter, and collector resistances, which are specified by RB, RE, and RC, respectively. It is especially important to specify RB (which might be 200 Ω to 500 Ω).

The important capacitance parameters and their corresponding SPICE parameters include the depletion capacitances at 0-V bias voltage, CJE, CJC, CJS; their grading coefficients, MJE, MJC, MJS; and their built-in voltages, VJE, VJC, VJS, for base-emitter, base-collector, and collector-substrate junctions. Again, the 0-V depletion capacitances should be specified in absolute values for a unit-sized transistor. Normally the base-emitter and base-collector junctions are graded (i.e., MJE, MJC = 0.33), whereas the collector-substrate junction may be either abrupt (MJS = 0.5) or graded (MJS = 0.5), depending on processing details. Typical built-in voltages might be 0.75 V to 0.8 V. In addition, for accurate simulations, one should specify the forward-base transit time, τ_F, specified by TF, and, if the transistor is to be operated in reverse mode or under saturated conditions, the reverse-base transit time, τ_R, specified by TR.

The most important of the model parameters just described are summarized in Table 1.3.

Once again, many other parameters can be specified if accurate simulation is desired. Other parameters might include those to model β degradation under high or

Table 1.3 The most important SPICE parameters for modelling BJTs

SPICE Parameter	Model Constant	Brief Description	Typical Value
BF	β	Transistor current gain in forward direction	100
ISS	I_{CS}	Transport saturation current in forward direction	2×10^{-18} A
VAF	V_A	Early voltage in forward direction	50 V
RB	r_b	Series base resistance	500 Ω
RE	R_E	Series emitter resistance	30 Ω
CJE	C_{je0}	Base-emitter depletion capacitance at 0 V	0.015 pF
CJC	C_{je0}	Base-collector depletion capacitance at 0 V	0.018 pF
CJS	C_{je0}	Collector-substrate depletion capacitance at 0 V	0.040 pF
MJE	m_e	Base-emitter junction exponent (grading factor)	0.30
MJC	m_c	Base-collector junction exponent (grading factor)	0.35
MJS	m_s	Collector-substrate junction exponent (grading factor)	0.29
TF	τ_F	Forward-base transit time	12 ps
TR	τ_R	Reverse-base transit time	4 ns

low current applications and parameters for accurate noise and temperature analysis. Readers should refer to their SPICE manuals for descriptions of these parameters.

1.7 APPENDIX

The purpose of this appendix is to present derivations for device equations that rely heavily on device physics knowledge. Specifically, equations are derived for the exponential relationship and diffusion capacitance of diodes, for the threshold voltage and triode relationship for MOS transistors, and for the exponential relationship and base charge storage for bipolar transistors.

Diode Exponential Relationship

The concentration of minority carriers in the bulk, far from the junction, is given by Eqs. (1.2) and (1.4). Close to the junction, the minority-carrier concentrations are much larger. Indeed, the concentration next to the junction increases exponentially with the external voltage, V_D, that is applied in the forward direction. The concentration of holes in the n side next to the junction, p_n, is given by [Sze, 1981]

$$p_n = p_{n0} e^{V_D/V_T} = \frac{n_i^2}{N_D} e^{V_D/V_T} \tag{1.178}$$

Similarly, the concentration of electrons in the p side next to the junction is given by

$$n_p = n_{p0}e^{V_D/V_T} = \frac{n_i^2}{N_A}e^{V_D/V_T} \tag{1.179}$$

As the carriers diffuse away from the junction, their concentration exponentially decreases. The relationship for holes in the n side is

$$p_n(x) = p_n(0)e^{-x/L_p} \tag{1.180}$$

where x is the distance from the junction and L_p is a constant known as the diffusion length for holes in the n side. Similarly, for electrons in the p side we have

$$n_p(x) = n_p(0)e^{-x/L_n} \tag{1.181}$$

where L_n is a constant known as the diffusion length of electrons in the p side. Note that p_n (0) and n_p (0) are given by (1.178) and (1.179), respectively. Note also that the constants L_n and L_p are dependent on the doping concentrations N_A and N_D, respectively.

The current density of diffusing carriers moving away from the junction is given by the well-known diffusion equations [Sze, 1981]. For example, the current density of diffusing electrons is given by

$$J_{D\text{-}n} = -qD_n\frac{dn_p(x)}{dx} \tag{1.182}$$

where D_n is the diffusion constant of electrons in the p side of the junction. The negative sign is present because electrons have negative charge. Note that $D_n = (kT/q)\mu_n$, where μ_n is the mobility of electrons. Using (1.181), we have

$$\frac{dn_p(x)}{dx} = \frac{n_p(0)}{L_n}e^{-x/L_n} = -\frac{n_p(x)}{L_n} \tag{1.183}$$

Therefore

$$J_{D\text{-}n} = \frac{qD_n}{L_n}n_p(x) \tag{1.184}$$

Thus, the current density due to diffusion is proportional to the minority-carrier concentration. Next to the junction, all the current flow results from the diffusion of minority carriers. Further away from the junction, some of the current flow is due to diffusion and some is due to majority carriers drifting by to replace carriers that recombined with minority carriers or diffused across the junction.

Continuing, we use (1.179) and (1.184) to determine the current density next to the junction of electrons in the p side:

$$\begin{aligned} J_{D\text{-}n} &= \frac{qD_n}{L_n}n_p(0) \\ &= \frac{qD_n}{L_n}\frac{n_i^2}{N_A}e^{V_D/V_T} \end{aligned} \tag{1.185}$$

For the total current of electrons in the p side, we multiply (1.185) by the effective junction area, A_D. The total current remains constant as we move away from the junction since, in the steady state, the minority carrier concentration at any particular location remains constant with time. In other words, if the current changed as we moved away from the junction, the charge concentrations would change with time.

Using a similar derivation, we obtain the total current of holes in the n side, $I_{D\text{-}p}$, as

$$I_{D\text{-}p} = \frac{A_D q D_n n_i^2}{L_p N_D} e^{V_D/V_T} \tag{1.186}$$

where D_n is the diffusion constant of electrons in the p side of the junction, L_p is the diffusion length of holes in the n side, and N_D is the impurity concentration of donors in the n side. This current, consisting of positive carriers, flows in the direction opposite to that of the flow of minority electrons in the p side. However, since electron carriers are negatively charged, the direction of the current flow is the same. Note also that if the p side is more heavily doped than the n side, most of the carriers will be holes, whereas if the n side is more heavily doped than the p side, most of the carriers will be electrons.

The total current is the sum of the minority currents at the junction edges:

$$I_D = A_D q n_i^2 \left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right) e^{V_D/V_T} \tag{1.187}$$

Equation (1.187) is often expressed as

$$I_D = I_S e^{V_D/V_T} \tag{1.188}$$

where

$$I_S = A_D q n_i^2 \left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right) \tag{1.189}$$

Equation (1.188) is the well-known exponential current-voltage relationship of forward-biased diodes.

The concentrations of minority carriers near the junction and the direction of current flow are shown in Fig. 1.35.

Diode-Diffusion Capacitance

To find the diffusion capacitance, C_d, we first find the minority charge close to the junction, Q_d, and then differentiate it with respect to V_D. The minority charge close to the junction, Q_d, can be found by integrating either (1.180) or (1.181) over a few diffusion lengths. For example, if we assume n_{p0}, the minority electron concentration in the p side far from the junction is much less than $n_p(0)$, the minority electron con-

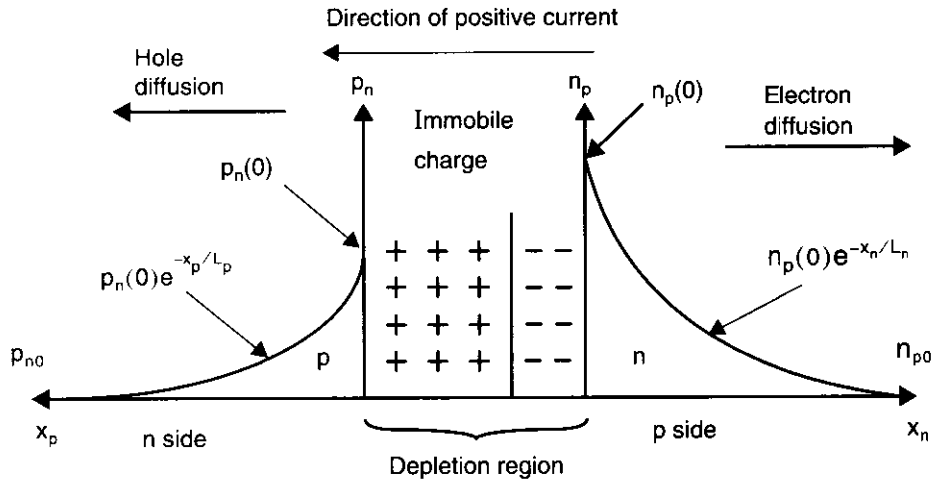


Fig. 1.35 The concentration of minority carriers and the direction of diffusing carriers near a forward-biased junction.

centration at the junction edge, we can use (1.181) to obtain

$$\begin{aligned} Q_n &= qA_D\int_0^\infty n_p(x)\,dx \\ &= qA_D\int_0^\infty n_p(0)e^{-X/L_n}dx \\ &= qA_DL_nn_p(0) \end{aligned} \tag{1.190}$$

Using (1.4) for $n_p(0)$ results in

$$Q_n = \frac{qA_DL_nn_i^2}{N_A}e^{V_D/V_T} \tag{1.191}$$

In a similar manner, we also have

$$Q_p = \frac{qA_DL_nn_i^2}{N_D}e^{V_D/V_T} \tag{1.192}$$

For a typical junction, one side will be much more heavily doped than the other side, and therefore the minority charge storage in the heavily doped side can be ignored since it will be much less than that in the lightly doped side. Assuming the n side is heavily doped, we find the total charge, Q_d, to be approximately given by Q_n, the minority charge in the p side. Thus, the small-signal diffusion capacitance, C_d, is given by

$$C_d = \frac{dQ_d}{dV_D} \cong \frac{dQ_n}{dV_D} = \frac{qA_DL_nn_i^2}{N_AV_T}e^{V_D/V_T} \tag{1.193}$$

Using (1.187) and again noting that $N_D >> N_A$, we have

$$C_d = \frac{L_n^2 I_D}{D_n V_T} \tag{1.194}$$

Equation (1.194) is often expressed as

$$C_d = \tau_T \frac{I_D}{V_T} \tag{1.195}$$

where τ_T is the transit time of the diode given by

$$\tau_T = \frac{L_n^2}{D_n} \tag{1.196}$$

for a single-sided diode in which the n side is more heavily doped.

MOS Threshold Voltage and the Body Effect

Many factors affect the gate-source voltage at which the channel becomes conductive. These factors are as follows:

1. The work-function difference between the gate material and the substrate material
2. The voltage drop between the channel and the substrate required for the channel to exist
3. The voltage drop across the thin oxide required for the depletion region, with its immobile charge, to exist
4. The voltage drop across the thin oxide due to unavoidable charge trapped in the thin oxide
5. The voltage drop across the thin oxide due to implanted charge at the surface of the silicon. The amount of implanted charge is adjusted in order to realize the desired threshold voltage.

The first factor affecting the transistor threshold voltage, V_{th}, is the built-in *Fermi potential* due to the different materials and doping concentrations used for the gate material and the substrate material. If one refers these potentials to that of intrinsic silicon [Tsividis, 1987], we have

$$\phi_{F\text{-}Gate} = \frac{kT}{q} \ln\left(\frac{N_D}{n_i}\right) \tag{1.197}$$

for a polysilicon gate with doping concentration N_D, and

$$\phi_{F\text{-}Sub} = \frac{kT}{q} \ln\left(\frac{n_i}{N_A}\right) \tag{1.198}$$

for a p substrate with doping concentration N_A. The work-function difference is then given by

$$\phi_{MS} = \phi_{F\text{-}Sub} - \phi_{F\text{-}Gate}$$
$$= \frac{kT}{q} \ln\left(\frac{N_D N_A}{n_i^2}\right) \tag{1.199}$$

The next factor that determines the transistor threshold voltage is the voltage drop from the channel to the substrate, which is required for the channel to exist. The question of exactly when the channel exists does not have a precise answer. Rather, the channel is said to exist when the concentration of electron carriers in the channel is equal to the concentration of holes in the substrate. At this gate voltage, the channel is said to be *inverted.* As the gate voltage changes from a low value to the value at which the channel becomes inverted, the voltage drop in the silicon also changes, as does the voltage drop in the depletion region between the channel and the bulk. After the channel becomes inverted, any additional increase in gate voltage is closely equal to the increase in voltage drop across the thin oxide. In other words, after channel inversion, gate voltage variations have little effect on the voltage drop in the silicon or the depletion region between the channel and the substrate.

The electron concentration in the channel is equal to the hole concentration in the substrate when the voltage drop from the channel to the substrate is equal to two times the difference between the Fermi potential of the substrate and intrinsic silicon, ϕ_F, where

$$\phi_F = -\frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right) \tag{1.200}$$

Equation (1.200) is a factor in several equations used in modelling MOS transistors. For typical processes, ϕ_F can usually be approximated as 0.35 V for typical doping levels at room temperature.

The third factor that affects the threshold voltage is due to the immobile negative charge in the depletion region left behind after the p mobile carriers are repelled. This effect gives rise to a voltage drop across the thin oxide of $-Q_B/C_{ox}$, where

$$Q_B = -qN_A x_d \tag{1.201}$$

and x_d is the width of the depletion region. Since

$$x_d = \sqrt{\frac{2K_s\varepsilon_0|2\phi_F|}{qN_A}} \tag{1.202}$$

we have

$$Q_B = -\sqrt{2qN_A K_s \varepsilon_0 |2\phi_F|} \tag{1.203}$$

The fourth factor that determines V_{tn} is due to the unavoidable charge trapped in the thin oxide. Typical values for the effective ion density of this charge, N_{ox}, might be 2×10^{14} to 10^{15} ions/m^3. These ions are almost always positive. This effect gives rise to a voltage drop across the thin oxide, V_{ox}, given by

$$V_{ox} = \frac{-Q_{ox}}{C_{ox}} = \frac{-qN_{ox}}{C_{ox}} \tag{1.204}$$

The *native transistor threshold voltage* is the threshold voltage that would occur naturally if one did not include a special ion implant used to adjust the threshold voltage. This value is given by

$$V_{t\text{-native}} = \phi_{MS} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} \tag{1.205}$$

A typical native threshold value might be around –0.1 V. It should be noted that transistors that have native transistor threshold voltages are becoming more important in circuit design where they might be used in transmission gates or in source-follower buffers.

The fifth factor that affects threshold voltage is a charge implanted in the silicon under the gate to change the threshold voltage from that given by (1.205) to the desired value, which might be 0.7 V for an n-channel transistor.

For the case in which the source-to-substrate voltage is increased, the effective threshold voltage is increased. This is known as the *body effect.* The body effect occurs because, as the source-bulk voltage, V_{SB}, becomes larger, the depletion region between the channel and the substrate becomes wider, and therefore more immobile negative charge becomes uncovered. This increase in charge changes the third factor in determining the transistor threshold voltage. Specifically, instead of using (1.203) to determine Q_B, one should now use

$$Q_B = -\sqrt{2qN_AK_s\varepsilon_0(V_{SB} + |2\phi_F|)} \tag{1.206}$$

If the threshold voltage when $V_{SB} = 0$ is denoted V_{tn0}, then, using (1.205) and (1.206), one can show that

$$\begin{aligned} V_{tn} &= V_{tn0} + \Delta V_{tn} \\ &= V_{tn0} + \frac{\sqrt{2qN_AK_s\varepsilon_0}}{C_{ox}}\left[\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right] \\ &= V_{tn0} + \gamma\left(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \end{aligned} \tag{1.207}$$

where

$$\gamma = \frac{\sqrt{2qN_AK_s\varepsilon_0}}{C_{ox}} \tag{1.208}$$

The factor γ is often called the *body-effect constant.*

MOS Triode Relationship

The current flow in a MOS transistor is due to *drift* current rather than diffusion current. This type of current flow is the same mechanism that determines the current in a resistor. The current density, J, is proportional to the electrical field, E, where the constant of proportionality, σ, is called the *electrical permittivity.* Thus,

$$J = \sigma E \tag{1.209}$$

This constant for an n-type material is given by

$$\sigma = qn\mu_n \tag{1.210}$$

where n is the concentration per unit volume of negative carriers and μ_n is the mobility of electrons. Thus, the current density is given by

$$J = qn\mu_n E \tag{1.211}$$

Next, consider the current flow through the volume shown in Fig. 1.36, where the volume has height H and width W. The current is flowing perpendicular to the plane $H \times W$ down the length of the volume, L. The current, I, everywhere along the length of the volume is given by

$$I = JWH \tag{1.212}$$

The voltage drop along the length of the volume in the direction of L for a distance dx is denoted dV and is given by

$$dV = E(x)\,dx \tag{1.213}$$

Combining (1.211), (1.212), and (1.213), we obtain

$$q\mu_n WHn(x)\,dV = I\,dx \tag{1.214}$$

where the carrier density n(x) is now assumed to change along the length L and is therefore a function of x.

As an aside, we examine the case of a resistor where n(x) is usually constant. A resistor of length L would therefore have a current given by

$$I = \frac{q\mu_n WH}{L}\Delta V \tag{1.215}$$

Thus, the resistance is given by

$$R = \frac{L}{q\mu_n WH} \tag{1.216}$$

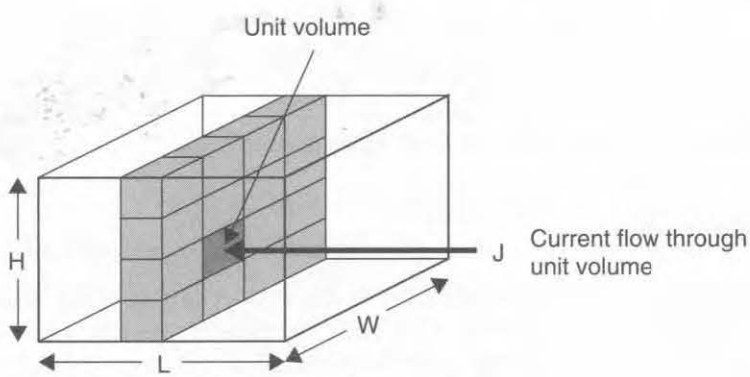


Fig. 1.36 Current flowing through a unit volume.

Often this resistance is presented in a relative manner, in which the length and width are removed (since they can be design parameters) but the height remains included. In this case, the resulting expression is commonly referred to as the *resistance per square* and designated as $R_\square$ where

$$R_\square = q\mu_n H \tag{1.217}$$

The total resistance is then given by

$$R_{total} = R_\square \frac{L}{W} \tag{1.218}$$

This equation is important when calculating the resistance of interconnects used in integrated circuits.

In the case of a MOS transistor, the charge density is not constant down the channel. If, instead of the carrier density per unit volume, one expresses $n(x)$ as a function of charge density per square area from the top looking down, we have

$$Q_n(x) = qHn(x) \tag{1.219}$$

Substituting (1.219) into (1.214) results in

$$\mu_n W Q_n(x)\, dV = I\, dx \tag{1.220}$$

Equation (1.220) applies to drift current through any structure that has varying charge density in the direction of the current flow. It can also be applied to a MOS transistor in the triode region to derive its I-V relationship. It should be noted here that in this derivation, it is assumed the source voltage is the same as the substrate voltage.

Since the transistor is in the triode region, we have $V_{DG} < -V_{tn}$. This requirement is equivalent to $V_{DS} < V_{GS} - V_{tn} = V_{eff}$. It is assumed that the effective channel length is L. Assuming the voltage in the channel at distance x from the source is given by $V_{ch}(x)$, from Fig. 1.37, we have

$$Q_n(x) = C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}] \tag{1.221}$$

Substituting (1.221) into (1.220) results in

$$\mu_n W C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}]dV_{ch} = I_D\, dx \tag{1.222}$$

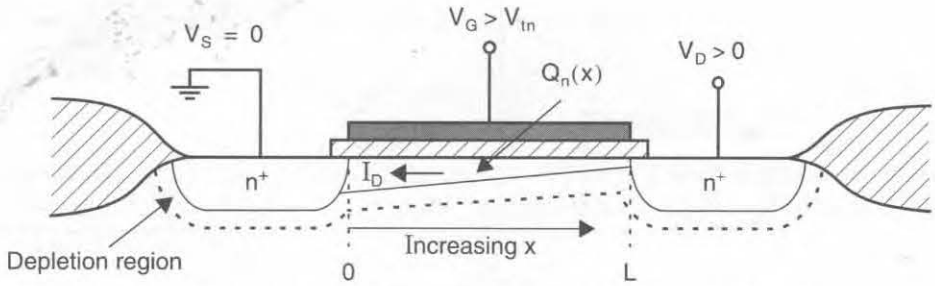


Fig. 1.37 The transistor definitions used in developing the transistor's I-V relationship.

Integrating both sides of (1.222), and noting that the total voltage along the channel of length L is V_{DS}, we obtain

$$\int_0^{V_{DS}} \mu_n W C_{ox}[V_{GS} - V_{ch}(x) - V_{tn}]\, dV_{ch} = \int_0^L I_D\, dx \tag{1.223}$$

which results in

$$\mu_n W C_{ox}\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] = I_D L \tag{1.224}$$

Thus, solving for I_D results in the well-known triode relationship for a MOS transistor:

$$I_D = \mu_n C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{1.225}$$

It should be noted that taking into account the body effect along the channel, the triode model of (1.225) is modified to

$$I_D = \mu C_{ox}\left(\frac{W}{L}\right)\left[(V_{GS} - V_{tn})V_{DS} - \alpha\frac{V_{DS}^2}{2}\right] \tag{1.226}$$

where $\alpha \cong 1.7$ [Tsividis, 1987].

Bipolar Transistor Exponential Relationship

The various components of the base, collector, and emitter currents were shown in Fig. 1.28, on page 44. Figure 1.38 shows plots of the minority-carrier concentrations

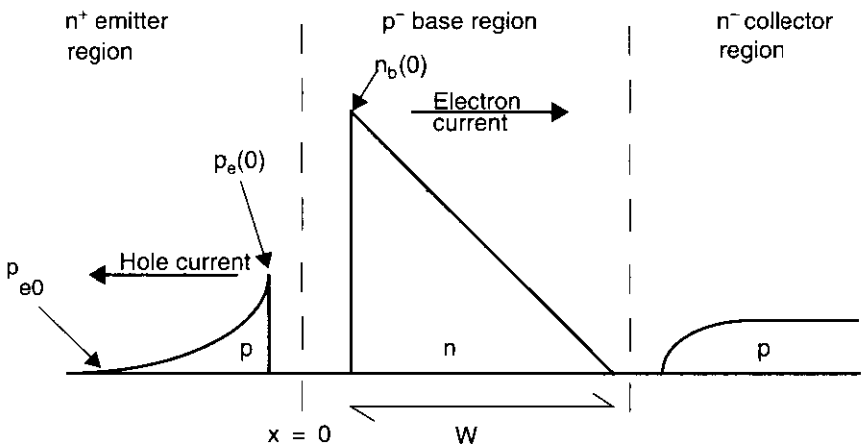


Fig. 1.38 The concentrations of minority carriers in the emitter, base, and collector.

in the emitter, base, and collector regions. The current flow of these minority carriers is due to diffusion. By calculating the gradient of the minority-carrier concentrations near the base-emitter junction in a manner similar to that used for diodes, it is possible to derive a relationship between the electron current and the hole current of Fig. 1.28.

The concentration of holes in the emitter at the edge of the base-emitter depletion region is denoted $p_e(0)$. This concentration decreases exponentially the farther one gets from the junction, in a manner similar to that described for diodes. The concentration far from the junction, p_e0, is given by

$$p_{e0} = \frac{n_i^2}{N_D} \tag{1.227}$$

where N_D is the doping density of the n^+ emitter. At a distance x from the edge of the emitter-base depletion region, we have

$$p_e(x) \cong p_e(0)e^{-x/L_p} \tag{1.228}$$

where

$$\begin{aligned} p_e(0) &= p_{e0}e^{V_{BE}/V_T} \\ &= \frac{n_i^2}{N_D}e^{V_{BE}/V_T} \end{aligned} \tag{1.229}$$

and where V_{BE} is the forward-bias voltage of the base-emitter junction.

At the edge of the base-emitter depletion region, the gradient of the hole concentration in the emitter is found, using (1.228), to be

$$\left.\frac{dp_e(x)}{dx}\right|_{x=0} = \frac{p_e(0)}{L_p} \tag{1.230}$$

Using (1.229), we can rewrite this as

$$\left.\frac{dp_e(0)}{dx}\right|_{x=0} = \frac{n_i^2}{L_pN_D}e^{V_{BE}/V_T} \tag{1.231}$$

The hole current is now found using the diffusion equation

$$I_{pe} = A_EqD_p\left.\frac{dp_e(x)}{dx}\right|_{x=0} \tag{1.232}$$

where A_E is the effective area of the emitter. Recall that the minority-hole current in the emitter, I_{pe}, is closely equal to the base current, I_B. After combining (1.231) and (1.232), we obtain

$$I_B = \frac{A_EqD_pn_i^2}{L_pN_D}e^{V_{BE}/V_T} \tag{1.233}$$

The situation on the base side of the base-emitter junction is somewhat different. The concentration of the minority carriers, in this case electrons that diffused from the emitter, is given by a similar equation,

$$n_b(0) = \frac{n_i^2}{N_A} e^{V_{BE}/V_T} \tag{1.234}$$

However, the gradient of this concentration at the edge of the base-emitter depletion region is calculated differently. This difference in gradient concentration is due to the close proximity of the collector-base junction, where the minority carrier (electron) concentration, $n_b(W)$, must be zero. This zero concentration at the collector-base junction occurs because any electrons diffusing to the edge of the collector-base depletion region immediately drift across the junction to the collector, as stated previously. If the base "width" is much shorter than the diffusion length of electrons in the base, L_n, then almost no electrons will recombine with base majority carriers (holes) before they diffuse to the collector-base junction. Given this fact, the decrease in electron or minority concentration from the base-emitter junction to the collector-base junction is a linear relationship decreasing from $n_b(0)$ at the emitter junction to zero at the collector junction in distance W. This assumption ignores any recombination of electrons in the base as they travel to the collector, which is a reasonable assumption for modern transistors that have very narrow bases. Thus, throughout the base region, the gradient of the minority-carrier concentration is closely given by

$$\begin{aligned} \frac{dn_b(x)}{dx} &= -\frac{n_b(0)}{W} \\ &= -\frac{n_i^2}{WN_A} e^{V_{BE}/V_T} \end{aligned} \tag{1.235}$$

Combining (1.235) with the diffusion equation, we obtain

$$\begin{aligned} I_{nb} &= -A_E q D_n \frac{dn_b(0)}{dx} \\ &= \frac{A_E q D_n n_i^2}{WN_A} e^{V_{BE}/V_T} \end{aligned} \tag{1.236}$$

Remembering that I_{nb} is closely equal to the collector current I_C, we have

$$I_C \cong I_{CS} e^{V_{BE}/V_T} \tag{1.237}$$

where

$$I_{CS} = \frac{A_E q D_n n_i^2}{WN_A} \tag{1.238}$$

The ratio of the collector current to the base current, commonly called the *transistor common-emitter current gain* and denoted β, is found using (1.237), (1.238), and

(1.233). We have

$$\beta = \frac{I_C}{I_B} = \frac{D_n N_D L_p}{D_p N_A W} \cong 2.5 \frac{N_D L_p}{N_A W} \tag{1.239}$$

which is a constant independent of voltage and current. Noting that $N_D >> N_A$, $L_p > W$, and $D_n \cong 2.5\ D_p$, we have $\beta >> 1$. A typical value might be between 50 and 200. The derivation of β just presented ignores many second-order effects that make β somewhat current and voltage dependent and are beyond the scope of this book. Interested readers should see Roulston, 1990, for more details. Regardless of second-order effects, equation (1.239) does reflect the approximate relationships among β, doping levels, and base width. For example, (1.239) explains why heavily doped emitters are important to achieve large current gain.

Base Charge Storage of an Active BJT

Figure 1.38 shows a minority-carrier storage in the base region, Q_b, given by

$$Q_b = A_E q \frac{n_b(0) W}{2} \tag{1.240}$$

Using (1.234) for $n_b(0)$, we have

$$Q_b = \frac{A_E q n_i^2 W}{2N_A} e^{V_{BE}/V_T} \tag{1.241}$$

This equation can be rewritten using (1.237) and (1.238) to obtain

$$Q_b = \frac{W^2}{2D_n} I_C = \tau_b I_C \tag{1.242}$$

where τ_b, called the *base-transit time constant,* is given approximately by

$$\tau_b = \frac{W^2}{2D_n} \tag{1.243}$$

ignoring second-order effects. Normally, the base-transit time constant is specified for a given technology and takes into account other charge-storage effects not considered here, and is therefore often denoted τ_T. However, since the base storage of electrons dominates the other effects, we have $\tau_T \cong \tau_b$.

If the current in a BJT changes, the base charge storage must also change. This change can be modelled by a diffusion capacitance, C_d, between the base and the emitter terminals. Using (1.242), we have

$$C_d \cong \frac{dQ_b}{dV_{BE}} = \frac{d(\tau_b I_C)}{dV_{BE}} \tag{1.244}$$

Using (1.244) and $I_C = I_{CS}e^{V_{BE}/V_T}$ results in

$$C_d = \tau_b \frac{I_C}{V_T} \tag{1.245}$$

This equation is similar to that for a diode.

1.8 REFERENCES

R. Geiger, P. Allen, and N. Strader. *VLSI: Design Techniques for Analog and Digital Circuits*. McGraw-Hill, New York, 1990.

P. Gray and R. G. Meyer. *Analog Integrated Circuits*, 3rd. ed. John Wiley & Sons, New York, 1993.

D. Hodges and H. Jackson. *Analysis and Design of Digital Integrated Circuits*, 2nd ed. McGraw-Hill, New York, 1988.

D. Roulston. *Semiconductor Devices*. McGraw-Hill, New York, 1990.

S. M. Sze. *Physics of Semiconductor Devices*. Wiley Interscience, New York, 1981.

Y. Tsividis. *Operation and Modeling of the MOS Transistor*. McGraw-Hill, New York, 1987.

S. Wolf. *Silicon Processing for the VLSI Era—Volume 3: The Submicron MOSFET*. Lattice Press, Sunset Beach, California, 1995.

1.9 PROBLEMS

Unless otherwise stated, assume the following hold throughout the problems section:

- Room temperature = 300 °K
- npn bipolar transistors:
  $\beta = 100$
  $V_A = 80$ V
  $\tau_b = 13$ ps
  $\tau_s = 4$ ns
  $r_b = 330\ \Omega$
- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8$ V
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}\ (\Omega) = 8000L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9$ V
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}\ (\Omega) = 12{,}000\ L\ (\mu m)/I_D\ (mA)$ in active region

$C_j = 4.5 \times 10^{-4}$ pF/(μm)^2
$C_{j\text{-}sw} = 2.5 \times 10^{-4}$ pF/μm
$C_{ox} = 1.9 \times 10^{-3}$ pF/(μm)^2
$C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}$ pF/μm

1.1 Estimate the hole and electron concentrations in silicon doped with arsenic at a concentration of 10^{25} atoms/m^3 at a temperature 22 °C above room temperature. Is the resulting material n type or p type?

1.2 For the pn junction of Example 1.2, does the built-in potential, Φ_0, increase or decrease when the temperature is increased 11 °C above room temperature?

1.3 Calculate the amount of charge per (μm)^2 in each of the n and p regions of the pn junction of Example 1.2 for a 5-V reverse-bias voltage. How much charge on each side would be present in a 10 μm × 10 μm diode?

1.4 A silicon diode has τ_t = 12 ps and C_{j0} = 15 fF. It is biased by a 43-kΩ resistor connected between the cathode of the diode and the input signal, as shown in Fig. P1.4. Initially, the input is 5 V, and then at time 0 it changes to 0 V. Estimate the time it takes for the output voltage to change from 5 V to 1.5 V (i.e., the $\Delta t_{-70\%}$ time). Repeat for an input voltage change from 0 V to 5 V and an output voltage change from 0 V to 3.5 V.

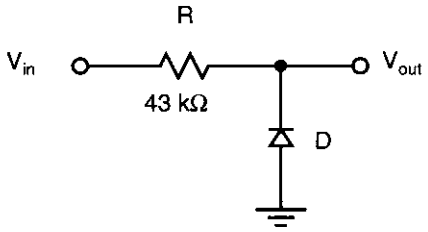


Fig. P1.4

1.5 Compare your answers for Problem 1.4 to those obtained using a SPICE simulation.

1.6 Verify that when $V_{DS} = V_{eff}$ is used in the triode equation for a MOS transistor, the current equals that of the active region equation given in (1.67).

1.7 Find I_D for an n-channel transistor having doping concentrations of $N_D = 10^{25}$ and $N_A = 10^{22}$ with $W = 50$ μm, $L = 1.5$ μm, $V_{GS} = 1.1$ V, and $V_{DS} = V_{eff}$. Assuming λ remains constant, estimate the new value of I_D if V_{DS} is increased by 0.3 V.

1.8 A MOS transistor in the active region is measured to have a drain current of 20 μA when $V_{DS} = V_{eff}$. When V_{DS} is increased by 0.5 V, I_D increases to 23 μA. Estimate the output impedance, r_{ds}, and the output impedance constant, λ.

1.9 Derive the low-frequency model parameters for an n-channel transistor having doping concentrations of $N_D = 10^{25}$ and $N_A = 10^{22}$ with $W = 10$ μm, $L = 1.2$ μm, $V_{GS} = 1.1$ V, and $V_{DS} = V_{eff}$. Assume that $V_{SB} = 1.0$ V.

1.10 Find the capacitances C_{gs}, C_{gd}, C_{db}, and C_{sb} for a transistor having $W = 50\ \mu m$ and $L = 1.2\ \mu m$. Assume that the source and drain junctions extend $4\ \mu m$ beyond the gate, resulting in source and drain areas being $A_s = A_d = 200\ (\mu m)^2$ and the perimeter of each being $P_s = P_d = 58\ \mu m$.

1.11 Consider the circuit shown in Fig. P1.11, where V_{in} is a dc signal of 1 V. Taking into account only the channel charge storage, determine the final value of V_{out} when the transistor is turned off, assuming half the channel charge goes to C_L.

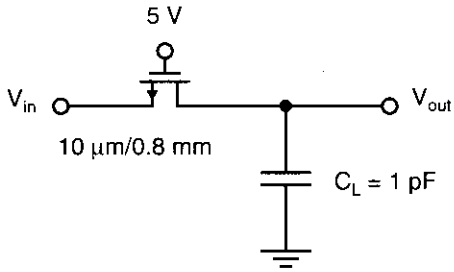


Fig. P1.11

1.12 For the same circuit as in Problem 1.11, the input voltage has a step voltage change at time 0 from 1 V to 1.2 V (the gate voltage remains at 5 V). Find its 99 percent settling time (the time it takes to settle to within 1 percent of the total voltage change). You may ignore the body effect and all capacitances except C_L. Also assume that $V_{tn} = V_{tn0}$. Repeat the question for V_{in} changing from 3 V to 3.1 V.

1.13 Repeat Problem 1.11, but now take into account the body effect on V_{tn}.

1.14 For an npn transistor having $I_C = 0.1$ mA, calculate g_m, r_π, r_e, r_0, and $g_m r_0$.

1.15 A bipolar junction transistor has the following SPICE parameters (the SPICE name for the parameter is included in parentheses):

I_S (IS) $= 2.0 \times 10^{-18}$
B_F (BF) $= 100$
B_R (BR) $= 1$
V_A (VA) $= 50$ V
τ_F (TF) $= 12 \times 10^{-12}$ s
τ_R (TR) $= 4 \times 10^{-9}$ s
C_{je0} (CJE) $= 15 \times 10^{-15}$ F
Φ_e (VJE) $= 0.9$ V
m_e (MJE) $= 0.27$
C_{jc0} (CJC) $= 18 \times 10^{-15}$ F
Φ_c (VJC) $= 0.7$ V
m_c (MJC) $= 0.37$
C_{js0} (CJS) $= 40 \times 10^{-15}$ F
Φ_s (VJS) $= 0.64$ V
m_s (MJS) $= 0.29$
R_e (RE) $= 30$

R_b (RB) = 500
R_c (RC) = 90

Initially, the circuit shown in Fig. P1.15 has a 0-V input. At time 0 its input changes to 5 V. Estimate the time it takes its output voltage to saturate, using the concepts of average capacitance and first-order transient solutions for each node. The time constants of the individual nodes can be added to arrive at an overall time constant for the approximate first-order transient response of the circuit. Next, assume that the input changes from 5 V to 0 V at time 0. How long does it take the output voltage to change to 3.5 V?

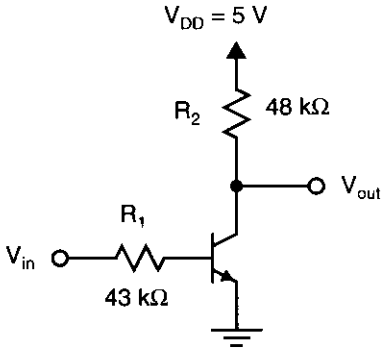


Fig. P1.15

1.16 Compare your answers to Problem 1.15 to those obtained using SPICE.

1.17 Verify that for $I_{BR} >> I_B$ and $I_{BR} >> I_C / \beta$, Eq. (1.140) simplifies to Eq. (1.139).

CHAPTER

2 Processing and Layout

This chapter describes the steps and processes used in realizing modern integrated circuits. While emphasis is placed on CMOS processing, the technology required for BJT circuits is also described. After processing is presented, circuit layout is covered. Layout is the design portion of integrated-circuit manufacturing, in which the geometry of circuit elements and wiring connections is defined. This process leads to the development of photographic masks used in manufacturing a microcircuit. The concepts of design rules and their relationship to microcircuits are emphasized. Next, circuit layout is related to the transistor models. Here, it is shown that once the layout is completed, the values of certain elements in the transistor models can be determined. This knowledge is necessary for accurate computer simulation of integrated circuits. It is also shown that, by using typical design rules, one can make reasonable assumptions to approximate transistor parasitic components before the layout has been done. Analog layout issues are then discussed, including matching and noise considerations. Finally, this chapter concludes with a discussion of a destructive mechanism in CMOS circuits known as latch-up.

2.1 CMOS PROCESSING

In this section, the basic steps involved in processing a CMOS microcircuit are presented. For illustrative purposes, we describe here an example n-well process (with, of course, a p substrate) and two layers of metal. Although the list is not complete, this section also describes many of the possible variations during processing.

The Silicon Wafer

The first step in realizing a microcircuit is to fabricate a defect-free, single-crystalline, lightly doped wafer. To realize such a wafer, one starts by creating metallurgical-grade silicon through the use of a high-temperature chemical process in an electrode-arc furnace. Although metallurgical-grade silicon is about 98 percent pure, it has far too many impurities for use in realizing microcircuits. Next, a silicon-containing gas is formed and then reduced. Pure silicon is precipitated onto thin rods of single-crystalline silicon. This deposited electronic-grade silicon is very pure

but, unfortunately, it is also polycrystalline. To obtain single-crystalline silicon, the silicon is melted once again and allowed to cool. As it cools, a single-crystalline *ingot* is slowly pulled and turned from the molten silicon using the *Czochralski method.* The Czochralski method starts with a seed of single crystal silicon, and the pull rate and speed of rotation determine the diameter of the crystalline rod or ingot. Typical diameters might be 10 to 20 cm (i.e., 4 to 8 inches) with lengths usually longer than 1 meter. Producing a silicon ingot can take several days.

Normally, heavily doped silicon is added to the *melt* before the single-crystalline ingot is pulled. After the doped silicon diffuses through the molten silicon, a lightly doped silicon ingot results. In our example process, boron impurities would be added to produce a p^- ingot.

The ingot is cut into wafers using a large diamond saw. A typical wafer might have a thickness of about 1 mm. After the ingot is sawed into wafers, each wafer is polished with Al_2O_3, chemically etched to remove mechanically damaged material, and then fine-polished again with SiO_2 particles in an aqueous solution of NaOH.

Very often, the company that produces the silicon wafers is not the same company that eventually patterns them into monolithic circuits. Sometimes, the surface of the wafer might be doped more heavily, and a single-crystal epitaxial layer of the opposite type might be grown over its surface before the wafer-manufacturing company delivers the wafers to the processing company. This layered approach results in an *epitaxial wafer.*

A starting wafer of p^- might be doped around the level of $N_A \cong 2 \times 10^{21}$ donor/m^3. Such a doping level would give a *resistivity* of 10 to 20 $\Omega \cdot$ cm.

Photolithography and Well Definition

Photolithography is a technique in which selected portions of a silicon wafer can be *masked out* so that some type of processing step can be applied to the remaining areas. Although photolithography is used throughout the manufacturing of an integrated circuit, here we describe this photographic process in the context of preparing the wafer for defining the well regions.[1]

Selective coverage for well definition is performed as follows. First, a glass mask, M_1, is created, which defines where the well regions will be located. The glass mask is created by covering the mask in photographic materials and exposing it to an electron beam, or *e beam,* in the regions corresponding to the well locations. Such exposure results in the well regions on the glass mask turning opaque, or dark. As a result, the glass mask can be thought of as a negative of one layer of the microcircuit. In a typical microcircuit process, ten to twenty different masks might be required. The typical cost for these masks is currently around $50,000. Because a high degree of precision is required in manufacturing these masks, often a company other than the processing

1. Wells are doped regions that will contain one of the two types of transistors realized in a CMOS process. Nowadays, wells are normally n type and contain p-channel transistors.

company makes the masks. The exposure of the opaque regions of the mask, by the electron beam, is controlled by a computer dependent on the contents of a database. The database required for the e beam is derived from the layout database produced by the designer, using a computer and a layout CAD software program.

The first step in masking the surface of the wafer is to thermally grow a thin layer of silicon-dioxide (SiO_2) to protect the surface of the microcircuit. Details of this step are discussed later. On top of the SiO_2, a negative *photoresist*, PR_1, is evenly applied, while spinning the microcircuit to a thickness of around 1 μm. Photoresist is a light-sensitive polymer (similar to latex). In the next step, the mask, M_1, is placed in close proximity to the wafer, and ultraviolet light is projected through the mask onto the photoresist. Wherever the light strikes, the polymers cross-link, or *polymerize*. This change makes these regions insoluble to an organic solvent. This step is shown in Fig. 2.1. The regions where the mask was opaque (i.e., the well regions) are not exposed. The photoresist is removed in these areas using an organic solvent. Next, the remaining photoresist is baked to harden it. After the photoresist in the well regions is removed, the uncovered SiO_2 may also be removed using an acid etch. (However, in some processes where this layer is very thin, it may not be removed.) In the next step, the dopants needed to form the well are introduced into the silicon using either diffusion or ion implantation (directly through the thin oxide, in cases where it has not been removed).

The procedure just described involves a *negative photoresist,* where the exposed photoresist remains after the masking. There are also *positive photoresists,* in which the exposed photoresist is dissolved by the organic solvents. In this case, the photoresist remains where the mask was opaque. By using both positive and negative resists,

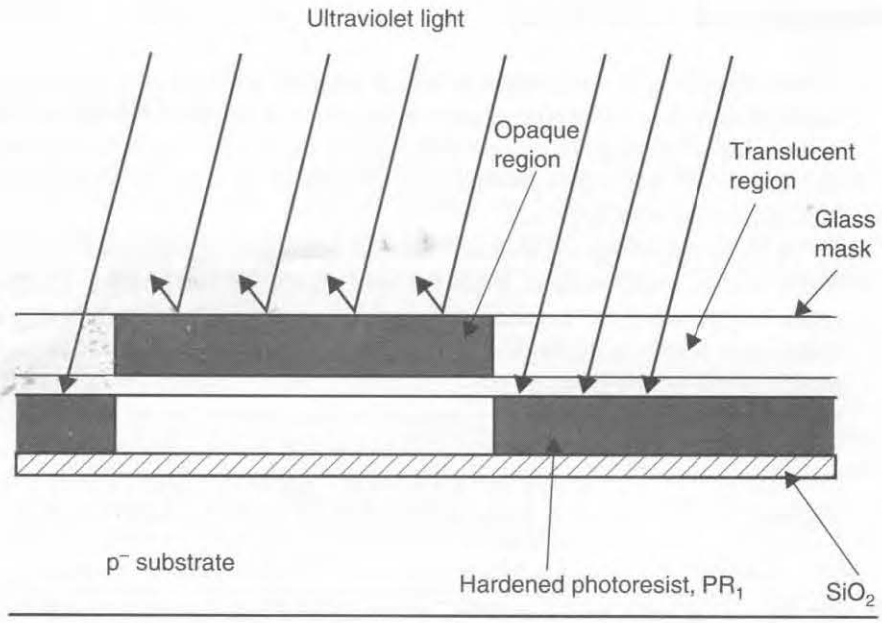


Fig. 2.1 Selectively hardening a region of photoresist using a glass mask.

a single mask can sometimes be used for two steps—first, to protect one region and implant the complementary region and second, to protect the complementary region and implant the original region.

Diffusion and Ion Implantation

After the photoresist over the well region has been removed, the next step is to introduce dopants through the opening where the well region will be located. As just mentioned, there are two approaches for introducing these dopants—diffusion and ion implantation.

In both implantation methods, usually the SiO_2 in the well region will first be removed using an acid etch. Next, the remaining hardened photoresist is stripped using acetone. This leaves SiO_2 that was protected by the hardened photoresist to mask all of the nonwell (i.e., substrate) regions.

In diffusion implantation, the wafers are placed in a quartz tube inside a heated furnace. A gas containing the dopant is introduced into the tube. In the case of forming an n well, the dopant in the gas would probably be phosphorus. Arsenic could also be used, but it takes a much longer time to diffuse. The high temperature of the diffusion furnace, typically 900 to 1,100 °C, causes the dopants to diffuse into the silicon both vertically and horizontally. The dopant concentration will be greatest at the surface and will decrease in a *Gaussian* profile further into the silicon. If a p well had been desired, then boron would have been used as the dopant. The resulting cross section, after diffusing the n well, is shown in Fig. 2.2.

An alternative technique for introducing dopants into the silicon wafer is ion implantation. This technique is now largely replacing diffusion because it allows more independent control over the dopant concentration and the thickness of the doped region. In ion implantation, dopants are introduced as ions into the wafer, as shown in the functional representation of an ion implanter in Fig. 2.3. The ions are generated by bombarding a gas with electrons from an arc-discharge or cold-cathode source. The ions are then focused and sent through a mass separator. This mass sepa-

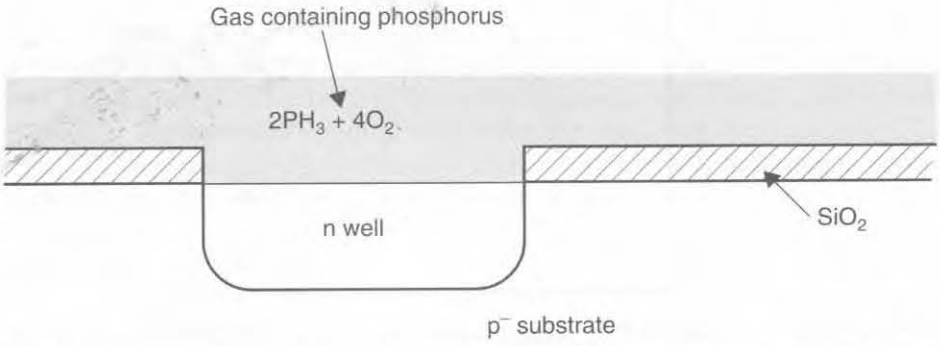


Fig. 2.2 Forming an n well by diffusing phosphorus from a gas into the silicon, through the opening in the SiO_2.

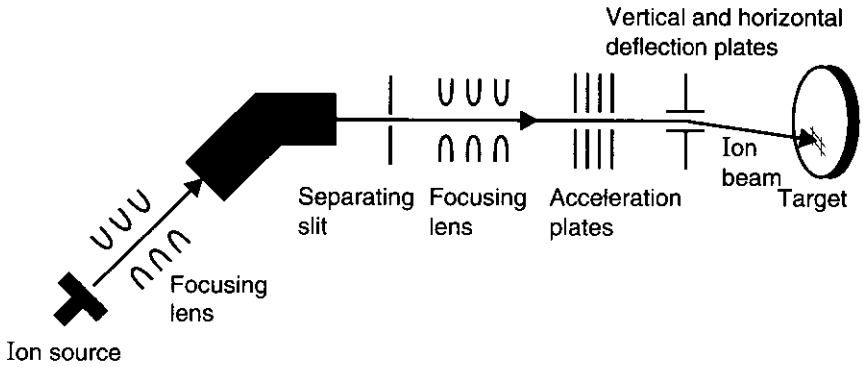


Fig. 2.3 An ion-implantation system.

rator bends the ion beam and sends it through a narrow slit. Since only ions of a specific mass pass through the slit, the beam is purified. Next, the beam is again focused and accelerated to between 10 keV and 1 MeV. The ion current might range from 10 μA to 2 mA. The deflection plates sweep the beam across the wafer (which is often rotated at the same time) and the acceleration potential controls how deeply the ions are implanted. The beam current and time of implantation determine the amount of dosage. Thus, depth and dosage are controlled independently. Two problems that occur with ion implantation are lattice damage and a narrow doping profile. The lattice damage is due to nuclear collisions that result in the displacement of substrate atoms. The narrow profile results in a heavy concentration over a narrow distance, as is shown in Fig. 2.4. For example, arsenic ions with an acceleration voltage of 100 keV might penetrate approximately 0.06 μm into the silicon, with the majority of

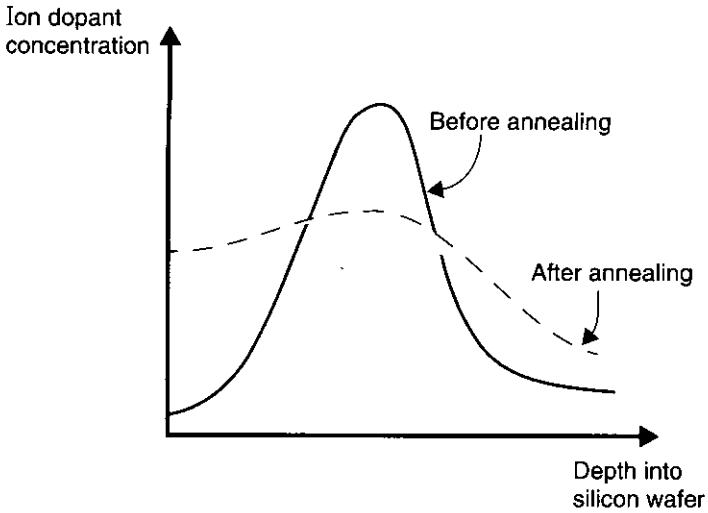


Fig. 2.4 Dopant profiles after ion implantation both before and after annealing.

ions being at 0.06 μm ± 0.02 μm. Both of these problems are largely solved by annealing.

Annealing is a step in which the wafer is heated to about 1,000 °C, perhaps for 15 to 30 minutes, and then allowed to cool slowly. This heating stage thermally vibrates the atoms, which allows the bonds to reform. Annealing also broadens the concentration profile, making the doping levels more uniform, as shown in Fig. 2.4. It should be noted that annealing is performed only once during processing, after all the implantation steps have been performed but before any metal layers have been created.[2]

For n-type dopants, arsenic is used for shallow implantations, such as the source or drain junctions. Phosphorus might be used for the well. Boron is always used to form the p regions.

Although more expensive, ion implantation has been largely replacing diffusion for forming n and p regions because of its greater control over doping levels. Another important advantage of ion implantation is the much smaller sideways diffusion, which allows devices to be more closely spaced and, more importantly for MOS transistors, minimizes the overlap between the gate-source or gate-drain regions.

Chemical Vapor Deposition and Defining the Active Regions

The next few steps use the field-oxide mask, M_2, to form the thick field-oxide as well as the field implants (used to isolate devices). These steps result in a thin layer of thermal SiO_2, as well as a layer of silicon nitride (Si_3N_4), everywhere the field-oxide is not desired.

Often, this step will be done using positive photoresist such that, wherever the mask M_2 is not opaque, the photoresist will be softened. In other words, the photoresist is left intact after the organic dissolution step, under the opaque regions of the mask where the field-oxide is not desired. A thin layer of thermal SiO_2 is first grown everywhere to protect the surface of the silicon lattice.

Next, Si_3N_4 is deposited everywhere during a gas-phase reaction in which energy is supplied by heat (at about 850 °C). This process is called *chemical vapor deposition,* or CVD. After this step, the positive photoresist is deposited, exposed through the mask, M_2, dissolved, and hardened. The hardened photoresist is left on top of the Si_3N_4 to protect it where the field-oxide is not desired. Next, the Si_3N_4, wherever it is not protected by the photoresist, is removed by etching it away with a hot phosphoric acid. The SiO_2 is then removed with a hydrofluoric acid etch. Finally, the remaining photoresist is chemically removed with a process that leaves the remaining Si_3N_4 intact. The remaining Si_3N_4 will act as a mask to protect the active regions when the thick field-oxide is being grown in the field-oxide regions.

Field-Implants and the Field-Oxide

The next step in our example process is to implant the *field-implants* under where the *field-oxide* will be grown. For example, boron will be implanted under the field-oxide

2. If annealing were done after deposition of a metal layer, the metal would melt.

everywhere except in the well regions. This implant guarantees that the silicon under the field-oxide will never invert (or become n) when a conductor over the field-oxide has a large voltage. If this implant were not performed, leakage currents could occur between the junctions of separate n-channel transistors in the substrate region, which are intended to be unconnected.

For the field-oxide in the well regions, where p-channel transistors will eventually reside, an n-type implant such as arsenic (As) could be used. Often, it is not necessary to include field-implants under the field-oxide of the well regions because the heavier doping of the well (compared to that of the substrate) normally guarantees that the silicon will never invert under the field-oxide in these regions.

When implanting the field-implants in the substrate regions, it is necessary to first cover the wells with a protective photoresist, PR_3, so the n-well regions do not receive the p implant. This can be done using the same mask, M_1, that was originally used for implanting the n wells, but now a positive photoresist is used. This positive photoresist remains where the mask is opaque (i.e., dark), which corresponds to the well regions.

After the exposed photoresist has been dissolved, we now have the cross section shown in Fig. 2.5. Notice that at this step, all the active regions, where eventually the transistors will reside, are protected from the field implants by the Si_3N_4, SiO_2, and PR_2. Additionally, the complete well regions are also protected by PR_3. The field-implant will be a high-energy implant with a fairly high doping level.

Growing the Field-Oxide

The next step is to grow the field-oxide, SiO_2. There are two different ways that SiO_2 can be grown. In a wet process, water vapor is introduced over the surface at a moderately high temperature. The water vapor diffuses into the silicon and, after some intermediate steps, reacts according to the formula

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \tag{2.1}$$

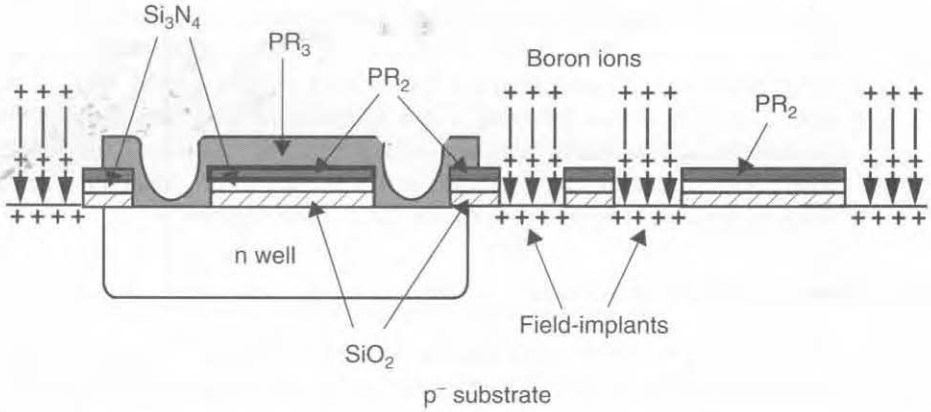


Fig. 2.5 The cross section when the field-implants are being formed.

In a dry process, oxygen is introduced over the wafer, normally at a slightly higher temperature than that used in the wet process, and reacts according to the formula

$$Si + O_2 \rightarrow SiO_2 \tag{2.2}$$

Since both of these processes occur at high temperatures, around 800 to 1200 °C, the oxide that results is sometimes called a *thermal oxide*.

Before the field-oxide is grown, PR_2 and PR_3 are removed, but the silicon-nitride–silicon-dioxide sandwich is left. The reaction does not occur wherever CVD-deposited Si_3N_4 remains, because the Si_3N_4 is relatively inert to both water and oxygen. Wherever the process does occur, the volume increases because oxygen atoms have been added. Specifically, SiO_2 takes up approximately 2.2 times the volume of the original silicon. This increase will cause the SiO_2 to extend approximately 45 percent into, and 55 percent above, what previously was the surface of the silicon. The resulting cross section is shown in Fig. 2.6. Note that in our example process, the field-oxide in the substrate region has field-implants under it, whereas the field-oxide in the wells does not.

When growing thermal SiO_2, the wet process is faster because H_2O diffuses faster in silicon than O_2 does, but the dry process results in denser, higher-quality SiO_2 that is less porous. Sometimes, growing the field-oxide starts with a dry process, changes to a wet process, and finishes with a dry process. When thin layers of SiO_2 are grown, as the next section describes, usually only a dry process is used.

Gate-Oxide and Threshold-Voltage Adjusts

In the next step, the Si_3N_4 is removed using hot phosphoric acid. If a thin layer of SiO_2 is under the Si_3N_4, protecting the surface, as shown in Fig. 2.6, this SiO_2 is also removed, usually with hydrofluoric acid. The high-quality, thin gate-oxide is then grown using a dry process. It is grown everywhere over the wafer to a thickness of about 0.01 to 0.03 μm.

After the gate-oxide has been grown, donors are implanted so that the final threshold voltages of the transistors are correct. Note that this implantation is performed directly through the thin gate-oxide since it now covers the entire surface. It should be mentioned here that many processes differ while realizing the threshold-adjust step.

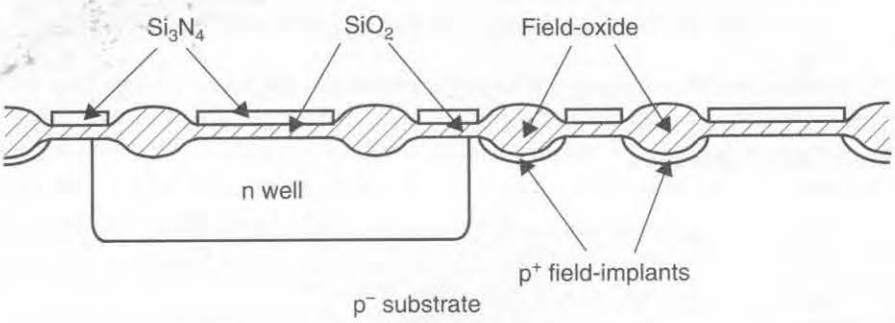


Fig. 2.6 The cross section after the field-oxide has been grown.

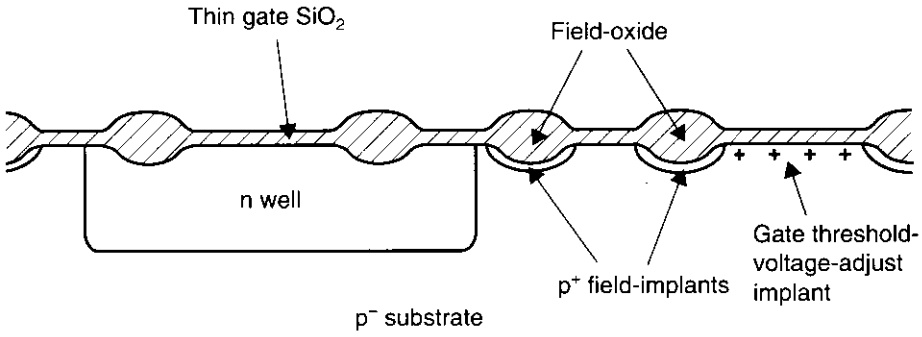


Fig. 2.7 Cross section after the thin gate-oxide growth and threshold-adjust implant.

In a simple process, the threshold voltages of both the p- and n-channel transistors are adjusted at the same time. We saw in the Appendix of Chapter 1 that the n-channel transistors require a boron implant to change V_{tn} from its native value of around −0.1 V to its desired value of 0.7 to 0.8 V. If the n wells are doped a little heavier than ideal, the native threshold voltage of the p-channel transistors in the well will be around −1.6 V. As a result, the same single boron threshold-adjust implant will bring it to around −0.8 to −0.9 V.

By using a single threshold-voltage-adjust implant for both n-channel and p-channel transistors, two photoresist masking steps are eliminated. If the different transistors are individually implanted, then the second of two types of transistors has to be protected by, say, a negative photoresist while the first type is being implanted. Next, a positive photoresist can be used with the same mask to protect the first type of transistor while the second type is being implanted. The mask used is normally the same mask used in forming the n wells, in other words, M_1. Thus, no additional mask is required, but a number of additional processing steps are needed. The major problem with using a single threshold-adjust implant is that the doping level of the n well is higher than optimum. This higher doping level increases the junction capacitances and the body effect of the transistors in the well. A double threshold adjust allows optimum well doping. Both approaches are currently in commercial use, although the double threshold-adjust implant is growing in favor as device dimensions decrease. The cross section at this stage is shown in Fig. 2.7.

Polysilicon Gate Formation

The next step in the process is the chemical deposition of the polysilicon gate material. One method to create polysilicon is to heat a wafer with silane gas flowing over it so the following reaction occurs

$$SiH_4 \rightarrow Si + 2H_2 \tag{2.3}$$

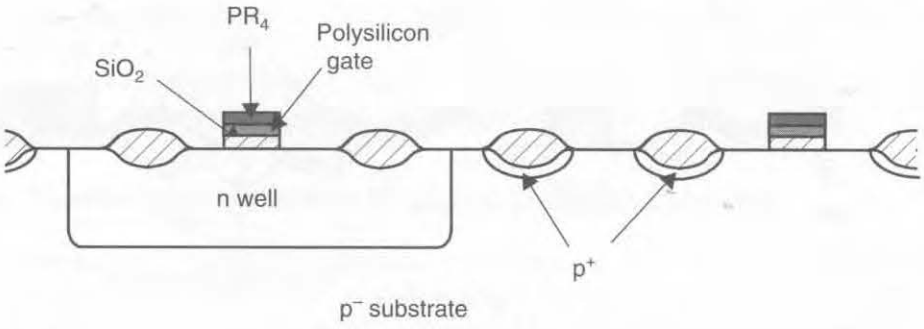


Fig. 2.8 Cross section after depositing and patterning the polysilicon gates.

If this reaction occurs at high temperatures, say, around 1,000 to 1,250 °C, and the original surface of the wafer was single crystal, the deposited silicon will also be single crystal. This approach is used both when epitaxial layers are grown in bipolar processes and in some of the more modern CMOS processes. However, when depositing the polysilicon gates, the original surface is SiO_2 and the wafer is heated only to about 650 °C. As a result, the silicon that is deposited is noncrystalline, or amorphous. Thus, this silicon is often referred to as *polysilicon.* Very often, after the polysilicon is deposited, it will be ion implanted with arsenic to increase its conductivity. A typical final resistivity for polysilicon might be 10 to 30 $\Omega/\square$, and its thickness might be around 0.25 μm.

After the deposition just described, the polysilicon gate material covers the entire wafer. This polysilicon is then patterned using a new mask, M_3, and a positive photoresist, PR_4. The mask is opaque where hardened polysilicon should remain. After the nonhardened photoresist is removed, the polysilicon is etched away using a reactive plasma etch. This etch removes all of the polysilicon not protected by photoresist but removes very little of the underlying SiO_2. This thin gate-oxide layer is used to protect the surface during the next step of junction implantation. The cross section at this phase is shown in Fig. 2.8.

Implanting the Junctions, Depositing SiO_2, and Opening Contact Holes

The next step involves the ion implantation of the junctions. In our example process, the p^+ junctions are formed first by placing positive photoresist, PR_5, everywhere except where the p^+ regions are desired. A new mask, M_4, is used in this step. The p^+ regions are then ion implanted, possibly through a thin oxide in some processes. The cross section at this stage is shown in Fig. 2.9.

Notice that the p^+ junctions of the p-channel transistors are defined on one edge by the field-oxide and, more importantly, next to the active gate area by the edge of the polysilicon gate. During the implantation of the boron, it was the gate polysilicon and the photoresist over it that protected the channel region from the p^+ implant.

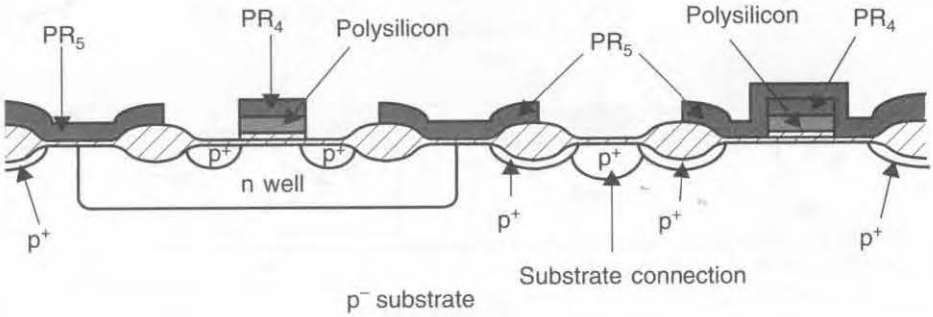


Fig. 2.9 Cross section after ion-implanting the p^+ junctions.

Thus, the p^+ junctions are *self-aligned* to the polysilicon gates, resulting in very little overlap (i.e., a small L_{ov}, as defined in Chapter 1). Also, note that the effective channel areas of the transistors are defined by the intersection of the gate-defining mask, M_3, and the mask used in defining the active regions, M_2 (i.e., the mask used in defining where Si_3N_4 remains). Thus, these are the two most important masks in any MOS process. The development of this self-aligned process has proven to be an important milestone in realizing small high-speed transistors, both in MOS processes and in BJT processes.

Also notice that a p^+ junction has been implanted in the substrate region. This junction, called a *substrate tie,* is used to connect the substrate to ground in microcircuits. These substrate ties are liberally placed throughout the microcircuit to help prevent latch-up, a problem discussed at the end of this chapter. In addition, the underside of the wafer would normally be connected to ground as well, through a package connection.

Next, the photoresists are all removed using acetone. The p^+ active regions are then protected using the same mask, M_4, that was used for the previous step, but now a negative photoresist, PR_6, is used. The n^+ junctions are then implanted using arsenic. The cross section at the end of this stage is shown in Fig. 2.10.

After the junctions have been implanted and PR_6 has been removed, the complete wafer is covered in CVD SiO_2. This protective *glass* layer can be deposited at moder-

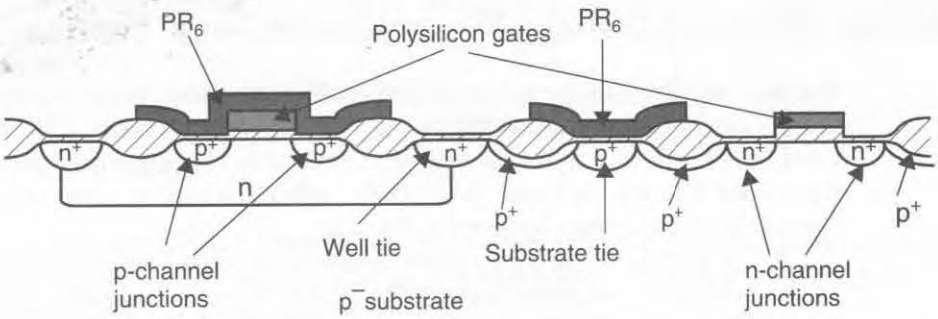


Fig. 2.10 Cross section after ion-implanting the n^+ junctions.

ately low temperatures of 500 °C or lower. The deposited SiO_2 might be 0.25 to 0.5 μm.

The next step is to open contact holes through the deposited SiO_2. The contact holes are defined using mask M_5 and positive resist PR_7.

Annealing, Depositing and Patterning Metal, and Overglass Deposition

After the first layer of CVD SiO_2 has been deposited, the wafer is annealed. As mentioned earlier in this section, annealing entails heating the wafer in an inert gas (such as nitrogen) for some period of time (say, 15 to 30 minutes) at temperatures up to 1,000 °C. The resulting thermal vibrations heal the lattice damage sustained during all the ion implantations, broaden the concentration profiles of the implanted dopants, and increase the density of the deposited SiO_2.

Next, interconnect metal is deposited everywhere. Historically, aluminum (Al) has been used for the interconnect. However, recently other metals have been used that have less of a tendency to diffuse into the silicon during electrical operation of the microcircuit. The metal is deposited using evaporation techniques in a vacuum. The heat required for evaporation is normally produced by using electron-beam bombarding, or possibly ion bombarding in a sputtering system. After the metal is deposited on the entire wafer, it is patterned using mask M_6 and positive photoresist PR_8, and then it is etched.

At this time, a low-temperature annealing might take place to give better bonds between the metal and the silicon. The temperature of this annealing must be less than 550 °C so the aluminum doesn't melt.

Next, an additional layer of CVD SiO_2 is deposited, additional contact holes are formed using mask M_7 and photoresist PR_9, and then a second layer of metal is deposited and etched using mask M_8 and photoresist PR_{10}. Often the primary use of this top layer of metal might be to distribute the power supply voltages. The bottom layer would be used more often for local interconnects in gates. In some modern processes, this process is repeated a third and possibly a fourth time to give up to four levels of metal, which allow for much denser interconnect.

After the last level of metal is deposited, a final *passivation*, or *overglass*, is deposited for protection. This layer would be CVD SiO_2, although often an additional layer of Si_3N_4 might be deposited because it is more impervious to moisture.

The final microcircuit processing step is to etch openings to the pads used for wire bonding. This final step would use mask M_9 and photoresist PR_{11}. A cross section of the final microcircuit for our example process is shown in Fig. 2.11.

This example process is a fairly representative CMOS process. However, many variations, often involving additional masks, are possible. Some of the possible variations are as follows:

1. Two wells may exist—one for p-channel transistors and one for n-channel transistors. This *twin-tub process* allows both wells to be optimally doped.
2. An additional polysilicon layer may be deposited over the first layer. This extra poly layer can be used to realize highly linear poly-to-poly capacitors in which a thin thermal oxide is used to separate the two layers.

3. An additional polysilicon layer might be formed that has an extremely high resistivity (say, 1 GΩ/□). This high resistivity is used to realize resistor loads in four-transistor, static random-access memory (SRAM) cells.
4. Field-implants may exist under the field-oxide in the well regions as well as under the field-oxide in the substrate regions.
5. Often, the n-channel and the p-channel transistors might have separate threshold-voltage-adjust implants.
6. The microcircuit might have three, four, or five layers of metal.
7. In a multimetal-layer process, it is usually necessary to add several additional steps so that the surface is made smoother, or *planarized,* after each metal-patterning step. This is normally done by a reactive etching process in which the metal is covered with SiO_2 and the hills are etched faster than the valleys.
8. Different metals might be used for the contacts than for the interconnect, to obtain better fill in and less diffusion into the silicon surface.
9. Thin-film nichrome resistors may exist under the top layer of metal.
10. The transistors may be realized in an epitaxial layer. In this case, the substrate would be n^-, and a p^--epitaxial layer would be grown. Before the epitaxial layer is grown, the top of the substrate would be doped p^+. This type of wafer is similar to that used in processing bipolar transistors, but of the opposite type, and is becoming more common in CMOS processing. The advantages of this type of transistor are that it is more immune to a destructive phenomenon called latch-up (described at the end of this chapter), and it is also more immune to gamma radiation in space. Finally, it greatly minimizes substrate noise in microcircuits that have both analog and digital circuits, (i.e., mixed-mode microcircuits).
11. Additional processing steps may be used to ensure that bipolar transistors can be included in the same microcircuit as MOS transistors. This type of process is called a BiCMOS process and is becoming particularly popular for both analog and digital high-speed microcircuits.

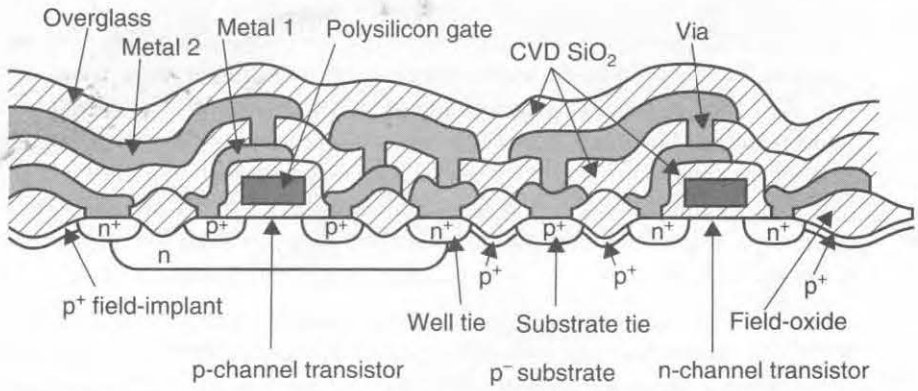


Fig. 2.11 Final cross section of an example CMOS microcircuit.

2.2 BIPOLAR PROCESSING

The processing steps required for realizing bipolar transistors are similar to those used for realizing MOS transistors, with some modifications. Thus, rather than presenting the complete realization of modern bipolar transistors, we briefly discuss some of the modifications needed for realizing them.

A bipolar process normally starts with a p^- substrate. The first masking step involves the diffusion (or ion implantation) of n^+ regions into the substrate wherever transistors are desired. These n^+ regions are used to lower the series collector resistance. Next, an n^- single-crystal epitaxial layer is deposited.

In the next basic step, the field-oxide is formed for isolation. However, before the field-oxide is grown and after the openings in the Si_3N_4-SiO_2-photoresist sandwich have been made, the surface of the silicon is typically etched to form empty cavities. This extra etching step allows the field-oxide to extend further down into the epitaxial region.

After the field oxide is grown, the n^+-collector contact region is implanted. This region extends from the surface down to the n^+-buried region under the transistor.

In a modern process, polysilicon is used to contact the emitter, the base, and possibly the collector, as Fig. 2.12 shows for a typical npn BJT structure. Normally, the base p^+ polysilicon is deposited first. This polysilicon is heavily doped p^+ so that later, during a high-temperature step, the boron dopant from the polysilicon contact diffuses into the silicon underneath the base polysilicon to make the underlying region p^+. The base polysilicon is removed in the active area of the transistor. Next, using one of a variety of possible methods, the base polysilicon is covered with a thin layer of

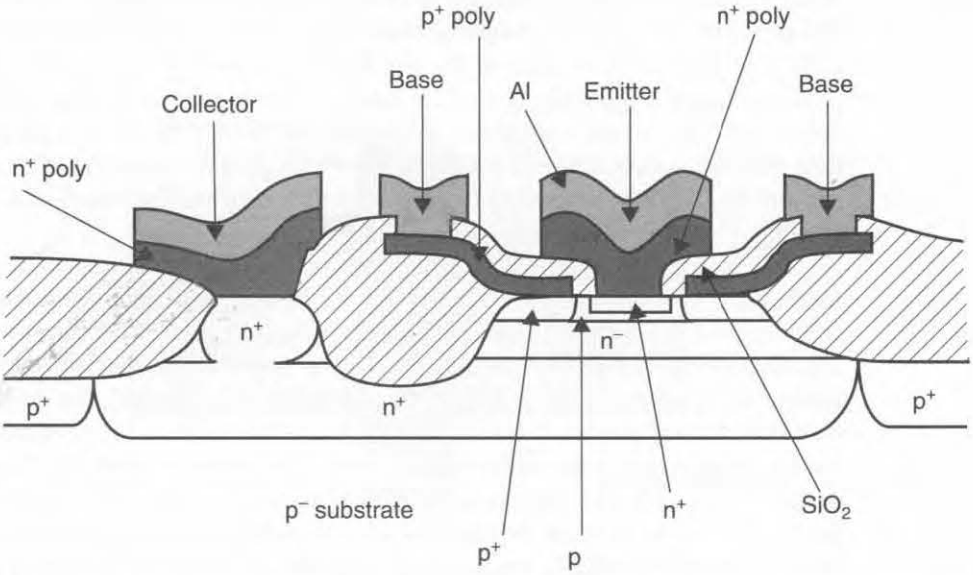


Fig. 2.12 Cross section of a modern, self-aligned bipolar transistor with oxide isolation. The term "poly" refers to polysilicon.

SiO_2, perhaps 0.5 μm in thickness, leaving an opening over the active area of the transistor. This SiO_2 *spacer*, which separates the base polysilicon from the emitter polysilicon, as shown in Fig. 2.12, allows the base polysilicon (or contact) to be very close to the emitter polysilicon (or contact), thereby minimizing the base resistance. Next, the base is ion-implanted to p -type silicon, and then n^+ polysilicon is deposited for the emitter. At this point, the true emitter has not yet been formed—only the emitter n^+ polysilicon has been laid down. However, when the wafer is annealed, the n^+ from the emitter polysilicon diffuses into the base p silicon to form the true emitter region. During annealing, the p^+ dopants from the base polysilicon also diffuse into the extrinsic base region. As a result, this procedure results in a self-aligned process since the use of the SiO_2 spacer allows the base polysilicon to determine where the emitter is finally located. The importance of this process is that, through the use of self-aligned contacts and field-oxide isolation, very small, high-frequency bipolar transistors can be realized using methods similar to those used in realizing modern MOS transistors.

2.3 CMOS LAYOUT AND DESIGN RULES

It is the designer's responsibility to determine the geometry of the various masks required during processing. The process of defining the geometry of these masks is known as *layout* and is done using a computer and a CAD program. Here, we describe some typical layout design rules and the reasons for these rules.

When designing the layout, typically the designer does not need to produce the geometry for all of the masks because some of the masks are automatically produced by the layout program. For example, the p^+ and n^+ masks used for the source and drain regions are usually generated automatically. Also, the program might allow the designer to work in the final desired dimensions. The layout program then automatically sizes the masks to account for any lateral diffusion or etching loss; this sizing produces larger- or smaller-dimension masks. For example, a designer might draw a polysilicon line so that a transistor would have a 1.2-μm length. The program might then produce a mask that had a 1.6-μm line width. This increased mask sizing would account for the junction overlap due to lateral diffusion and the polysilicon loss due to etching.

In a modern layout program, the layout of some circuit *cells* might already be performed and stored in a library. During overall layout, these cells are then parametrically adapted to a required size, and the corresponding geometries for every layer are automatically generated. Often, when the cells are being connected, they might be automatically *placed and routed*, or connected, by the program. The designer might then interactively modify this automatically generated layout. Thus, as time goes on, the layout becomes more automated as more cells become available. However, the designer must still take direct control of the layout of critical cells, especially when the layout must be small or the resulting circuits must be fast. For example, one would rarely allow a computer to automatically generate the layout of a memory cell where space and capacitive loading of the connecting buses are critical. Thus, a digital microcircuit designer must be knowledgeable about the design rules that govern the layout required for the process used.

The two most important masks are those for the active region and for the gate polysilicon. The intersection of these two masks becomes the channel region of MOS transistors. For example, consider Fig. 2.13(*a*), which shows a simplified view of a MOS transistor, and Fig. 2.13(*b*), which shows the corresponding layout of the active mask and the polysilicon, or poly, mask. In Fig. 2.13(*b*), the poly mask runs vertically. The length of the poly that intersects the active-region mask is the transistor width, W, and the width of the poly line is the transistor length, L, as Fig. 2.13 shows.

The design rules for laying out transistors are often expressed in terms of a quantity, λ, where λ is 1/2 the gate length. This generalization allows many of the design rules to be simply expressed, independent of the true value for the minimum channel length (i.e., 2λ). Figure 2.13(*b*) shows the smallest possible transistor that can be realized when a contact must be made to each junction. Also shown are many of the minimum dimensions in terms of λ.

When we express design rules in terms of λ, we assume that each mask has a worst-case alignment of under 0.75λ. Thus, we can guarantee that the relative misalignment between any two masks is under 1.5λ. If an overlap between any two regions of a microcircuit would cause a destructive short circuit, then a separation

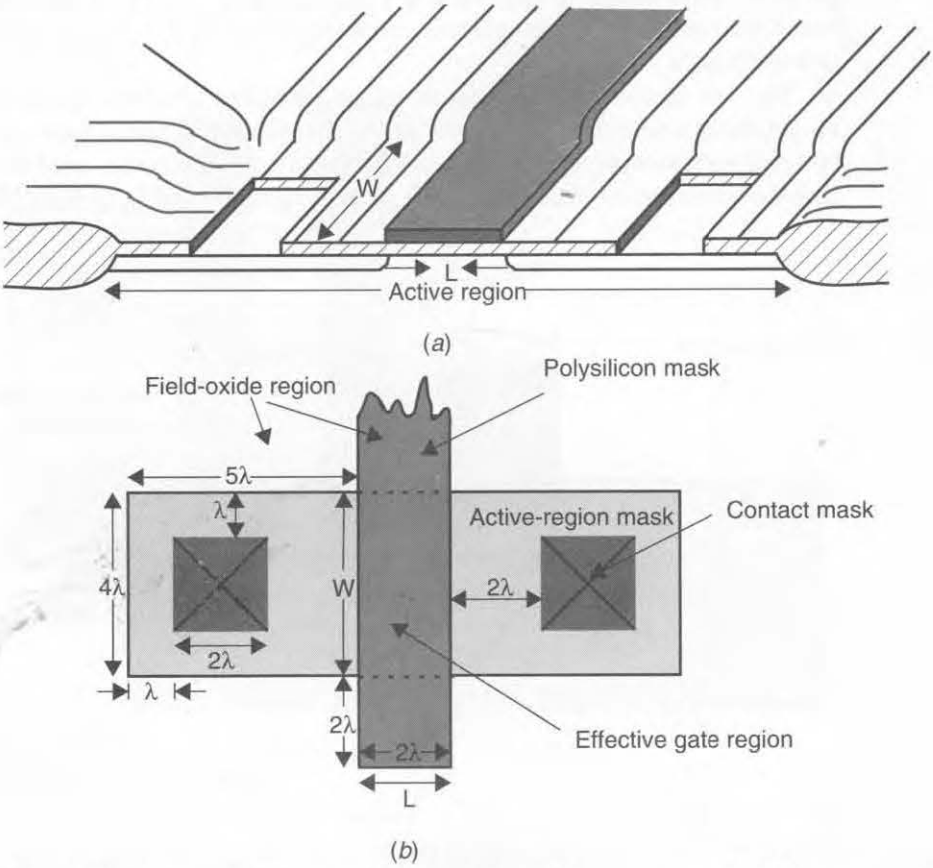


Fig. 2.13 (*a*) A simplified view of a partially finished transistor and (*b*) the corresponding layout of the active, polysilicon, and contact masks.

between the corresponding regions in a layout of 2λ guarantees this will never happen. For example, consider the poly mask and the contact mask in Fig. 2.13(*b*). If these two regions overlap in the microcircuit, then the metal used to contact the source junction is also short-circuited to the gate poly, causing the transistor to be always turned off, as shown in Fig. 2.14. If the source happens to be connected to ground, this error also short-circuits the gate-to-ground. To prevent this type of short, the contact openings must be kept at least 2λ away from the polysilicon gates.

Another example of a catastrophic failure due to misalignment is a gate that does not fully cross the active region (also shown in Fig. 2.14). Since the junctions are implanted everywhere in the active region except under the gate, this misalignment causes a short circuit between the source and the drain—thus the design rule that polysilicon must always extend at least 2λ past the active region.

Another design rule is that active regions should surround contacts by at least 1λ. If, in reality, an overlap exists between the edge of the active-region mask and the contact mask, no disastrous shorts occur. The circuit still works correctly as long as sufficient overlap exists between the contact and the active masks so that a good connection is made between the aluminum interconnect and the junction. Since the maximum relative misalignment is 1.5λ, having the source (or drain) region surround the contact by 1λ guarantees an overlap of at least 1.5λ (the minimum contact width is 2λ).

The few design rules just described are sufficient to allow one to estimate the minimum dimensions of a junction area and perimeter before a transistor has been laid out. For example, assume that in Fig. 2.13 a contact is to be made to a junction; then the active region must extend past the polysilicon region by at least 5λ. Thus, the

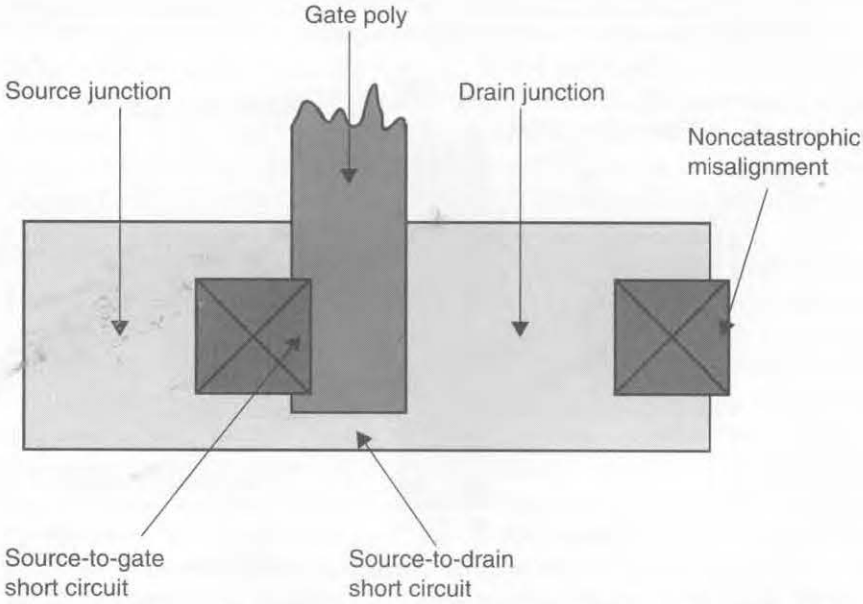


Fig. 2.14 Mask misalignment that results in catastrophic short circuits and an example of a noncatastrophic misalignment.

minimum area of a small junction with a contact to it is

$$A_s = A_d = 5\lambda W \tag{2.4}$$

where W is the transistor width. Similarly, in Fig. 2.13, the perimeter of a junction[3] with a contact is given by

$$P_s = P_d = 10\lambda + W \tag{2.5}$$

These estimates may be used when estimating the parasitic capacitances in the transistor models. They may also be used in SPICE to simulate circuits so the parasitic capacitances are determined more accurately. However, note that they are only estimates; the true layout will differ somewhat from these rough estimates.

Sometimes, when it is important to minimize the capacitance of a junction, a single junction can be shared between two transistors. For example, consider the series connection of two transistors shown in Fig. 2.15(*a*). The active, poly, and contact

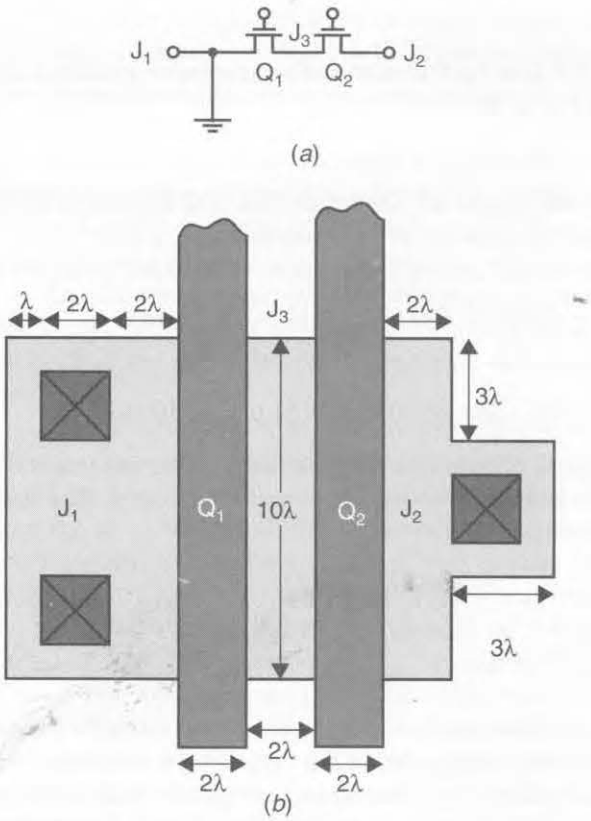


Fig. 2.15 (*a*) A series connection of two transistors and (*b*) a possible layout.

3. Note that the perimeter does not include the edge between the junction and the active channel separating the junction and the gate because there is no field-implant along this edge, and the sidewall capacitance is therefore smaller along that edge.

masks might be laid out as shown in Fig. 2.15(*b*). Notice that a single junction is shared between transistors Q_1 and Q_2. The area, and especially the perimeter of this junction, are much smaller than those given by equations (2.4) and (2.5). Also, in a SPICE simulation, the area and perimeter should be divided by 2 when they are specified in each transistor description, since the junction is shared. Alternatively, all of the area and perimeter could be specified in one transistor description, and the area and perimeter of the other junction could be specified as zero.

Since the junction sidewall capacitance is directly proportional to the junction perimeter, and since this capacitance can be a major part of the total junction capacitance (because of the heavily doped field-implants), minimizing the perimeter is important. It is of interest to note that as transistor dimensions shrink, the ratio of the perimeter to the area increases and the sidewall capacitance becomes more important.

EXAMPLE 2.1

Assuming $\lambda = 0.5\ \mu m$, find the area and perimeters of junctions J_1, J_2, and J_3 for the circuit in Fig. 2.15.

Solution

Since the width and length are shown as 10λ and 2λ, respectively, and $\lambda = 0.5\ \mu m$, the physical sizes are $W = 5\ \mu m$ and $L = 1\ \mu m$.

Thus, for junction J_1, using the formulas of (2.4) and (2.5), we have

$$A_{J1} = 5\lambda W = 5(0.5)5(\mu m)^2 = 12.5(\mu m)^2 \tag{2.6}$$

and

$$P_{J1} = 10\lambda + W = [10(0.5) + 5]\ \mu m = 10\ \mu m \tag{2.7}$$

Since this junction is connected to ground, its parasitic capacitance is unimportant and little has been done to minimize its area. Contrast this case with junction J_2, where we have

$$\begin{aligned} A_{J2} &= 2\lambda W + 12\lambda^2 \\ &= W + 12(0.5)^2 \\ &= 8(\mu m)^2 \end{aligned} \tag{2.8}$$

The perimeter is unchanged, resulting in $P_{J2} = 10\ \mu m$. Thus, we have decreased the junction area by using the fact that the transistor is much wider than the single contact used. However, sometimes wide transistors require additional contacts to minimize the contact impedance. For example, the two contacts used for junction J_1 result in roughly half the contact impedance of junction J_2.

Next, consider the shared junction. Here we have a junction area given by

$$\begin{aligned} A_{J3} &= 2\lambda W \\ &= 5(\mu m)^2 \end{aligned} \tag{2.9}$$

Since this is a shared junction, in a SPICE simulation we would use

$$\begin{aligned} A_s &= A_d = \lambda W \\ &= 2.5(\mu m)^2 \end{aligned} \tag{2.10}$$

for each of the two transistors, which is much less than 12.5 $(\mu m)^2$. The reduction in the perimeter is even more substantial. Here we have

$$\begin{aligned} P_{J3} &= 4\lambda \\ &= 2\ \mu m \end{aligned} \tag{2.11}$$

for the shared junction; so sharing this perimeter value over the two transistors would result in

$$\begin{aligned} P_S &= P_d = 2\lambda \\ &= 1\ \mu m \end{aligned} \tag{2.12}$$

for the appropriate junction of each transistor when simulating it in SPICE. This result is much less than the 10-μm perimeter for node J_1.

Because minimizing the junction capacitance is so important, one of the first steps an experienced designer takes before laying out important high-speed cells is first to identify the most critical nodes and then to investigate possible layouts that minimize the junction capacitance of these nodes.[4]

An additional design rule has been implicitly introduced in the previous example. Notice that for junction J_2 in Fig. 2.15, part of the active region boundary is only 2λ away from the gate. This minimum junction area is the typical design rule for this case.

Several design rules are required in addition to those just mentioned. Some of these are described next, with reference to the layout of a digital inverter, shown in Fig. 2.16. Notice that the n well surrounds the p-channel active region, and therefore the p^+ junctions of the p-channel transistors, by at least 3λ. Notice also that the minimum spacing between the n well and the junctions of n-channel transistors, in the substrate, is 5λ. This large spacing is required because of the large lateral diffusion of the n well and the fact that if the n-channel junction became short-circuited to the n well, which is connected to V_{DD}, the circuit would not work. Conversely, a p^+-substrate tie can be much closer to a well because it is always connected to ground and is separated from the well by a reverse-biased junction. A typical dimension here might be 2λ. Since a p-channel junction must be inside the well by at least 3λ and an n-channel junction must be outside the well by 5λ, the closest an n-channel transistor can be placed to a p-channel transistor is 8λ.

Notice in Fig. 2.16 that metal is used to connect the junctions of the p-channel and n-channel transistors. Normally, the metal must overlap any underlying contacts

4. Note that it is not possible to share junctions between n-channel and p-channel transistors. This limitation is one of the reasons for the larger parasitic capacitances sometimes encountered in CMOS microcircuits as opposed to NMOS microcircuits, where only n-channel transistors are used.

by at least λ. A typical minimum width for first-level metal might be 2λ, the same as the minimum width for polysilicon. However, it can be wider as in Fig. 2.16, where it is 4λ wide.

Notice also in Fig. 2.16 that a single contact opening, known as a *butting contact*, can be used to contact both the p-channel transistor source and an n^+-well tie, because both will be connected to V_{DD}. Although the outlines of the p^+ and n^+ masks are not shown in Fig. 2.16, under the contact, one half will be doped p^+ (the p-channel junction) and one half will be doped n^+ (the well tie). Also, for the n-channel transistor, a

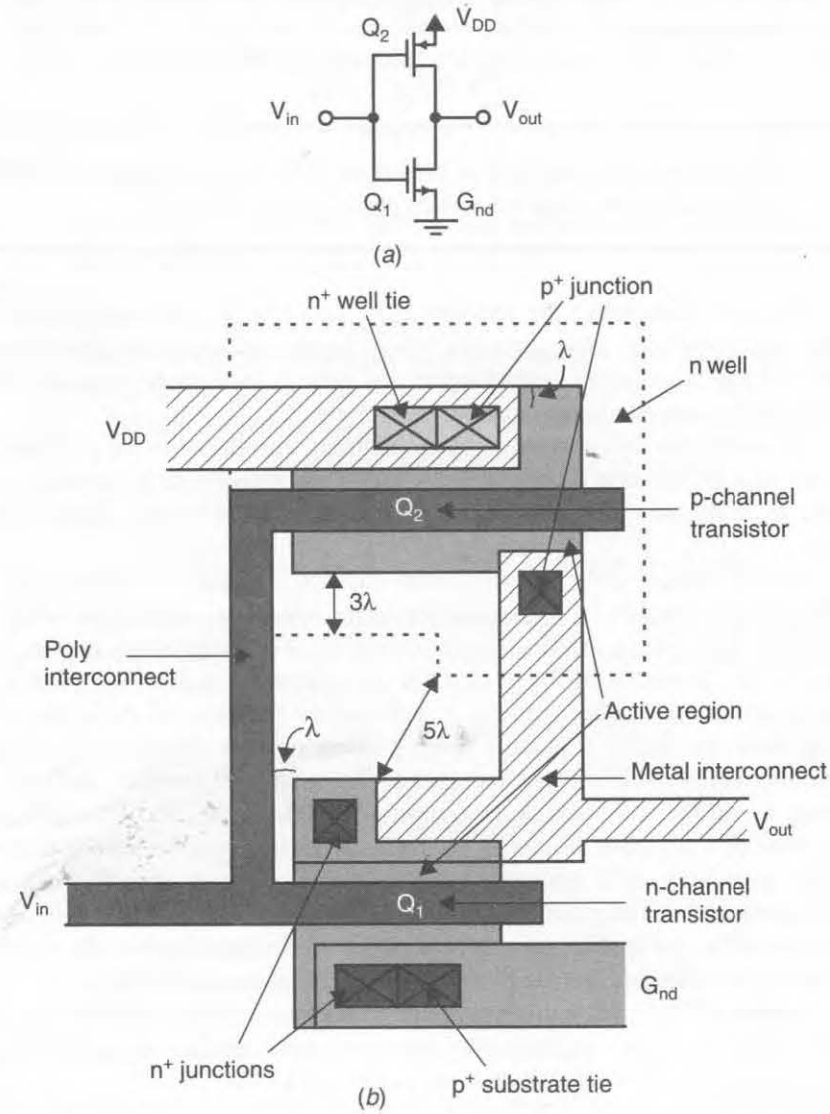


Fig. 2.16 (*a*) A CMOS digital inverter and (*b*) a possible layout with several design rules illustrated.

butting contact was used to connect the n-channel source to a p^+-substrate tie, and both will be connected to ground. In a typical set of design rules, a maximum distance between transistors and well (or substrate) ties is specified, and a maximum distance between substrate ties is also specified. For example, the rules might specify that no transistor can be more than 100λ from a substrate tie. These rules are necessary to prevent latch-up, a phenomenon described at the end of this chapter.

As a final example, we describe the layout of a large transistor. Normally, a wide transistor is composed of smaller transistors connected in parallel. A simplified layout of this approach is shown in Fig. 2.17(*a*), where four transistors that have a common gate are connected in parallel. Figure 2.17(*b*) shows the circuit corresponding to the layout in Fig. 2.17(*a*), where the transistors have been drawn in the same relative positions. Figure 2.17(*c*) shows the same circuit redrawn differently, where it is clear that the circuit consists of four transistors connected in parallel. Notice that the second and fourth junction regions are connected by metal to node 1, whereas the first, third, and fifth junction regions are connected by metal to realize node 2. Because it has a larger total junction area and especially a larger perimeter, node 2 will have a much greater junction capacitance than node 1. Thus, when the equivalent transistor is connected to a circuit, node 1 should be connected to the more critical node. Also notice the large number of contacts used to minimize the contact impedance. Normally, some of these are butting contacts if either node 1 or node 2 is connected to an appropriate power supply. The use of many contacts in wide junction regions greatly minimizes voltage drops that would otherwise occur due to the relatively high resistivity of silicon junctions compared to the resistivity of the metal that overlays the junctions and connects them.[5]

Design rules also specify the minimum pitch between polysilicon interconnects, metal 1 interconnects, and metal 2 interconnects. These might be 2λ, 2λ, and 3λ, respectively. Metal 2 requires a larger minimum pitch because it resides further from the silicon surface where the topography is less even. The minimum widths of poly, metal 1, and metal 2 might also be 2λ, 2λ, and 3λ, respectively.

This concludes our brief introduction to layout and design rules. In a modern process, many more design rules are used than those just described. However, the reasons for using and the methods of applying these rules is similar to that which has been described. Finally, note that when one does modern microcircuit layout, the design rules are usually available to the layout program and are automatically checked as layout progresses.

EXAMPLE 2.2

Consider the transistor shown in Fig. 2.17, where the total width of the five parallel transistors is 80λ, its length is 2λ, and $\lambda = 0.5\ \mu m$. Assuming node 2 is the source, node 1 is the drain, and the device is in the active region, find the

5. The use of metal to overlay a higher-resistivity interconnect, such as polysilicon or heavily-doped silicon, is recommended to lower the resistivity of the interconnect.

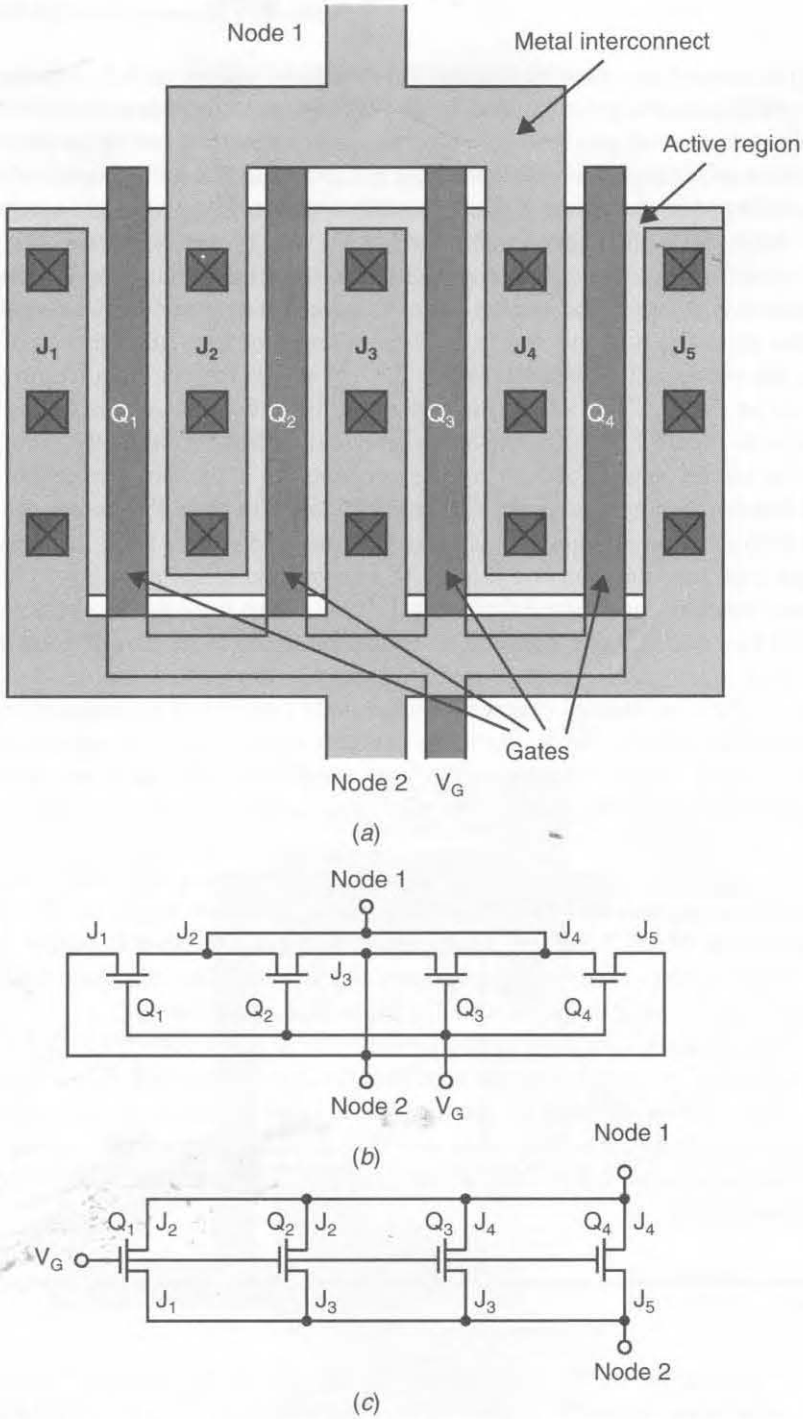


Fig. 2.17 Connecting four transistors in parallel to realize a single large transistor: (*a*) the layout, (*b*) the schematic drawn in the same relative positions as the layout, and (*c*) the circuit redrawn to make the parallel transistors more obvious.

source-bulk and drain-bulk capacitances given the parameters $C_j = 2.4 \times 10^{-4}$ pF/(μm)^2 and $C_{j\text{-}sw} = 2.0 \times 10^{-4}$ pF/μm. Also find the equivalent capacitances if the transistor were realized as a single device with source and drain contacts still evenly placed.

Solution

Starting with node 1, the drain, we find that the areas of the junctions are equal to

$$A_{J2} = A_{J4} = 6\lambda \times 20\lambda = 120\lambda^2 = 30\ (\mu m)^2$$

Ignoring the gate side, the perimeters are given by

$$P_{J2} = P_{J4} = 6\lambda + 6\lambda = 12\lambda = 6\ \mu m$$

As a result, C_{db} can be estimated to be

$$C_{db} = 2(A_{J2}C_j + P_{J2}C_{j\text{-}sw}) = 0.017\ pF$$

For node 2, the source, we have

$$A_{J1} = A_{J5} = 5\lambda \times 20\lambda = 100\lambda^2 = 25\ (\mu m)^2$$

and

$$A_{J3} = A_{J2} = 30\ (\mu m)^2$$

The perimeters are found to be

$$P_{J1} = P_{J5} = 5\lambda + 5\lambda + 20\lambda = 30\lambda = 15\ \mu m$$

and

$$P_{J3} = P_{J2} = 6\ \mu m$$

resulting in an estimate for C_{sb} of

$$C_{sb} = (A_{J1} + A_{J3} + A_{J5} + WL)C_j + (P_{J1} + P_{J3} + P_{J5})C_{j\text{-}sw} = 0.036\ pF$$

It should be noted that, even without the additional capacitance due to the WL gate area, node 1 has less capacitance than node 2 since it has less area and perimeter.

In the case where the transistor is a single wide device, rather than four transistors in parallel, we find

$$A_J = 5\lambda \times 80\lambda = 400\lambda^2 = 100\ (\mu m)^2$$

and

$$P_J = 5\lambda + 5\lambda + 80\lambda = 90\lambda = 45\ \mu m$$

resulting in $C_{db} = 0.033$ pF and $C_{sb} = 0.043$ pF. Note that in this case, C_{db} is nearly twice what it is when four parallel transistors are used.

2.4 ANALOG LAYOUT CONSIDERATIONS

When one designs analog circuits, several important layout issues should be considered to realize high-quality circuits. These issues can be broadly divided into two categories—matching and noise issues.

Matching Issues

When integrated components are realized using lithographic techniques, a variety of two-dimensional effects can cause the effective sizes of the components to differ from the sizes of the glass layout masks. Some examples of these effects are illustrated in Fig. 2.18. For example, Fig. 2.18(*a*) shows how an effective well area will typically be larger than its mask due to the lateral diffusion that occurs not just during ion implantation but also during later high-temperature steps, such as annealing. Another effect, known as *overetching*, occurs when layers such as polysilicon or metal are being etched. Figure 2.18(*b*), for example, shows overetching that occurs under the SiO_2 protective layer at the polysilicon edges and causes the polysilicon layer to be smaller than the corresponding mask layout. A third effect is shown in Fig. 2.18(*c*), where an n-channel transistor is shown as we look along the channel from the drain to the source. The width of the transistor is defined by the width of the active region (as opposed to the width of the polysilicon line), and this width is determined by the field-oxide. The p^+ field implant under the field-oxide causes the effective substrate doping to be greater at the sides of the transistors than elsewhere. This increased doping raises the effective transistor threshold voltage near the sides of the transistors and therefore decreases the channel-charge density at the edges. The result is that the effective width of the transistor is less than the width drawn on the layout mask. The examples shown illustrate typical sizing effects, but many other second-order effects influence the realized compo-

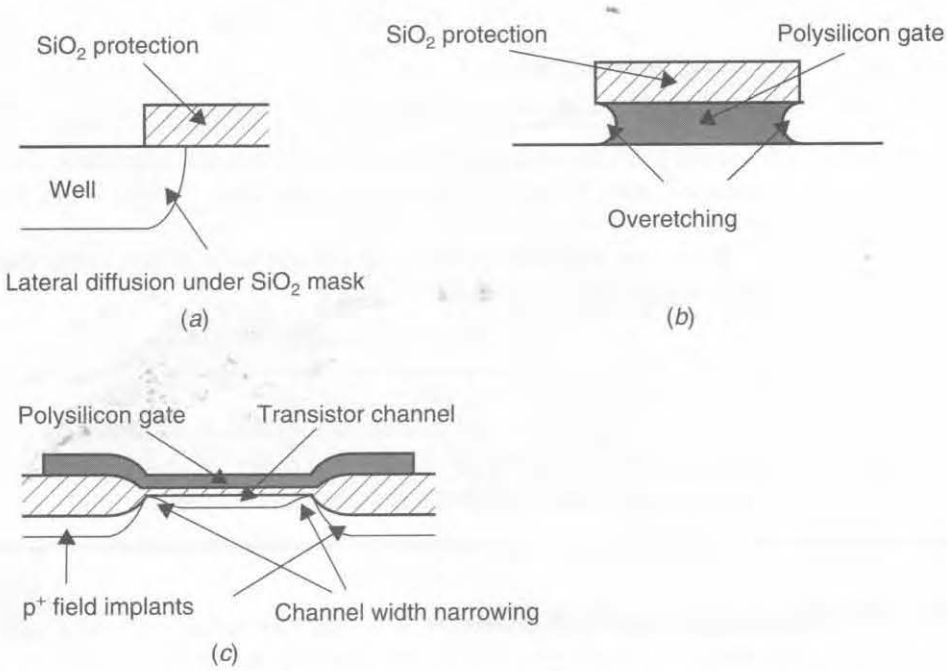


Fig. 2.18 Various two-dimensional effects causing sizes of realized microcircuit components to differ from sizes of layout masks.

nent sizes. These other effects include those caused by boundary conditions of an object, the size of the opening in a protective layout through which etching occurs, and the unevenness of the surface of the microcircuit [Maloberti, 1994]. For these reasons, the absolute sizes of microcircuit components can seldom be accurately determined. These inaccuracies also affect the ratios of sizes, although to a lesser degree, when the ratio is not unity (unless the second-order size error effects are matched). *Matching second-order size error effects is done mainly by making larger objects out of several unit-sized components connected together. Also, for best accuracy, the boundary conditions around all objects should be matched, even when this means adding extra unused components.* Next, we give some examples of how to apply these principles with regard to realizing transistors, capacitors, and resistors.

Transistor Layouts

Transistors in analog circuits are typically much wider than transistors in digital circuits. For this reason, they are commonly laid out using multiple-gate fingers similar to the layout shown in Fig. 2.17. When precision matching between transistors is required, then not only should the individual transistors be realized by combining a single-sized unit transistor, but the fingers for one transistor should be interdigitated with the fingers of the second transistor. This approach, known as *common-centroid layout,* helps match errors caused by gradient effects across a microcircuit, such as the temperature or the gate-oxide thickness changing across the microcircuit. An example of a common-centroid layout of two identical matched transistors whose sources are connected is shown in Fig. 2.19 in simplified form [O'Leary, 1991; Maloberti, 1994].

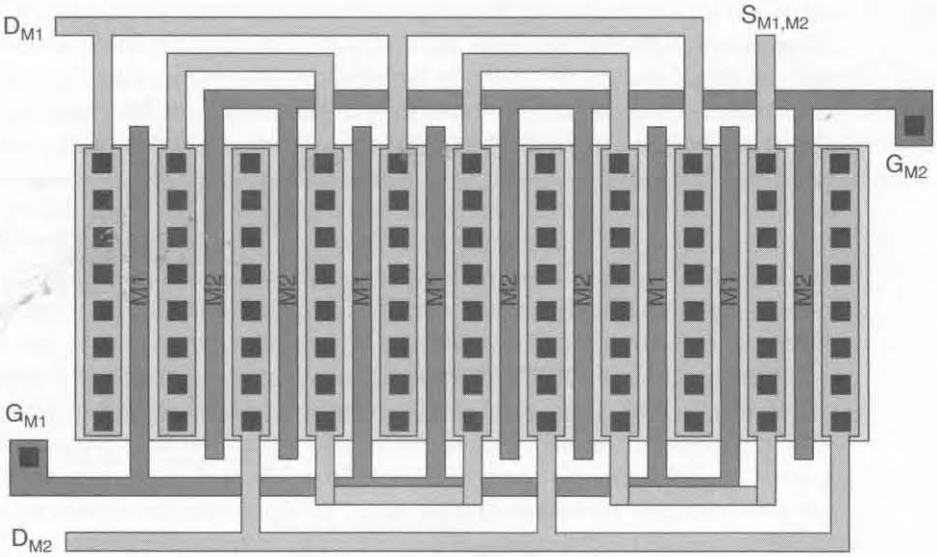


Fig. 2.19 A common-centroid layout for a differential source-coupled pair.

Each of the two transistors is composed of five separate transistor fingers connected in parallel. The outside fingers have separate second-order size effects, and therefore one outside finger is used for M_1 and one is used for M_2. Inside the structure, the fingers occur in doubles—two for M_2, two for M_1, two for M_2, and so on. Thus, the layout is symmetric in both the x and y axes, and any gradients across the microcircuit would affect both M_1 and M_2 in the same way. This layout technique greatly minimizes nonidealities such as opamp input-offset voltage errors when using a differential pair in the input stage of an opamp.

When current mirrors with ratios other than unity are required, again, each of the individual transistors should be realized from a single unit-sized transistor. For example, if a current ratio of 1:2 were desired, then the input transistor might be made from four fingers, whereas the output transistor might be realized using eight identical fingers. In addition, for the greatest accuracy, all fingers should be inside fingers only. Outside, or dummy, fingers would only be included for better matching accuracy and would have no other function. The gates of these dummy fingers are normally connected to the most negative power-supply voltage to ensure they are always turned off (or they are connected to the positive power supply in the case of p-channel transistors).

Capacitor Matching

Very often, analog circuits require precise ratios of capacitors. Ideally, capacitor size is given by

$$C_1 = \frac{\varepsilon_{ox}}{t_{ox}} A_1 = C_{ox} x_1 y_1 \tag{2.13}$$

The major sources of errors in realizing capacitors are due to overetching (which causes the area to be smaller than the area of the layout masks) and an oxide-thickness gradient across the surface of the microcircuit. The former effect is usually dominant and can be minimized by realizing larger capacitors from a parallel combination of smaller, unit-sized capacitors, similar to what is usually done for transistors. For example, to realize two capacitors that have a ratio of 4:6, the first capacitor might be realized from four unit-sized capacitors, whereas the second capacitor might be realized by six unit-sized capacitors. Errors due to the gradient of the oxide thickness can then be minimized by interspersing the unit-sized capacitors in a common-centroid layout so the gradient changes affect both capacitors in the same way. Since oxide-thickness variations are not usually large in a reasonably small area, this common-centroid layout is not typically used except where very accurate capacitors are required.

If only unit-sized capacitors are used, then any overetching will leave the capacitor ratio unaffected. Thus, good designers strive to realize circuits in which only unit-sized capacitors are needed. Unfortunately, this situation is not always possible. When it is not, *overetching error can still be minimized by realizing a nonunit-sized capacitor with a specific perimeter-to-area ratio.* To determine the correct ratio, first note that the error due to overetching is roughly proportional to the perimeter of the capacitor. Specifically, if we assume that a capacitor has an absolute overetching given by

Δe and that its ideal dimensions are given by x_1 and y_1, then its true dimensions are given by $x_{1a} = x_1 - 2\Delta e$ and $y_{1a} = y_1 - 2\Delta e$, and the true capacitor size is given by

$$C_a = C_{ox}x_{1a}y_{1a} = C_{ox}(x_1 - 2\Delta e)(y_1 - 2\Delta e) \tag{2.14}$$

This situation is illustrated in Fig. 2.20. Thus, the error in the true capacitance is given by

$$\Delta C_t = C_{ox}x_{1a}y_{1a} - C_{ox}x_1y_1 = C_{ox}[-2\Delta e(x_1 + y_1) + 4\Delta e^2] \tag{2.15}$$

When this error is small, then the second-order error term can be ignored and (2.15) can be approximated by

$$\Delta C_t \cong -2\Delta e(x_1 + y_1)C_{ox} \tag{2.16}$$

The relative error in the capacitor is therefore given by

$$\varepsilon_r = \frac{\Delta C_t}{C_{ideal}} \cong \frac{-2\Delta e(x_1 + y_1)}{x_1y_1} \tag{2.17}$$

Thus, the relative capacitor error is approximately proportional to the negative of the ratio of the ideal perimeter to the ideal area (assuming only small errors exist, which is reasonable since, if the errors were not small, then that capacitor sizing would probably not be used). When we realize two capacitors that have different sizes, usually the ratio of one capacitor to the other is important, rather than their absolute sizes. This ratio is given by

$$\frac{C_{1a}}{C_{2a}} = \frac{C_1(1 + \varepsilon_{r1})}{C_2(1 + \varepsilon_{r2})} \tag{2.18}$$

If the two capacitors have the same relative errors (i.e., $\varepsilon_{r1} = \varepsilon_{r2}$), then their true ratio is equal to their ideal ratio even when they are not the same sizes. Using (2.17), we see that the relative errors are the same if they both have the same perimeter-to-area ratio. This leads to the following result: *To minimize errors in capacitor ratios due to overetching, their perimeter-to-area ratios should be kept the same, even when the capacitors are different sizes.*

Normally, the unit-sized capacitor will be taken square. *When a non-unit-sized capacitor is required, it is usually set to between one and two times the unit-sized capacitor and is rectangular in shape, so that it has the same number of corners as*

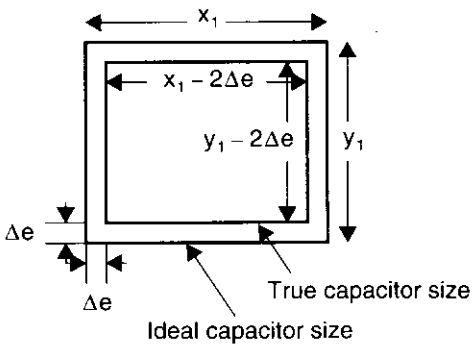


Fig. 2.20 Capacitor errors due to overetching.

the unit-sized capacitor. Defining K to be a desired non-unit-sized capacitor ratio, we have

$$K \equiv \frac{C_2}{C_1} = \frac{A_2}{A_1} = \frac{x_2 y_2}{x_1^2} \tag{2.19}$$

where C_1, A_1, and x_1 represent the capacitance, area, and side-length of a unit-sized capacitor, respectively. Variables C_2, A_2, x_2, and y_2 are similarly defined, except this non-unit-sized capacitor is now rectangular. Equating the ratios of the perimeters-to-areas implies that

$$\frac{P_2}{A_2} = \frac{P_1}{A_1} \tag{2.20}$$

where P_1 and P_2 represent the perimeters of the two capacitors. Rearranging (2.20), we have

$$\frac{P_2}{P_1} = \frac{A_2}{A_1} = K \tag{2.21}$$

which implies that K can also be written as the ratio of perimeters,

$$K = \frac{x_2 + y_2}{2x_1} \tag{2.22}$$

This can be rearranged to become

$$x_2 + y_2 = 2Kx_1 \tag{2.23}$$

Also rearranging (2.19), we have

$$x_2 = \frac{Kx_1^2}{y_2} \tag{2.24}$$

Combining (2.23) and (2.24), we find the quadratic equation

$$y_2^2 - 2Kx_1y_2 + Kx_1^2 = 0 \tag{2.25}$$

which can be solved to give

$$y_2 = x_1(K \pm \sqrt{K^2 - K}) \tag{2.26}$$

Recall that K is assumed to be greater than one, which ensures that the square root in (2.26) is applied to a positive number. The value for x_2 is then given by (2.24).

EXAMPLE 2.3

Show a layout that might be used to match two capacitors of size 4 and 2.314 units, where a unit-sized capacitor is $10\ \mu m \times 10\ \mu m$.

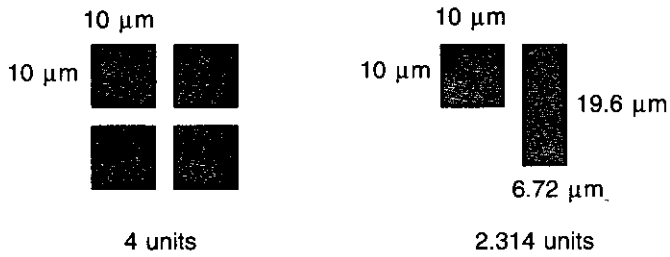


Fig. 2.21 A capacitor layout with equal perimeter-to-area ratios of 4 units and 2.314 units.

Solution

Four units are simply laid out as four unit-sized capacitors in parallel. We break the 2.314-unit capacitor up into one unit-sized capacitor in parallel with another rectangular capacitor of size 1.314 units. The lengths of the sides for this rectangular capacitor are found from (2.26), resulting in

$$y_2 = 10\,\mu\text{m}\left(1.314 \pm \sqrt{1.314^2 - 1.314}\right) = 19.56\,\mu\text{m or } 6.717\,\mu\text{m}$$

Either of these results can be chosen for y_2, and the other result becomes x_2; in other words, the choice of sign affects only the rectangle orientation. Thus, we have the capacitor layout as shown in Fig. 2.21 Note that the ratio of the area of the rectangular capacitor to its perimeter equals 2.5, which is the same as the ratio for the unit-sized capacitor.

Several other considerations should be followed when realizing accurate capacitor ratios. Usually the bottom plate of capacitors will be the first layer of polysilicon or, in some technologies, an ion-implanted region. This region is usually common to many unit-size capacitors.[6] The interconnection of the top plates, which are almost always polysilicon, can often be done in first-level metal with contacts to the polysilicon plates. In some technologies this is not possible, and the polysilicon plates must be interconnected through polysilicon tabs at the sides. These tabs will contact to metal in a region where the bottom-plate polysilicon is not present. The parasitic capacitances of these tabs should be matched as much as possible. This matching often entails adding additional tabs that are not connected anywhere. Another common matching technique is to ensure that the boundary conditions around the unit-sized capacitors match. This boundary-condition matching is accomplished by adding additional top-plate polysilicon around the outside boundaries of unit-sized capacitors at the edge of an array. Many of these principles are illustrated in the simplified layout of two capacitors shown in Fig. 2.22. Each capacitor in the

6. It is usually possible to realize the bottom plate over the top of a well region that has many contacts connected to a low-noise power-supply voltage. This well region acts as a shield to help keep substrate noise out of the bottom plate of the capacitor.

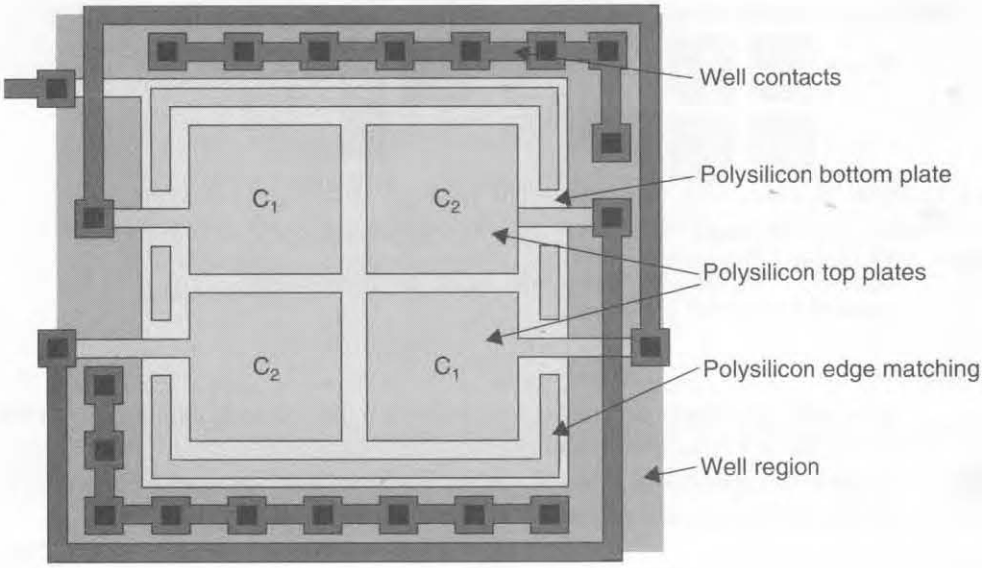


Fig. 2.22 A simplified layout of a capacitor array.

figure consists of two unit-sized capacitors. An additional technique sometimes used (not shown in Fig. 2.22) is to cover the top plates with the first layer of metal, which is then connected to the bottom plate. This sandwich-like structure not only gives additional capacitance per area but more importantly, shields the top level of polysilicon from electromagnetic interference in the air. The capacitor plate formed from the "sandwiched" top level of polysilicon is then connected to critical nodes, such as the virtual inputs of amplifiers.

An important consideration when using capacitors in switched-capacitor circuits is that the bottom plate usually has more noise coupled into it than the top plate because of the large parasitic capacitance between the bottom plate and the silicon region. Therefore the top plate should be connected to critical nodes such as the virtual inputs of opamps, whereas the bottom plate can be connected to less critical nodes such as opamp outputs.

For more details concerning the realization of integrated capacitors, the interested reader can see [Allstot, 1983; O'Leary, 1991; Maloberti, 1994].

Resistor Layout

Integrated resistors can be realized using a wide variety of different conductors. A popular choice is polysilicon, which is a deposited and etched material. Other choices include diffused or ion-implanted regions such as junctions, wells, or base regions. Another possibility is deposited and etched thin-film resistors such as nichrome (consisting of 80 percent nickel and 20 percent chromium) or tantalum. The temperature coefficient of ion-implanted or diffused resistors tends to be positive and large (espe-

cially for larger resistivities) with values as large as 1,000 to 3,000 ppm/°C. On the other hand, the temperature coefficient for thin-film resistors can be as small as 100 ppm/°C. Polysilicon resistors usually have large positive temperature coefficients (say, 1,000 ppm/°C) for low-resistivity polysilicon ranging to moderately large, negative temperature coefficients for specially doped, high-resistivity polysilicon. The positive temperature coefficients are primarily due to mobility degradation that results from temperature increases. In implanted and diffused resistors, nonlinear resistance varies greatly with voltage because the depletion-region width is dependent on voltage in the more heavily doped conductive region. This depletion-region width variation is substantially smaller in a polysilicon resistor, which is one of the major reasons polysilicon resistors are preferred over implanted resistors even though they often require more area due to the low resistivity. When thin-film resistors are available in a particular technology, they are almost always the preferred type—unfortunately, they are seldom available.

Regardless of the type of resistor used, the equations governing the resistance (see the Appendix of Chapter 1) are given by

$$R_{\square} = \frac{\rho}{t} \tag{2.27}$$

where $R_{\square}$ is the resistance per square, $\rho = 1/(q\mu_n N_D)$ is the resistivity,[7] t is the thickness of the conductor, and N_D is the concentration of carriers, which we assume are electrons. The total resistance is then given by

$$R = \frac{L}{W} R_{\square} \tag{2.28}$$

where L is the length of the resistor and W is the width of the resistor.

Often, the typical resistivity encountered in integrated circuits is small. For example, the typical resistivity of the polysilicon layer used to form transistor gates is about 20 Ω/□. Polysilicon is the layer most commonly used to form resistors inside microcircuits.[8] To obtain medium-sized resistors, one must usually use a serpentine layout similar to that shown in Fig. 2.23. When calculating the resistance of such a structure, one must make allowance for the bends and for the end contacts. For example, the contact structure chosen contributes 0.14 squares [Reinhard, 1987; Grebene, 1984], and each bend contributes 2.11 squares[9] [Glasser, 1977]. For the structure shown, each finger contributes 10 squares each for a total of 70 squares, the two contacts contribute 0.28 squares, and the six bends contribute 12.66 squares. Thus, the total is 82.94 squares. If the resistivity of the resistor is 20 Ω/□, then the total resistance is

$$R = 82.94\ \square \times 20\ \Omega/\square = 1.659\ \text{k}\Omega \tag{2.29}$$

7. This equation is valid for resistors with electron carriers. For resistors with holes as carriers, $\rho = 1/(q\mu_p N_A)$.
8. In a modern process, where the gates are formed from a sandwich of a refractory metal over polysilicon, i.e., a polycide, or silicide, then the resistivity is much smaller, perhaps 1–5 Ω/□, and the gate layer is not useful for realizing moderate-sized resistors.
9. If the bends had been semicircular rather than square, they would have contributed 2.96 □ each.

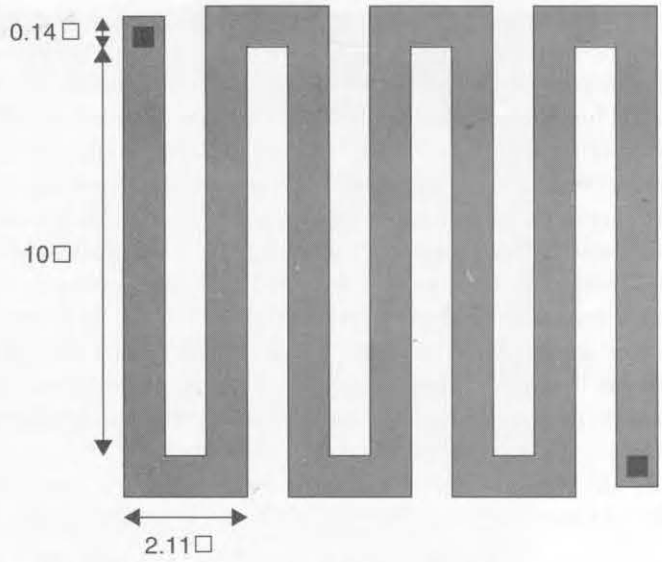


Fig. 2.23 A typical layout for an integrated resistor.

To this value, an additional impedance due to the contacts, about 20 Ω per contact, should be added to give a total resistance of about 1.70 kΩ. Note that quite large structures are required even for relatively small resistances when gate poly is used to realize the resistor.

When very accurate resistor ratios are required and the ratio can be expressed as a ratio of integer values, then an architecture similar to that shown in Fig. 2.24 can be

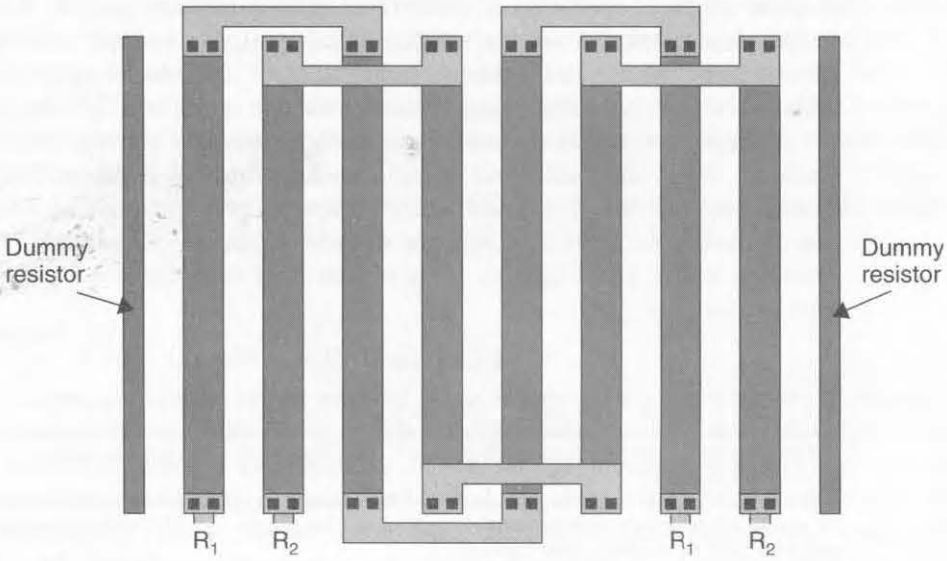


Fig. 2.24 A more accurate, but larger, layout for a resistor.

used. The resistor consists of several fingers connected at their ends using low-resistivity metal. This approach matches errors caused by the contact impedance between R_1 and R_2. Also, two dummy fingers have been included to match boundary conditions. This structure might result in about 0.1 percent matching accuracy of identical resistors if the finger widths are relatively wide (say, 10 μm in a 0.8-μm technology).

As with integrated capacitors, it is a good idea to place a shield under a resistor that is connected to a clean power supply. An appropriate shield might be a well region. This shielding helps keep substrate noise from being injected into the conductive layer. (Noise is due to capacitive coupling between the substrate and a large resistor structure.) Also, the parasitic capacitance between the resistor and the shield should be modelled during simulation. Its second-order effects on circuits such as RC filters can often be eliminated using optimization, which is available in many SPICE-like simulators. For low-noise designs, a metal shield over the top of a resistor may also be necessary, although it will result in a corresponding increase in capacitance.

For more information on realizing accurate resistor ratios, the reader is referred to [O'Leary, 1991; Maloberti, 1994].

Noise Considerations

Some additional layout issues help minimize noise in analog circuits. Most of these issues either attempt to minimize noise from digital circuits coupling into the substrate or analog power supplies, or try to minimize substrate noise that affects analog circuits.

With analog circuits, it is critical that different power-supply connections be used for analog circuits than for digital circuits. Ideally, these duplicate power supplies are connected only off the chip. Where a single I/O pin must be used for the power supply, it is still possible to use two different bonding wires extending from a single-package I/O pin to two separate bonding pads on the integrated circuit. At a very minimum, even if a single bonding pad is used for both analog and digital circuitry, two separated nets from the bonding pad out should be used for the different types of circuitry, as Fig. 2.25 shows. The reason the power-supply interconnects must be separated is that the interconnect does not have zero impedance. Every time a digital gate or buffer changes state, a glitch is injected on the digital power supply and in the surrounding substrate. By having the analog power supplies separate, we prevent this noise from affecting the analog circuitry. In the ideal case, separate pins are used for the positive power supply and for ground in both the digital and analog circuits. In addition, another pair of pins may be used for the supply voltage and ground for digital output buffers, which can inject very large current spikes. Finally, sometimes multiple pins are used for additional supply and grounds for very large microcircuits.

Another common precaution is to lay out the digital and analog circuitry in different sections of the microcircuit. The two sections should be separated by guard rings and wells connected to the power-supply voltages, as Fig. 2.26 shows. The p^+ connections to ground help keep a low-impedance path between the substrate and ground. For modelling purposes, the substrate can be modelled as a number of series-connected resistors with the p^+ ground connections modelled as resistor-dividers

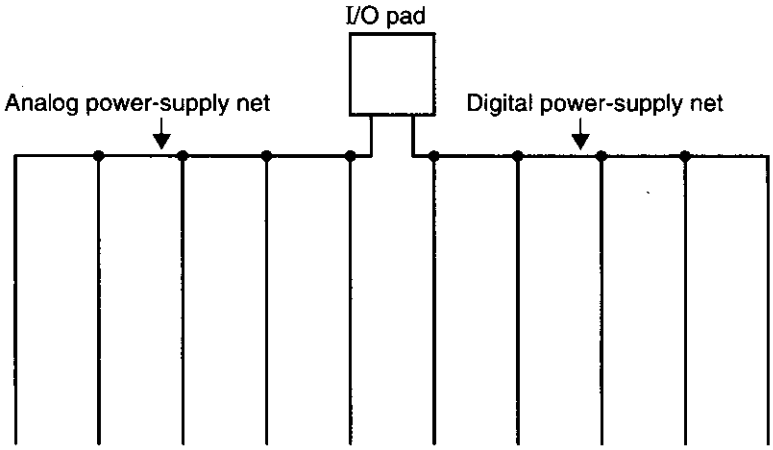


Fig. 2.25 Using separate nets for analog and digital power supplies.

having a small impedance to ground. These low-impedance ground connections help keep substrate noise from propagating through the resistive substrate. The use of the n well between p^+ connections helps to further increase the resistive impedance of the substrate between the analog and digital regions due to graded substrate doping. Specifically, the p^- substrate often has 10 times higher doping at the surface of the microcircuit compared to the doping level below the n well, which leads to a tenfold increase in substrate resistivity between the two p^+ connections. Finally, the n well also operates as a bypass capacitor to help lower the noise on V_{DD}.

Another important consideration when laying out a circuit that includes both analog and digital circuits is the use of shields connected to either ground or to a separate power-supply voltage. Figure 2.27 shows examples of the use of shields. In this example, an n well is used to shield the substrate from the digital interconnect line. The well is also used to shield an analog interconnect line from any substrate noise. This

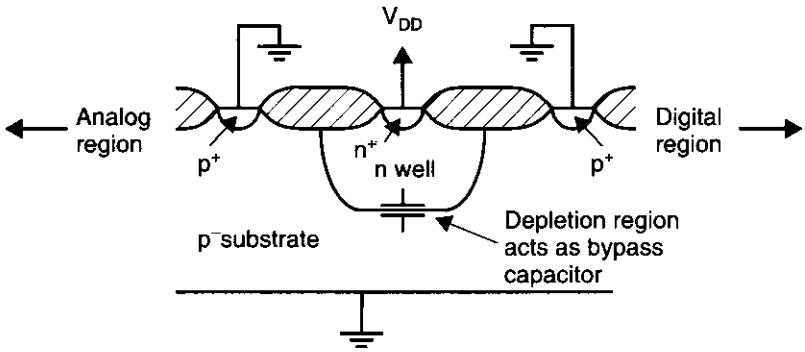


Fig. 2.26 Separating analog and digital areas with guard rings and wells in an attempt to minimize the injection of noise from digital circuits into the substrate under the analog circuit.

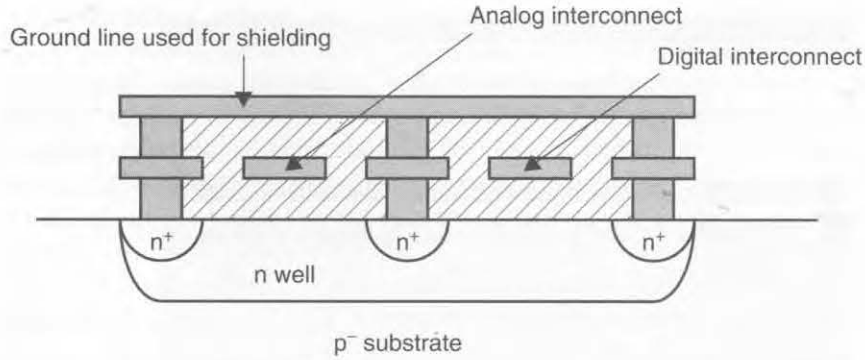


Fig. 2.27 Using shields helps keep noise from being capacitively coupled into and out of the substrate.

shield is ideally connected to a ground net that is used only for shields. If this type of connection is not possible due to layout and space constraints, then the digital ground can be used for the shields, although this is not ideal. In the example, the shield ground is also connected to metal lines that separate the analog and digital lines from each other and from other interconnect lines. Finally, an additional metal shield might be placed above the lines as well. This final shield may be somewhat excessive, but it can often be easily realized in many parts of the microcircuit if ground and power-supply lines are distributed in metal 2, perpendicular to the metal-1 interconnect lines. It should also be mentioned that the n well shield also acts as a bypass capacitor; this helps minimize noise in the substrate, which is connected to V_{DD}. Additional layers that are often used as shields are the polysilicon layers.

Perhaps the most effective technique for minimizing the propagation of substrate noise in a mixed-mode microcircuit containing both analog and digital circuitry is the use of an epitaxial process. An epitaxial process places a conductive layer under all transistors. Any charge flowing through the substrate is attracted to this layer and does not propagate into sensitive analogs regions. Although this process is more expensive, it also helps prevent latch-up. For submicron technologies, this epitaxial process is becoming more common because of its reduced latch-up sensitivity.

Careful thought should go into the overall placement of different blocks in a mixed-mode analog-digital microcircuit. A possible arrangement for an analog section containing switched-capacitor circuits is shown in Fig. 2.28. Notice that an n well shield is being used under the capacitors. Notice also that the clock lines are not only as far from the opamps as possible, but are also separated by two wells and a V_{SS} interconnect that is liberally connected to the substrate, two ground (Gnd) lines, and a V_{DD} line. A well is placed under the clock lines as a shield. This shield is connected to a separate ground line (perhaps digital ground) from the one used in the opamp region because this shield will likely have quite a bit of clock noise coupled into it. Also note that a separate V_{DD} line is used to connect to the n wells under the switches, a region where digital interconnects exist, as is used in the critical opamp section.

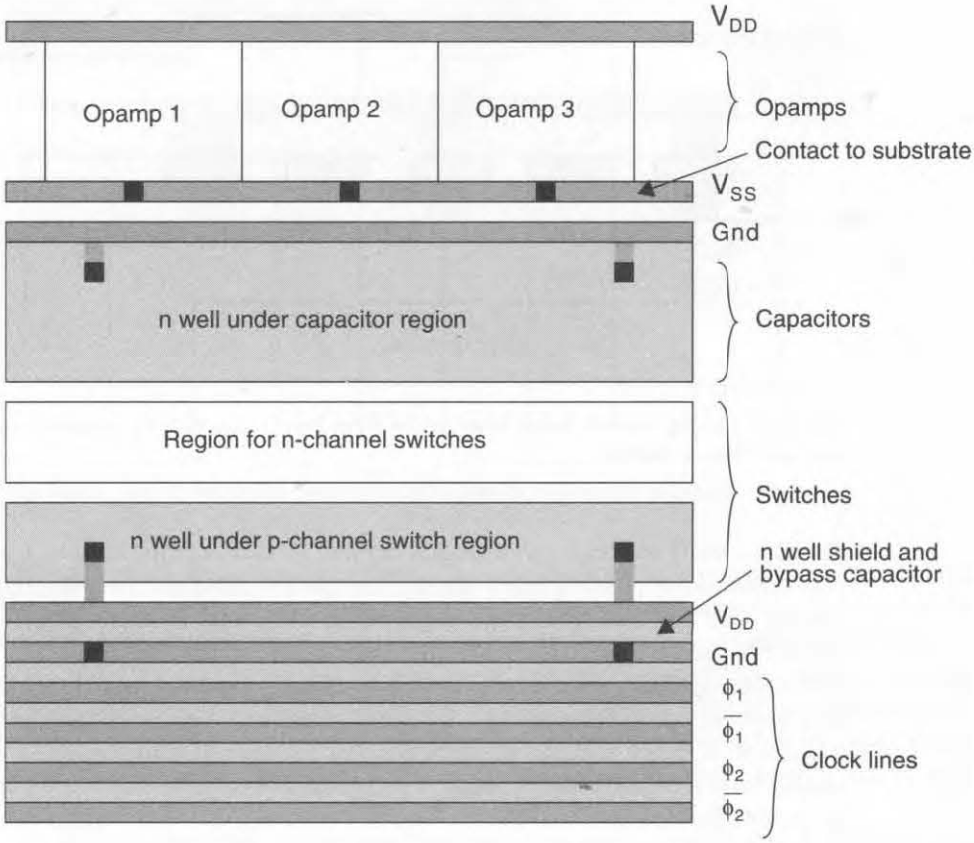


Fig. 2.28 A possible floor plan for an analog section containing switched-capacitor circuits.

One last technique for noise minimization in analog microcircuits should always be used: After layout has been finished, *any unused space should be filled with additional contacts to both the substrate and to the wells, which are used as bypass capacitors.* In a typical microcircuit, this results in a significant increase in bypass capacitance.

Many other techniques have been developed by various companies, but the preceding techniques give the reader a good idea of the types of practical considerations necessary when realizing high-performance analog microcircuits.

2.5 LATCH-UP

One of the effects that CMOS designers must be wary of, especially as dimensions shrink, is latch-up. This effect can occur when there are relatively large substrate or well currents or, equivalently, large substrate or well voltage drops, that might be caused by capacitive coupling. These triggering voltage drops often occur when power is first applied to a CMOS microcircuit.

A latched-up microcircuit is equivalent to a turned-on *silicon-controlled rectifier* (SCR) between the power supply and ground. This SCR effectively short-circuits the power supply on the microcircuit and, unless the supply-current is limited, irreparable damage will probably occur (such as a fused open bonding wire or interconnect line).

To understand latch-up, consider the cross section of the CMOS inverter shown in Fig. 2.29 with the parasitic bipolar transistors Q_1 and Q_2. Q_1 is a lateral npn, with the base being formed by the p^- substrate, whereas Q_2 is a vertical pnp, with the base being formed by the n-well region. The parasitic bipolar circuit has been redrawn in Fig. 2.30 along with some of the parasitic resistances due to the lightly doped substrate and well regions. The circuit realizes two cross-coupled common-emitter amplifiers in a positive feedback loop. This is the equivalent circuit of an SCR, which is sometimes referred to as a crowbar switch.

Normally, the parasitic bipolar transistors are off, and the voltages are as shown in Fig. 2.30(*a*). However, if latch-up is somehow triggered, they turn on when the

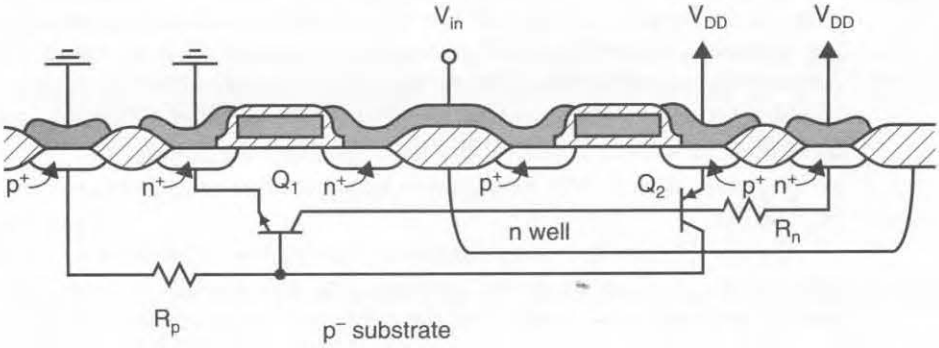


Fig. 2.29 Cross section of a CMOS inverter with superimposed schematic of the parasitic transistors responsible for the latch-up mechanism.

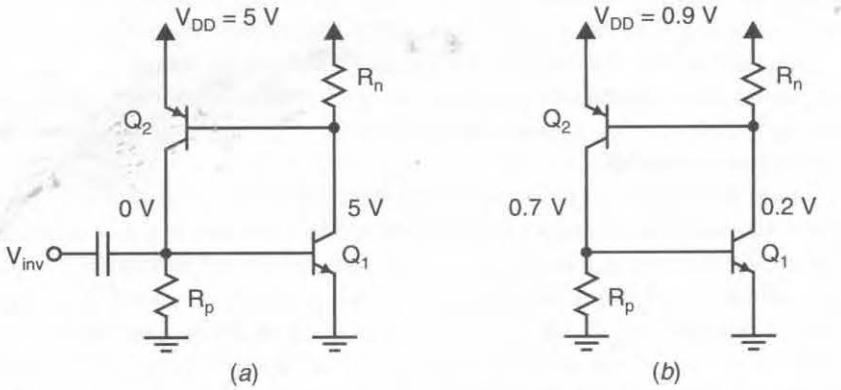


Fig. 2.30 (*a*) The equivalent circuit of the parasitic bipolar transistors, and (*b*) the voltages after latch-up has occurred.

loop gain is larger than unity, and as a result, the voltages are approximately those shown in Fig. 2.30(*b*). This turned-on SCR effectively places a short-circuit across the power-supply voltage and pulls V_{DD} down to approximately 0.9 V. If the power supply does not have a current limit, then excessive current will flow and some portion of the microcircuit may be destroyed.

Latch-up can be triggered in several ways. For example, the output of the CMOS inverter, V_{inv} in Fig. 2.29, is capacitively coupled to the bases of the bipolar transistors by the junction depletion capacitances of the MOS drains. If the inverter is large (as in the case of an output buffer), these capacitances will be large. When the output of the inverter changes, glitches will be capacitively coupled to the base nodes of the parasitic bipolars and may cause latch-up. Alternatively, substrate currents caused by hot-carrier electrons can also result in voltage drops large enough to trigger latch-up.

To prevent latch-up, the loop gain of the cross-coupled bipolar inverters must be kept less than unity. This low loop gain is normally achieved by keeping the current gains of the parasitic bipolars as low as is possible, and most importantly, by keeping shunting resistors R_n and R_p as small as possible. The current gain of the vertical pnp, Q_2, might be 50 to 100 and is difficult to minimize. The current gain of the lateral npn can be decreased by larger spacings between n-channel and p-channel transistors. However, with typically used spacings, the product $\beta_{npn}\beta_{pnp}$ is still normally greater than 1. The loop gain is kept less than unity primarily by decreasing R_n and R_p.

R_n and R_p are decreased mainly by having low-impedance paths between the substrate and well to the power supplies. One way to achieve these low-impedance paths is to have many contacts to the substrate. For example, with an n-well technology, the design rules normally specify a maximum distance between any place in the n-channel region of the microcircuit and the closest p^+ junction, which connects the substrate to ground. Similarly, in the p-channel regions, the maximum distance to the nearest n^+ junction, which connects the n wells to V_{DD}, is specified. Also, after layout is completed, a good designer fills any unused areas with extra connections to the substrate and well regions. In addition, any transistors that conduct large currents are usually surrounded by guard rings. These guard rings are connections to the substrate for n-channel transistors, or to the n well for p-channel transistors, that completely surround the high-current transistors. Also, ensuring that the back of the die is connected to ground through a eutectic gold bond to the package header is helpful.

One of the best ways of preventing latch-up is to use an epitaxial process, especially one with highly doped buried layers. For example, if a p^+ substrate has a p^- epitaxial layer in which the transistors are placed, device performance is only marginally affected, but the highly conductive p^+ substrate has very little impedance to ground contacts and to the package header. Alternatively, one might use a p^- substrate that has n^+- and p-buried regions and an intrinsic epitaxial region that is separately and optimally ion implanted to form the n-channel and p-channel regions. This self-aligned twin-tub technology is very immune to latch-up due to the highly conductive buried layers.

2.6 REFERENCES

D. Allstot and W. Black. "Technological Design Considerations for Monolithic MOS Switched-Capacitor Filtering Systems," *IEEE Proceedings*, Vol. 71, no. 8, pp. 967–986, August 1983.

R. Geiger, P. Allen, and N. Strader. *VLSI: Design Techniques for Analog and Digital Circuits*. McGraw-Hill, New York, 1990.

A. Glasser and G. Subak-Sharpe. *Integrated Circuit Engineering: Design, Fabrication, and Applications*. Addison-Wesley, Reading, Massachusetts, 1977.

A. Grebene, *Bipolar and MOS Analog Integrated Circuits*. Wiley-Interscience, Wiley, New York, 1984.

S. Muraka and M. Peckerar. *Electronic Materials*. Academic Press, New York, 1989, p. 326.

R. Haveman, et al. "A 0.8-μm, 256K BiCMOS SRAM Technology," *Digest of Technical Papers, 1987 Intern. Electron Devices Meeting*, pp. 841–843, December 1987.

R. Haken, R. Haveman, R. Ekund, and L. Hutter. "BiCMOS Process Design," in *BiCMOS Technology and Applications*, ed. A. Alvarez. Kluwer Academic Publishers, Nowell, Massachusetts, 1989.

F. Maloberti. "Layout of Analog and Mixed Analog-Digital Circuits," in *Design of Analog-Digital VLSI Circuits for Telecommunication and Signal Processing*, ed. J. Franca and Y. Tsividis. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

P. O'Leary. "Practical Aspects of Mixed Analogue and Digital Design," in *Analogue-Digital Asics, Circuit Techniques, Design Tools, and Applications*, ed. R. S. Soin, F. Maloberti, and J. Franca. Peter Peregrinus, Stevenage, England, 1991.

D. Reinhard. *Introduction to Integrated Circuit Engineering*. Houghton Mifflin, Boston, 1987.

2.7 PROBLEMS

Unless otherwise stated, assume the following hold throughout the problems section:

- npn bipolar transistors:
  $\beta = 100$
  $V_A = 80$ V
  $\tau_b = 13$ ps
  $\tau_s = 4$ ns
  $r_b = 330\ \Omega$
- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8$ V
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}\ (\Omega) = 8{,}000L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9$ V
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}\ (\Omega) = 12{,}000L\ (\mu m)/I_D\ (mA)$ in active region

$C_j = 4.5 \times 10^{-4}$ pF/(μm)^2
$C_{j\text{-}sw} = 2.5 \times 10^{-4}$ pF/μm
$C_{ox} = 1.9 \times 10^{-3}$ pF/(μm)^2
$C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}$ pF/μm

2.1 Discuss briefly the relationships between an ion beam's acceleration potential, the beam current, and the time of implantation on the resulting doping profile.

2.2 Place the following processing steps in their correct order: metal deposition and patterning, field implantation, junction implantation, well implantation, polysilicon deposition and patterning, field-oxide growth.

2.3 What are the major problems associated with a single threshold-voltage-adjust implant?

2.4 What is the reason for using a field implant and why is it often not needed in the well regions?

2.5 What are the major trade-offs in using a wet process or a dry process when growing thermal SiO_2?

2.6 Why is polysilicon rather than metal used to realize gates of MOS transistors?

2.7 Why can't a microcircuit be annealed after metal has been deposited?

2.8 What minimum distance, in terms of λ, would you expect that metal should be separated from polysilicon? Why?

2.9 Find the circuit that the layout shown in Fig. P2.9 realizes. Simplify the circuit, if possible, and give the sizes of all transistors. Assume L = 2λ, where λ = 1 μm.

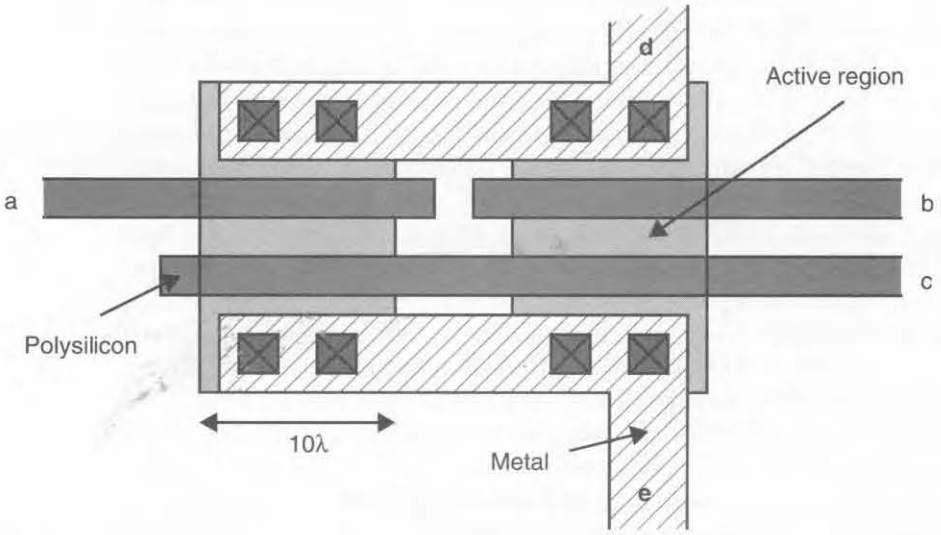


Fig. P2.9

2.10 Find the transistor schematic for the CMOS logic circuit realized by the layout shown in Fig. P2.10. Give the widths of all transistors. Assume L = 2λ,

where $\lambda = 0.4\,\mu m$. In tabular form, give the area and perimeter of each junction that is not connected to V_{DD} or to ground.

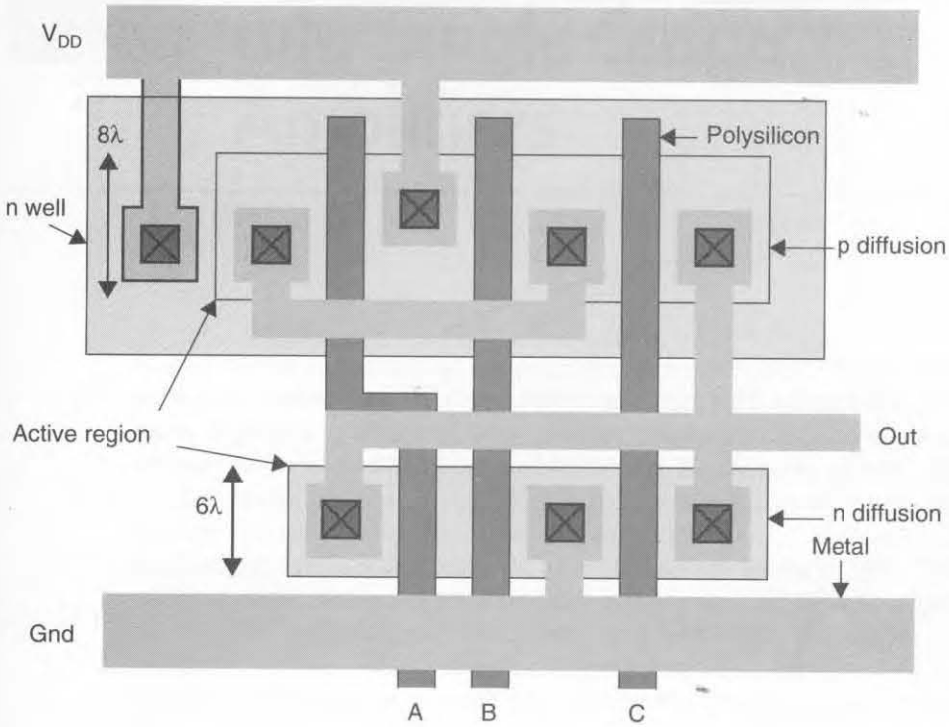


Fig. P2.10

2.11 Repeat Example 2.1 for the case in which the two transistors do not physically share any junction, but each junction is realized in a way similar to junction J_2.

2.12 Repeat Example 2.2 where an overall transistor width of 80λ is still desired, but assume 8 parallel transistors, each of width 10λ, are used.

2.13 Repeat Example 2.2 where an overall transistor width of 80λ is still desired, but assume 2 parallel transistors, each of width 40λ, are used.

2.14 We desire to match two capacitors of relative sizes 9 and 4.523. Sketch a layout for the two capacitors such that their ratio will be maintained during overetching.

2.15 Given that a unit-sized capacitor is $5\ \mu m \times 5\ \mu m$ and that the e-beam lithography rounds all sizes to $0.1\ \mu m$, what is the matching accuracy due to rounding of the capacitors found in Problem 2.14? What is the new matching accuracy if the capacitor sizes are doubled to 18 and 9.046?

2.16 Given that a polysilicon layer has $7\ \Omega/\square$, what is the resistance of a long line that is $2\ \mu m$ wide and $1{,}000\ \mu m$ long? (Ignore any contact resistance.)

2.17 Repeat Problem 2.16 for the case in which the resistive line is laid out in a serpentine manner, as shown in Fig. 2.23, where enough bends are used such that a line drawn along the middle of the serpentine resistor has length 1,000 μm (the last finger length might be short). What is the resulting height and width of the overall resistor layout?

CHAPTER

3 Basic Current Mirrors and Single-Stage Amplifiers

In this chapter, fundamental building blocks are described. These blocks include a variety of current mirrors, single-stage amplifiers with active loads, and differential pairs. A good knowledge of these building blocks is critical to understanding many subjects in the rest of this book and for analog IC design in general. Although some bipolar circuits are described, CMOS mirrors and gain stages are emphasized because modern designs mostly make use of CMOS technology. Fortunately, most of the small-signal analyses presented can be applied to bipolar circuits with little change. In addition, rather than using resistive loads and ac coupling, the gain stages covered are shown with current-mirror active loads since such loads are almost always used in integrated circuits. Finally, single-transistor amplifiers are described here for completeness, although it is assumed that most readers are already somewhat familiar with them.

3.1 SIMPLE CMOS CURRENT MIRROR

A simple CMOS current mirror is shown in Fig. 3.1, in which it is assumed that both transistors are in the active region. If the finite output impedances of the transistors are ignored, and it is assumed that both transistors are the same size, then Q_1 and Q_2 will have the same current since they both have the same gate-source voltage, V_{gs}. However, when finite output impedance is considered, whichever transistor has a larger drain-source voltage will also have a larger current. In addition, the finite output impedance of the transistors will cause the small-signal output impedance of the current mirror, that is, the small-signal impedance looking into the drain of Q_2, to be less than infinite. To find the output impedance of the current mirror, r_{out}, the small-signal circuit is analyzed after placing a signal source, v_x, at the output node. Then by definition, r_{out} is given by the ratio v_x/i_x, where i_x is the current flowing out of the source and into the drain of Q_2.

Before finding r_{out}, consider the small-signal model for Q_1 alone, as shown in Fig. 3.2(*a*). Note that Q_1 is diode connected (i.e., its drain and gate are connected) and that I_{in} does not exist in the small-signal model; I_{in} was replaced with an open

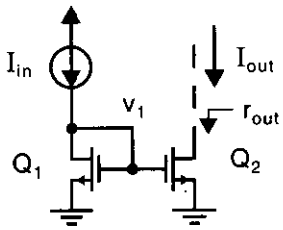


Fig. 3.1 A simple CMOS current mirror.

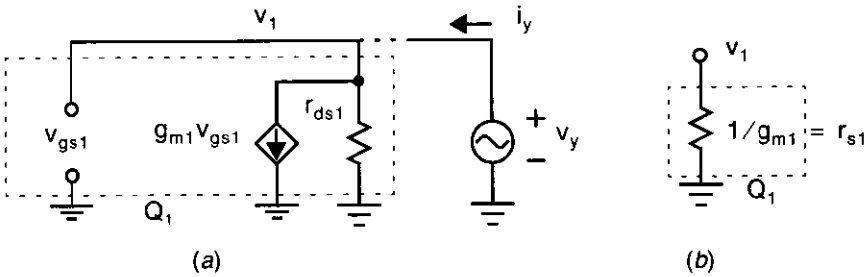


Fig. 3.2 (*a*) The small-signal model for Q_1 and (*b*) the equivalent small-signal model for Q_1.

circuit because it is an *independent* current source. Also note that a low-frequency small-signal model is used for Q_1 (i.e., all the capacitors are ignored in the model). This small-signal model can be further reduced by finding the *Thévenin-equivalent* circuit. The Thévenin-equivalent output voltage is 0 since the circuit is stable and contains no input signal. This circuit's Thévenin-equivalent output impedance is found by applying a test signal voltage, v_y, at v_1 and measuring the signal current, i_y, as shown. Here, the current i_y is given by

$$i_y = \frac{v_y}{r_{ds1}} + g_{m1}v_{gs1} = \frac{v_y}{r_{ds1}} + g_{m1}v_y \tag{3.1}$$

and recalling that the output impedance is given by v_y/i_y, the output impedance equals $1/g_{m1} \parallel r_{ds1}$. Because typically $r_{ds1} >> 1/g_{m1}$, we approximate the output impedance to be simply $1/g_{m1}$ (which is also defined to be r_{s1}), which results in the equivalent model shown in Fig. 3.2(*b*). This same result holds in the bipolar case and is also equivalent to the small-signal model for a diode—hence the name *diode-connected* transistor.

Using the model just described leads to the simplified small-signal model for the overall current mirror, as shown in Fig. 3.3(*a*), where v_{gs2} has been connected to ground via a resistance of $1/g_{m1}$. Since no current flows through the $1/g_{m1}$ resistor, v_{gs2} equals 0 no matter what voltage level v_x is applied to the current-mirror output. This should come as no surprise, since MOS transistors operate unilaterally at low frequencies. Thus, since $g_{m2}v_{gs2} = 0$, the circuit is simplified to the equivalent small-signal model shown in Fig. 3.3(*b*). The small-signal output impedance, r_{out}, is simply equal to r_{ds2}.

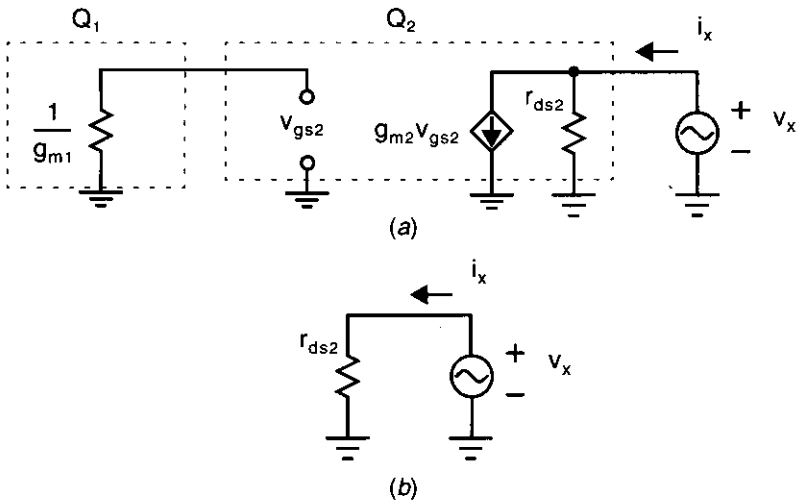


Fig. 3.3 (*a*) A small-signal model for the current mirror of Fig. 3.1 and (*b*) a simplified small-signal model.

EXAMPLE 3.1

Consider the current mirror shown in Fig. 3.1, where $I_{in} = 100\ \mu A$ and each transistor has $W/L = 100\ \mu m/1.6\ \mu m$. Given that $\mu_n C_{ox} = 92\ \mu A/V^2$, $V_{tn} = 0.8$ V and $r_{ds} = [8{,}000L\ (\mu m)]/[I_D\ (mA)]$, find r_{out} for the current mirror and the value of g_{m1}. Also, estimate the change in I_{out} for a 0.5 V change in the output voltage.

Solution

Since the W/L ratios of Q_1 and Q_2 are the same, the nominal value of I_{out} equals that of $I_{in} = 100\ \mu A$. Thus, we have

$$r_{out} = r_{ds2} = \frac{8{,}000 \times 1.6}{0.1} = 128\ k\Omega \tag{3.2}$$

The value of g_{m1} is given by

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)I_{D1}} = 1.07\ mA/V \tag{3.3}$$

resulting in $r_{s1} = 1/g_{m1} = 935\ \Omega$. Note that this r_{s1} value is significantly less than r_{ds1}, which equals r_{ds2} in this case.

The change in output current can be estimated, using r_{out}, as

$$\Delta I_{out} = \frac{\Delta V}{r_{out}} = \frac{0.5}{128\ k\Omega} = 3.9\ \mu A \tag{3.4}$$

In other words, if initially I_{out} is measured to be 101 μA (due to mismatch or a larger V_{DS} voltage), then a 0.5 V increase in output voltage would result in a

new output current of about 105 μA. Note that this estimate does not account for second-order effects such as the fact that r_{ds} changes as the output current changes.

3.2 COMMON-SOURCE AMPLIFIER

A common use of simple current mirrors is in a single-stage amplifier with an active load, as shown in Fig. 3.4. This common-source topology is the most popular gain stage, especially when high-input impedance is desired.

Here, an n-channel common-source amplifier has a p-channel current mirror used as an active load to supply the bias current for the drive transistor. By using an active load, a high-impedance output load can be realized without using excessively large resistors or a large power-supply voltage. As a result, for a given power-supply voltage, a larger voltage gain can be achieved using an active load than would be possible if a resistor were used for the load. For example, if a 1-MΩ load were required with a 100-μA bias current, a resistive-load approach would require a power-supply voltage of 1 MΩ × 100 μA = 100 V. An active load makes use of the nonlinear, large-signal transistor equations to create simultaneous conditions of large bias currents and large small-signal resistances.

A small-signal equivalent circuit for low-frequency analysis of the common-source amplifier of Fig. 3.4 is shown in Fig. 3.5. V_{in} and R_{in} are the Thévenin equivalent of the input source. It is assumed that the bias voltages are such that both transistors are in the active region. The output resistance, R_2, is made up of the parallel combination of the drain-to-source resistance of Q_1, that is, r_{ds1}, and the drain-

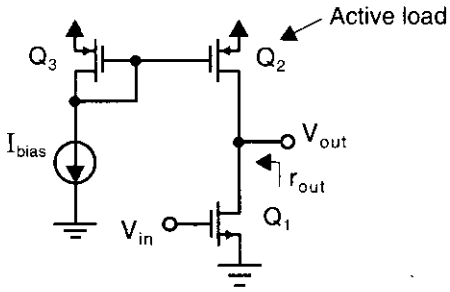


Fig. 3.4 A common-source amplifier with a current-mirror active load.

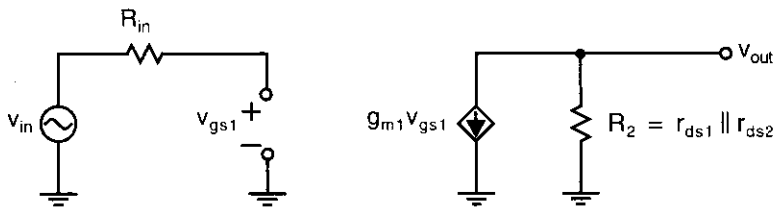


Fig. 3.5 A small-signal equivalent circuit for the common-source amplifier.

to-source resistance of Q_2, that is, r_{ds2}. Notice that the voltage-controlled current source modelling the body effect has not been included since the source is at a small-signal ground, and, therefore, this source always has 0 current.

Using small-signal analysis, we have $v_{gs1} = v_{in}$ and, therefore,

$$A_V = \frac{v_{out}}{v_{in}} = -g_{m1}R_2 = -g_{m1}(r_{ds1} \parallel r_{ds2}) \tag{3.5}$$

Depending on the device sizes, currents, and the technology used, a typical gain for this circuit is in the range of −10 to −100. To achieve similar gains with resistive loads, much larger power-supply voltages than 5 V must be used. This resistive-load approach also greatly increases the power dissipation. However, it should be mentioned here that for low-gain, high-frequency stages, it may be desirable to use resistor loads (if they do not require much silicon area) because they often have less parasitic capacitances associated with them. They are also typically less noisy than active loads.

EXAMPLE 3.2

Assume all transistors have $W/L = 100\ \mu m/1.6\ \mu m$ in Fig. 3.4, and that $\mu_n C_{ox} = 90\ \mu A/V^2$, $\mu_p C_{ox} = 30\ \mu A/V^2$, $I_{bias} = 100\ \mu A$, $r_{ds\text{-}n}(\Omega) = 8{,}000L\ (\mu m)/I_D\ (mA)$, and $r_{ds\text{-}p}(\Omega) = 12{,}000L\ (\mu m)/I_D\ (mA)$. What is the gain of the stage?

Solution

We have

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)_1 I_{bias}} = 1.06\ mA/V \tag{3.6}$$

Also,

$$r_{ds1} = \frac{8{,}000 \times 1.6\ \mu m}{0.1\ mA} = 128\ k\Omega \tag{3.7}$$

and

$$r_{ds2} = \frac{12{,}000 \times 1.6\ \mu m}{0.1\ mA} = 192\ k\Omega \tag{3.8}$$

Using Eq. (3.5), we have

$$A_V = -g_{m1}(r_{ds1} \parallel r_{ds2}) = -1.06(128 \parallel 192) = -81.4 \tag{3.9}$$

3.3 SOURCE-FOLLOWER OR COMMON-DRAIN AMPLIFIER

Another general use of current mirrors is to supply the bias current of source-follower amplifiers, as shown in Fig. 3.6. In this example, Q_1 is the source follower and Q_2 is

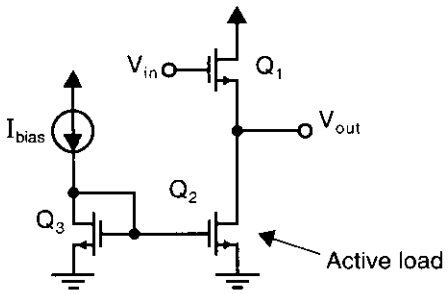


Fig. 3.6 A source-follower stage with a current mirror used to supply the bias current.

an active load that supplies the bias current of Q_1. These amplifiers are commonly used as *voltage buffers* and are therefore commonly called source followers. They are also referred to as common-drain amplifiers, since the input and output nodes are at the gate and source nodes, respectively, with the drain node being at small-signal ground. Although the dc level of the output voltage is not the same as the dc level of the input voltage, ideally the small-signal voltage gain is close to unity. In reality, it is somewhat less than unity. However, although this circuit does not generate voltage gain, it does have the ability to generate current gain.

A small-signal model for low-frequency analysis of this source-follower stage is shown in Fig. 3.7. Note that the voltage-controlled current source that models the body effect of MOS transistors has been included. This body effect is included because the source is not at small-signal ground and the body effect is a major limitation on the small-signal gain. Note that in Fig. 3.7, r_{ds1} is in parallel with r_{ds2}. Notice also that the voltage-controlled current source modelling the body effect produces a current that is proportional to the voltage across it. This relationship makes the body effect equivalent to a resistor of size $1/g_{s1}$, which is also in parallel with r_{ds1} and r_{ds2}. Thus, the small-signal model of Fig. 3.7 is equivalent to the simplified small-signal model of Fig. 3.8, in which $R_{s1} = r_{ds1} \parallel r_{ds2} \parallel 1/g_{s1}$. Writing the nodal equation at v_{out}, and noting that $v_{gs1} = v_{in} - v_{out}$, we have

$$v_{out}G_{s1} - g_{m1}(v_{in} - v_{out}) = 0 \tag{3.10}$$

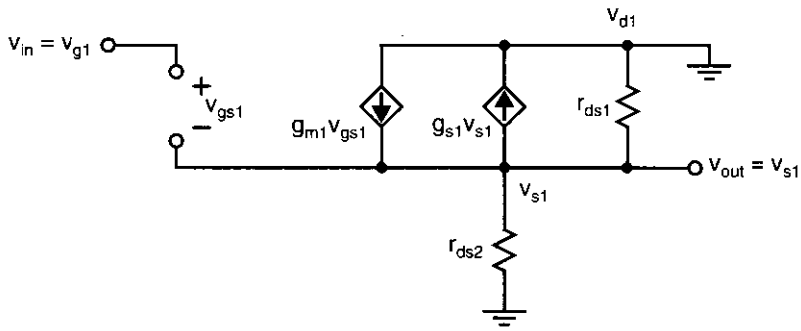


Fig. 3.7 The low-frequency model of the source-follower amplifier.

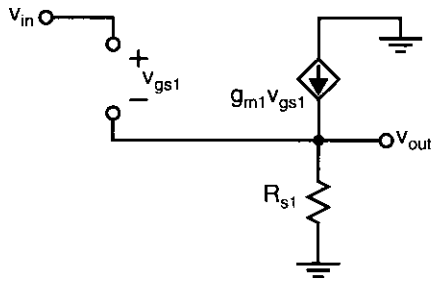


Fig. 3.8 An equivalent small-signal model for the source follower.

where $G_{s1} = 1/R_{s1}$.[1] Before proceeding, it is worth mentioning here that the authors have found that, to minimize circuit equation errors, a consistent methodology should be maintained when writing nodal equations. The methodology employed here is as follows: The first term is always the node at which the currents are being summed. This node voltage is multiplied by the sum of all admittances connected to the node. The next negative terms are the adjacent node voltages, and each is multiplied by the connecting admittance. The last terms are any current sources with a multiplying negative sign used if the current is shown to flow into the node.

Solving for v_{out}/v_{in}, we have

$$A_v = \frac{v_{out}}{v_{in}} = \frac{g_{m1}}{g_{m1} + G_{s1}} = \frac{g_{m1}}{g_{m1} + g_{s1} + g_{ds1} + g_{ds2}} \tag{3.11}$$

Normally, g_{s1} is on the order of one-tenth to one-fifth that of g_{m1}. Also, the transistor output admittances, g_{ds1} and g_{ds2}, might be one-tenth that of the body-effect parameter, g_{s1}. Therefore, it is seen that the body-effect parameter is the major source of error causing the gain to be less than unity. Notice also that at low frequencies the stage is completely unilateral. In other words, there is no signal flow from the output to the input.

EXAMPLE 3.3

Consider the source follower of Fig. 3.6, where all transistors have $W/L = 100\ \mu m/1.6\ \mu m$, $\mu_n C_{ox} = 90\ \mu A/V^2$, $\mu_p C_{ox} = 30\ \mu A/V^2$, $I_{bias} = 100\ \mu A$, $\gamma_n = 0.5\ V^{1/2}$, $r_{ds\text{-}n}(\Omega) = 8{,}000L\ (\mu m)/I_D\ (mA)$. What is the gain of the stage?

Solution

Notice that many parameters are the same as in Example 3.2. Repeating, we have

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)_1 I_{bias}} = 1.06\ mA/V \tag{3.12}$$

Also,

$$r_{ds1} = r_{ds2} = \frac{8{,}000 \cdot 1.6\ \mu m}{0.1\ mA} = 128\ k\Omega \tag{3.13}$$

1. Whenever a variable is designated G_i, it is assumed that the variable is an admittance and that $G_i = 1/R_i$, where R_i is the resistance of the same component.

The equation for the body-effect parameter, (Eq. 1.87) from Chapter 1, is

$$g_{s1} = \frac{\gamma g_m}{2\sqrt{V_{SB} + |2\phi_F|}} \tag{3.14}$$

To calculate this parameter, we need to know the source-bulk voltage, V_{SB}. Unfortunately, this voltage is dependent on the application and cannot be known accurately beforehand. Here we will assume that 5 V power supplies are being used, and that $V_{SB} \approx 2$ V. This is somewhat arbitrary, but it is the best one can do without more details. We therefore have

$$g_{s1} = \frac{0.5 \cdot g_m}{2\sqrt{2 + 0.7}} = 0.15 g_m = 0.16 \text{ mA/V} \tag{3.15}$$

Using (3.11), we have

$$A_V = \frac{1.06}{1.06 + 0.16 + 1/128 + 1/128} = 0.86 \text{ V/V} \tag{3.16}$$

Note that, as mentioned above, the fact that this result is so far below unity is mainly due to the body-effect parameter, g_s. If the body effect were not present, the gain would be around 0.99 V/V.

3.4 COMMON-GATE AMPLIFIER

A common-gate amplifier with an active load is shown in Fig. 3.9. This stage is commonly used as a gain stage when a relatively small input impedance is desired. For example, it might be designed to have an input impedance of 50 Ω to terminate a 50-Ω transmission line. Another common application for a common-gate amplifier is the first stage of an amplifier designed to amplify current rather than voltage.

If we use straightforward small-signal analysis, when the impedance seen at V_{out} (in this case, the output impedance of the current mirror formed by Q_2) is much less than r_{ds1}, the input impedance, r_{out}, is found to be $1/g_{m1}$ at low frequencies. However, in integrated applications, the impedance seen at V_{out} is often on the same order of magnitude or even much greater than r_{ds1}. In this case, the input impedance at low frequencies can be considerably larger than $1/g_{m1}$. To see this result, consider the small-signal model shown in Fig. 3.10. In this model, the voltage-dependent current source that models the body effect has been included. Notice that $v_{gs1} = -v_{s1}$ and therefore the two current sources can be combined into a single current source, as shown in Fig. 3.11. This simplification is always possible for a transistor that has a grounded gate in a small-signal model, and considerably simplifies taking the body effect into account. Specifically, one can simply ignore the body effect for transistors with grounded gates, and then, after the analysis is complete, simply replace the constants g_{mi} with $g_{mi} + g_{si}$. However, for this example, we include the body-effect parameter throughout the analysis.

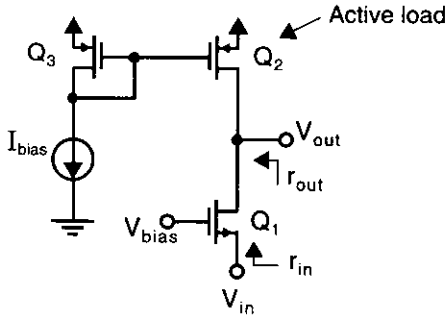


Fig. 3.9 A common-gate amplifier with a current-mirror active load.

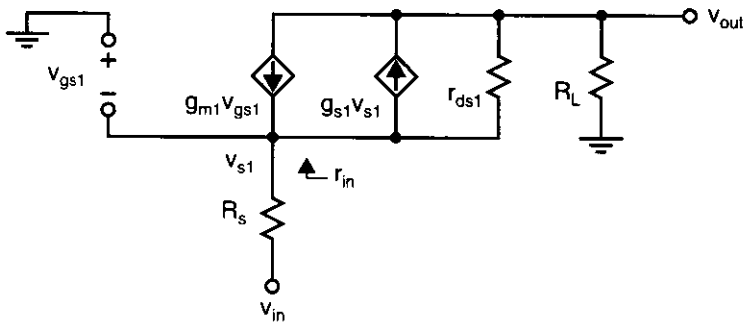


Fig. 3.10 The small-signal model of the common-gate amplifier at low frequencies.

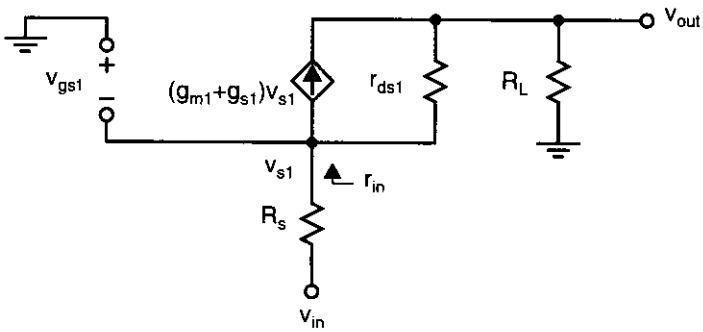


Fig. 3.11 A simplified small-signal model of the common-gate amplifier.

At node v_{out}, we have

$$v_{out}(G_L + g_{ds1}) - v_{s1}g_{ds1} - (g_{m1} + g_{s1})v_{s1} = 0 \tag{3.17}$$

Rearranging slightly, we have

$$\frac{v_{out}}{v_{s1}} = \frac{g_{m1} + g_{s1} + g_{ds1}}{G_L + g_{ds1}} \tag{3.18}$$

where it should be noted that this gain is approximately equal to $g_{m1}/(G_L + g_{ds1})$.

The current going into the source of Q_1 is given by

$$i_s = v_{s1}(g_{m1} + g_{s1} + g_{ds1}) - v_{out}g_{ds1} \tag{3.19}$$

Combining (3.18) and (3.19) to find the input admittance, $y_{in} = 1/r_{in}$, we have

$$y_{in} \equiv \frac{i_s}{v_{s1}} \equiv \frac{g_{m1} + g_{s1} + g_{ds1}}{1 + \dfrac{g_{ds1}}{G_L}} \cong \frac{g_{m1}}{1 + \dfrac{g_{ds1}}{G_L}} \tag{3.20}$$

Alternatively, we have

$$r_{in} = \frac{1}{y_{in}} \cong \frac{1}{g_{m1}}\left(1 + \frac{R_L}{r_{ds1}}\right) \tag{3.21}$$

With the p-channel active load shown in Fig. 3.9, $R_L = r_{ds2}$. Since, in this case, R_L is approximately the same magnitude as r_{ds1}, the input impedance, r_{in}, is about $2/g_{m1}$ for low frequencies—twice as large as the expected value of $1/g_{m1}$. This increased input impedance must be taken into account in applications such as transmission-line terminations. In some examples, the current-mirror output impedance realized by Q_2 is much larger than r_{ds1} (i.e., $R_L >> r_{ds1}$), and so *the input impedance for this common-gate amplifier is much larger than* $1/g_{m1}$. This increased input impedance often occurs in integrated circuits and is not commonly known.

The attenuation from the input to the transistor source can be considerable for a common-gate amplifier when R_s is large. This attenuation is given by

$$\frac{v_{s1}}{v_{in}} = \frac{G_s}{G_s + y_{in}} \tag{3.22}$$

using the admittance-divider rule.[2] Using (3.20) to replace y_{in}, we have

$$\frac{v_{s1}}{v_{in}} = \frac{G_s}{G_s + \dfrac{g_{m1} + g_{s1} + g_{ds1}}{1 + g_{ds1}/G_L}} \tag{3.23}$$

Using (3.18) and (3.23), we find that the overall dc gain is given by

$$\begin{aligned} A_V = \frac{v_{out}}{v_{in}} &= \left[\frac{G_s}{\left(G_s + \dfrac{g_{m1} + g_{s1} + g_{ds1}}{1 + g_{ds1}/G_L}\right)}\right]\frac{g_{m1} + g_{s1} + g_{ds1}}{G_L + g_{ds1}} \\ &\cong \left[\frac{G_s}{\left(G_s + \dfrac{g_{m1}}{1 + g_{ds1}/G_L}\right)}\right]\frac{g_{m1}}{G_L + g_{ds1}} \end{aligned} \tag{3.24}$$

2. This rule states that the gain is the ratio of the admittance connected between two nodes, divided by the sum of that admittance and the admittance between the second node and ground.

3.5 SOURCE-DEGENERATED CURRENT MIRRORS

We saw in Section 3.1 that a current mirror can be realized using only two transistors, where the output impedance of this current source was seen to be r_{ds2}. To increase this output impedance, a source-degenerated current mirror can be used, as shown in Fig. 3.12. The small-signal model for this current mirror is shown in Fig. 3.13. Since no current flows into the gate, the gate voltage is 0 V.

Note that the current i_x sourced by the applied voltage source is equal to the current through the degeneration resistor, R_s. Therefore, we have

$$v_s = i_x R_s \tag{3.25}$$

Also, note that

$$v_{gs} = -v_s \tag{3.26}$$

Setting i_x equal to the total current through $g_{m2}v_{gs}$ and r_{ds2} gives

$$i_x = g_{m2}v_{gs} + \frac{v_x - v_s}{r_{ds2}} \tag{3.27}$$

Substituting (3.25) and (3.26) into (3.27) gives

$$i_x = -i_x g_{m2} R_s + \frac{v_x - i_x R_s}{r_{ds2}} \tag{3.28}$$

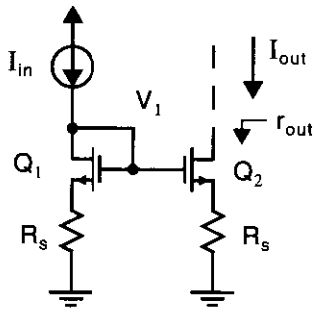


Fig. 3.12 A current mirror with source degeneration.

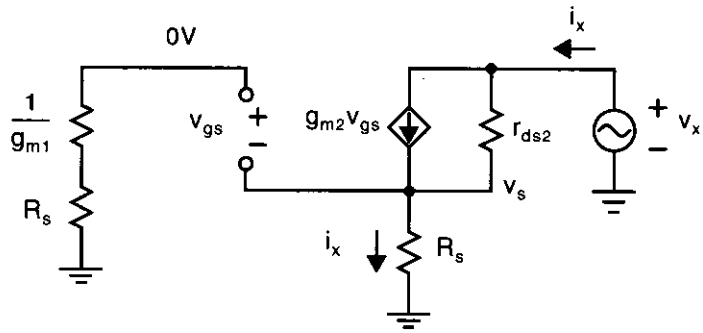


Fig. 3.13 The small-signal model for the source-degenerated current source.

Rearranging, we find the output impedance to be given by

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{ds2})] \cong r_{ds2}(1 + R_s g_{m2}) \tag{3.29}$$

where g_{ds2} is equal to $1/r_{ds2}$, which is much less than g_{m2} (recall that $g_m = 1/r_s$). Thus, the output impedance has been increased by a factor approximately equal to $(1 + R_s g_{m2})$.

This formula can often be applied to moderately complicated circuits to quickly estimate the impedances looking into a node. Such an example follows, in the derivation of the output impedance of cascode current mirrors.

It should be noted that the above derivation ignores the body effect of the transistor, even though the source of the transistor is not connected to a small-signal ground. As discussed earlier, in Section 3.4, since the gate is at a small-signal ground, the body effect can be taken into account by simply replacing g_{m2} in Eq. (3.29) with $g_{m2} + g_{s2}$. This substitution results in

$$r_{out} = \frac{v_x}{i_x} = r_{ds2}[1 + R_s(g_{m2} + g_{s2} + g_{ds2})] \cong r_{ds2}[1 + R_s(g_{m2} + g_{s2})] \tag{3.30}$$

where g_{s2} is the body-effect constant. This result is only slightly different since g_s is roughly one-fifth of g_m.

EXAMPLE 3.4

Consider the current mirror shown in Fig. 3.12, where $I_{in} = 100\ \mu A$, each transistor has $W/L = (100\ \mu m)/(1.6\ \mu m)$, and $R_s = 5\ k\Omega$. Given that $\mu_n C_{ox} = 92\ \mu A/V^2$, $V_{tn} = 0.8$ V, and $r_{ds} = [8{,}000L\ (\mu m)]/[I_D\ (mA)]$, find r_{out} for the current mirror. Assume the body effect can be approximated by $g_s = 0.2 g_m$.

Solution

Nominally, $I_{out} = I_{in}$, and thus we find the small-signal parameters for this current mirror to be

$$g_{m2} = \sqrt{2\mu_n C_{ox}(W/L)I_{out}} = 1.07\ mA/V \tag{3.31}$$

Also, as in Example 3.1, we have

$$r_{ds2} = \frac{8{,}000 \times 1.6}{0.1} = 128\ k\Omega \tag{3.32}$$

Now, making use of (3.29), the output impedance is given by

$$r_{out} = 128k\left[1 + 5\left(1.07 + 0.2 \times 1.07 + \frac{1}{128}\right)\right] = 955\ k\Omega \tag{3.33}$$

Note that this result is nearly eight times the output impedance for a simple current mirror as found in Example 3.1. Also note that the voltage drop across R_s equals $100\ \mu A \times 5\ k\Omega = 0.5$ V due to the dc bias current through it.

3.6 HIGH-OUTPUT-IMPEDANCE CURRENT MIRRORS

In this section two high-output-impedance current mirrors are described—the cascode and the Wilson current mirrors. These current mirrors have output impedances that are larger than that of a simple current mirror by a factor of $g_m r_{ds}$—the maximum gain of a single transistor.

Cascode Current Mirrors

A cascode current mirror is shown in Fig. 3.14. First, note that the output impedance looking into the drain of Q_2 is simply r_{ds2}, which is seen using an analysis very similar to that which was used for the simple current mirror. Thus, the output impedance can be immediately derived by considering Q_4 as a current source with a source-degeneration resistor of value r_{ds2}. Making use of Eq. (3.29), and noting that Q_4 is now the cascode transistor rather than Q_2, we have

$$r_{out} = r_{ds4}[1 + R_s(g_{m4} + g_{s4} + g_{ds4})] \tag{3.34}$$

where now $R_s = r_{ds2}$. Therefore, the output impedance is given by

$$\begin{aligned} r_{out} &= r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4} + g_{ds4})] \\ &\cong r_{ds4}[1 + r_{ds2}(g_{m4} + g_{s4})] \\ &\cong r_{ds4}(r_{ds2}g_{m4}) \end{aligned} \tag{3.35}$$

Thus, the output impedance has been increased by a factor of $g_{m4}r_{ds2}$, which is an upper limit on the gain of a single-transistor MOS gain-stage, and might be a value between 10 and 100, depending on the transistor sizes and currents and the technology being used. This significant increase in output impedance can be instrumental in realizing single-stage amplifiers with large low-frequency gains.

There is a disadvantage in using a cascode current mirror—it reduces the maximum output-signal swings possible before transistors enter the triode region. To understand this reduction, recall that for an n-channel transistor to be in the active region (also called the saturation or pinch-off region) its drain-source voltage must be greater than V_{eff}; V_{eff} was defined in (1.54) as

$$V_{eff} \equiv V_{GS} - V_{tn} \tag{3.36}$$

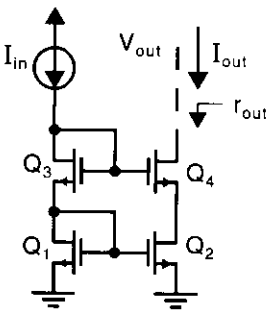


Fig. 3.14 A cascode current mirror.

which was shown in (1.80) to be given by

$$V_{eff} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{3.37}$$

If we assume all transistors have the same sizes and currents, then they also all have the same V_{eff} and, therefore, the same gate-source voltages, $V_{GSi} = V_{eff} + V_{tn}$. Also, from Fig. 3.14, we see that

$$V_{G3} = V_{GS1} + V_{GS3} = 2V_{eff} + 2V_{tn} \tag{3.38}$$

and

$$V_{DS2} = V_{G3} - V_{GS4} = V_{G3} - (V_{eff} + V_{tn}) = V_{eff} + V_{tn} \tag{3.39}$$

Thus, the drain-source voltage of Q_2 is larger than the minimum needed to place it at the edge of the active region. Specifically, the drain-source voltage of Q_2 is V_{tn} (about 0.8 V) greater than what is required. Since the smallest output voltage, V_{D4}, can be without Q_4 entering the triode region is given by $V_{DS2} + V_{eff}$, the minimum allowed voltage for V_{out} is given by

$$V_{out} > V_{DS2} + V_{eff} = 2V_{eff1} + V_{tn} \tag{3.40}$$

which, again, is V_{tn} greater than the minimum value of $2V_{eff}$. This loss of signal swing is a serious disadvantage when modern technologies are used that might have a maximum allowed power-supply voltage as small as 3 V. In the next chapter, we will see how the cascode current mirror can be modified to maintain large output impedances and yet still allow for near minimum voltages at the output of the mirror.

EXAMPLE 3.5

Consider the cascode current mirror shown in Fig. 3.14, where $I_{in} = 100\ \mu A$ and each transistor has $W/L = (100\ \mu m)/(1.6\ \mu m)$. Given that $\mu_n C_{ox} = 92\ \mu A/V^2$, $V_{tn} = 0.8$ V, and $r_{ds} = [8{,}000L\ (\mu m)]/[I_D\ (mA)]$, find r_{out} for the current mirror (approximate the body effect by $0.2g_m$). Also find the minimum output voltage at V_{out} such that the output transistors remain in the active region.

Solution

Nominally, $I_{out} = I_{in}$, and thus we can find the small-signal parameters for this current mirror to be

$$g_{m4} = \sqrt{2\mu_n C_{ox}(W/L)I_{out}} = 1.07\ mA/V \tag{3.41}$$

We also have

$$r_{ds2} = r_{ds4} = \frac{8{,}000 \times 1.6}{0.1} = 128\ k\Omega \tag{3.42}$$

Now, making use of (3.35), the output impedance is given by

$$r_{out} = 128k[128(1.07 + 0.2 \times 1.07)] = 21\ M\Omega \tag{3.43}$$

To find the minimum output voltage, we first need to determine V_{eff}:

$$V_{eff} = \sqrt{\frac{2I_{out}}{\mu_n C_{ox}(W/L)}} = 0.19\ V \tag{3.44}$$

Thus, the minimum output voltage is determined to be $2 \times 0.19 + 0.8 = 1.18$ V.

Wilson Current Mirror

Another commonly used current mirror is the Wilson current mirror, shown in Fig. 3.15.[3] It is an example of using shunt-series feedback to increase the output impedance [Sedra, 1991]. Basically, Q_2 senses the output current and then mirrors it to I_{D1}, which, in turn, is subtracted from the input current, I_{in}. Note that I_{D1} must precisely equal I_{in}; otherwise the voltage at the gate of Q_3, Q_4 would either increase or decrease, and the negative feedback loop forces this equality. This feedback arrangement increases the output impedance by an amount equal to 1 plus the loop gain. Assuming all devices are matched, the output impedance without the feedback due to Q_1, Q_3 would be $2r_{ds4}$, taking into account that Q_4 has source degeneration equal to $1/g_{m2}$ (i.e., the small-signal impedance of diode-connected Q_2), which is responsible for the 2 factor. The loop gain is approximately given by

$$A_L \cong \frac{g_{m1}(r_{ds1} \parallel r_{in})}{2} \tag{3.45}$$

where r_{in} is the input impedance of the biasing current source I_{in}. The factor of 1/2 is due to the voltage attenuation from the gate of Q_4 to its source, caused by the source degeneration of the diode-connected Q_2. Assuming r_{in} is approximately equal

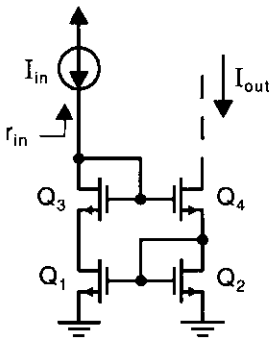


Fig. 3.15 The Wilson current mirror.

3. Actually, this is a modified Wilson current mirror, where Q_3 has been added to make the drain-source voltages of Q_1 and Q_2 approximately equal, which improves the absolute matching of I_{in} and I_{out}.

to r_{ds1}, then the loop gain is given by

$$A_L \cong \frac{g_{m1} r_{ds1}}{4} \tag{3.46}$$

and the output impedance is therefore given by

$$r_{out} \cong 2r_{ds4}\frac{g_{m1}(r_{ds1} \parallel r_{in})}{2} \cong r_{ds4}\left(\frac{g_{m1} r_{ds1}}{2}\right) \tag{3.47}$$

which is roughly one-half the output impedance for that of a cascode current mirror. For this reason, the cascode current mirror is often preferred over the Wilson current mirror. In terms of output voltage swing, the minimum allowed voltage across the current mirror, before Q_4 enters the triode region, is $2V_{eff1} + V_{tn}$, which is similar to that of the cascode current mirror.

Finally, it should be noted that Q_3 is not required in the Wilson current mirror. It has been included to give Q_1 and Q_2 the same drain-source bias voltages, and thus minimizes inaccuracies caused by the large-signal output impedances of the transistors. Without this transistor, the output current would be slightly smaller than the input current because V_{DS1} would be larger than V_{DS2}. However, the small-signal output impedance would remain the same.

3.7 CASCODE GAIN STAGE

In modern IC design, a commonly used configuration for a single-stage amplifier is a cascode configuration. This configuration consists of a common-source-connected transistor feeding into a common-gate-connected transistor. Two examples of cascode amplifiers are shown in Fig. 3.16. The configuration in Fig. 3.16(*a*) has both an n-channel common-source transistor, Q_1, and an n-channel common-gate cascode transistor, Q_2. This configuration is sometimes called a telescopic-cascode amplifier. The configuration shown in Fig. 3.16(*b*) has an n-channel input (or drive) transistor, but a p-channel transistor is used for the cascode (or common-gate) transistor. This configuration is usually called a folded-cascode stage. It allows the dc level of the output signal to be the same as the dc level of the input signal. Unfortunately, it is usually slower than the telescopic-cascode amplifier because although parasitic capacitances at the source of the cascode transistor are similar in both cases, the impedance levels of the folded-cascode stage are roughly three times larger due to the smaller transconductance of p-channel transistors as compared to n-channel transistors.

There are two major reasons for the popularity of cascode stages. The first is that they can have quite large gain for a single stage due to the large impedances at the output. To enable this high gain, the current sources connected to the output nodes are realized using high-quality cascode current mirrors. Normally this high gain is obtained without any degradation in speed, and sometimes with an improvement in speed. The second major reason for the use of cascode stages is that they limit the voltage across the input drive transistor. This minimizes any short-channel effects,

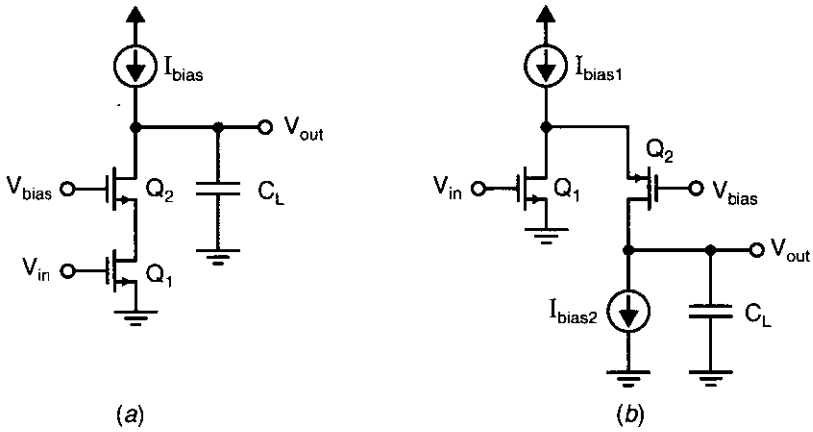


Fig. 3.16 (*a*) A telescopic-cascode amplifier and (*b*) a folded-cascode amplifier.

which becomes more important with modern technologies having very short channel-length transistors.

The analysis of the cascode gain stage is based on the telescopic stage of Fig. 3.16(*a*). The same analysis, with only minor modifications, also applies for the folded-cascode stage of Fig. 3.16(*b*).

From the section on cascode current mirrors, we know that the impedance looking into the drain of cascode transistor Q_2 is approximately given by

$$r_{d2} \cong g_{m2} r_{ds1} r_{ds2} \tag{3.48}$$

The total impedance at the output node is r_{d2} in parallel with R_L, where R_L is the output impedance of the bias current source, I_{bias}. Assuming I_{bias} is a high-quality source with an output impedance on the order of

$$R_L \cong g_{m\text{-}p} r_{ds\text{-}p}^2 \tag{3.49}$$

then the total impedance at the output node is

$$R_{out} \cong \frac{g_m r_{ds}^2}{2} \tag{3.50}$$

We have dropped the indices here under the assumption that the transistors are somewhat matched, and to simplify matters since we are only deriving an approximate solution.

To find the approximate low-frequency gain, we can use part of the analysis done previously for the common-gate stage. Repeating Eq. (3.20) here for convenience, we found that the low-frequency impedance looking into the source of the common-gate, or cascode, transistor, Q_2, was given by

$$Y_{in2} = \frac{g_{m2} + g_{s2} + g_{ds2}}{1 + \dfrac{g_{ds2}}{G_L}} \cong \frac{g_{m2}}{1 + \dfrac{g_{ds2}}{G_L}} \tag{3.51}$$

Note that the indices are changed to reflect the fact that Q_2 is the common-gate transistor rather than Q_1. Substituting (3.49) into (3.51), and again assuming all elements are somewhat matched so that the indices can be dropped, we have

$$Y_{in2} \cong \frac{g_m}{1 + \dfrac{g_{ds}}{g_{ds}^2/g_m}} \cong g_{ds} \tag{3.52}$$

The gain from the input to the source of Q_2 is therefore given by

$$\frac{v_{s2}}{v_{in}} = -\frac{g_{m1}}{g_{ds1} + Y_{in2}} \cong -\frac{g_m}{2g_{ds}} \tag{3.53}$$

The overall gain can then be found using (3.18) and (3.53). We have

$$A_V = \frac{v_{s2}}{v_{in}}\frac{v_{out}}{v_{s2}} \cong -\frac{g_m}{2g_{ds}}\frac{g_{m2}}{G_L + g_{ds2}} \cong -\frac{g_m}{2g_{ds}}\frac{g_{m2}}{g_{ds2}} \cong -\frac{1}{2}\left(\frac{g_m}{g_{ds}}\right)^2 \tag{3.54}$$

The reader should be cautioned that (3.54) is only approximate, primarily due to the difficulty of accurately determining the output admittance, g_{ds}, for the different transistors. For example, one problem in estimating g_{ds} is that it is voltage dependent. Therefore, prudent designers should never construct a design for which successful operation requires knowing the gain precisely, rather than just knowing that it will be greater than some minimum value.

EXAMPLE 3.6

Assuming g_m is on the order of 0.5 mA/V and r_{ds} is on the order of 100 kΩ, what is the gain of the cascode amplifier?

Solution

Using (3.54), we have $A_V = -1{,}250$. This is a fairly representative number.

3.8 MOS DIFFERENTIAL PAIR AND GAIN STAGE

Most integrated amplifiers have a differential input. To realize this differential input, almost all amplifiers use what is commonly called a differential transistor pair. A differential pair together with a biasing current source is shown in Fig. 3.17. A low-frequency, small-signal model of the differential pair is shown in Fig. 3.18. This small-signal equivalent circuit is based on the T model for a MOS transistor that was described in Chapter 1. To simplify the analysis, we ignore the output impedance of the transistors temporarily. Defining the differential input voltage as $v_{in} \equiv v^+ - v^-$, we have

$$i_{d1} = i_{s1} = \frac{v_{in}}{r_{s1} + r_{s2}} = \frac{v_{in}}{1/g_{m1} + 1/g_{m2}} \tag{3.55}$$

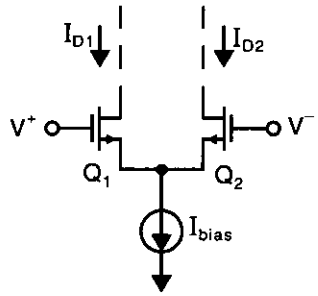


Fig. 3.17 A MOS differential pair.

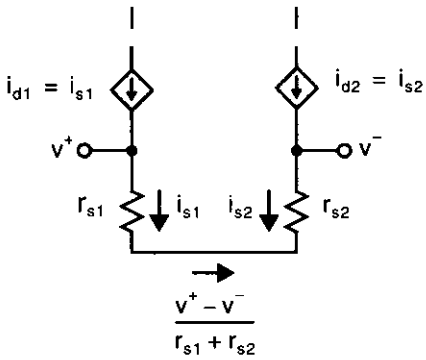


Fig. 3.18 The small-signal model of a MOS differential pair.

Since both Q_1 and Q_2 have the same bias currents, $g_{m1} = g_{m2}$. Therefore, we find

$$i_{d1} = \frac{g_{m1}}{2} v_{in} \tag{3.56}$$

Also, since $i_{d2} = i_{s2} = -i_{d1}$, we find that

$$i_{d2} = -\frac{g_{m1}}{2} v_{in} \tag{3.57}$$

Finally, defining a differential output current, $i_{out} \equiv i_{d1} - i_{d2}$, then the following relationship is obtained:

$$i_{out} = g_{m1} v_{in} \tag{3.58}$$

Thus, if a differential pair has a current mirror as an active load, a complete differential-input, single-ended-output gain stage can be realized, as shown in Fig. 3.19. This circuit is the typical first gain stage of a classical two-stage integrated opamp in which the input differential pair is realized using n-channel transistors and the active current-mirror load is realized using p-channel transistors. From the small-signal analysis of the differential pair, we have

$$i_{d1} = i_{s1} = \frac{g_{m1}}{2} v_{in} \tag{3.59}$$

Also, ignoring transistor output impedances, we have

$$i_{d4} = i_{d3} = -i_{s1} \tag{3.60}$$

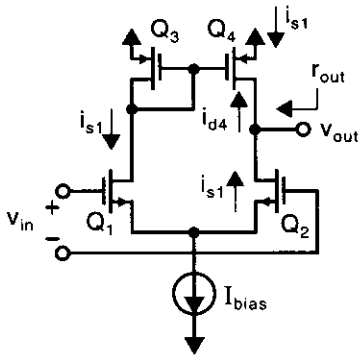


Fig. 3.19 A differential-input, single-ended-output MOS gain stage.

Note that a positive small-signal current is defined as the current going into the drain of a transistor. Using (3.60) and the fact that $i_{d2} = -i_{s1}$, we have

$$v_{out} = (-i_{d2} - i_{d4})r_{out} = 2i_{s1}r_{out} = g_{m1}r_{out}v_{in} \tag{3.61}$$

This result assumes that the output impedance is purely resistive. If there is also a capacitive load, C_L, then the gain is given by

$$A_v = \frac{v_{out}}{v_{in}} = g_{m1}z_{out} \tag{3.62}$$

where $z_{out} = r_{out} \parallel 1/(sC_L)$. Thus, for this differential stage, the very simple model shown in Fig. 3.20 is commonly used. This model implicitly assumes that the time constant at the output node is much larger than the time constant due to the parasitic capacitance at the node at the sources of Q_1 and Q_2. This assumption is usually justified, because the impedance at the output node, r_{out}, is much larger than the impedance at the Q_1, Q_2 source node (i.e., $1/g_{m1} \parallel 1/g_{m2}$). Also, the capacitance at the output node, C_L, is usually larger than the parasitic capacitance at the Q_1, Q_2 source node. However, when high-frequency effects are important (which may be the case when compensating an opamp to guarantee stability), then this assumption may not be justified.

The evaluation of the output resistance, r_{out}, is determined by using the small-signal equivalent circuit and applying a voltage to the output node, as seen in Fig. 3.21. Note that the T model was used for both Q_1 and Q_2, whereas Q_3 was replaced by an equivalent resistance (since it is diode-connected), and the hybrid-π model was used for Q_4.

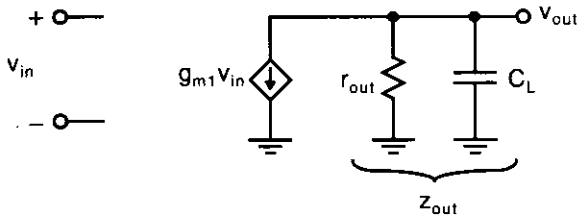


Fig. 3.20 A small-signal model for the differential-input amplifier.

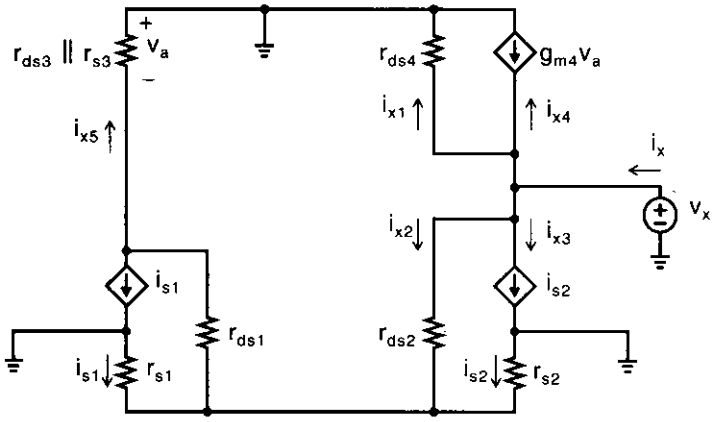


Fig. 3.21 The small-signal model for the calculation of the output impedance of the differential-input, single-ended-output MOS gain stage.

As usual, r_{out} is defined as the ratio v_x/i_x, where i_x is given by the sum $i_x = i_{x1} + i_{x2} + i_{x3} + i_{x4}$. Clearly,

$$i_{x1} = \frac{v_x}{r_{ds4}} \tag{3.63}$$

implying that the resistance seen in the path taken by i_{x1} is equal to r_{ds4}. Now, assuming that the effect of r_{ds1} can be ignored (since it is much larger than r_{s1}), we see that the current i_{x2} is given by

$$i_{x2} \cong \frac{v_x}{r_{ds2} + (r_{s1} \parallel r_{s2})} \cong \frac{v_x}{r_{ds2}} \tag{3.64}$$

where the second approximation is valid, since r_{ds2} is typically much greater than $r_{s1} \parallel r_{s2}$. This i_{x2} current splits equally between i_{s1} and i_{s2} (assuming $r_{s1} = r_{s2}$ and once again ignoring r_{ds1}), resulting in

$$i_{s1} = i_{s2} = \frac{-v_x}{2r_{ds2}} \tag{3.65}$$

However, since the current mirror realized by Q_3 and Q_4 results in $i_{x4} = i_{x5}$ (assuming $g_{m4} = 1/r_{s4} = 1/r_{s3}$ and r_{ds3} is much larger than r_{s3}), the current i_{x4} is given by

$$i_{x4} = -i_{s1} = -i_{s2} = -i_{x3} \tag{3.66}$$

In other words, when the current splits equally between r_{s1} and r_{s2}, the current mirror of Q_3 and Q_4 causes the two currents i_{x3} and i_{x4} to cancel each other. Finally, the output resistance, r_{out}, is given by

$$r_{out} = \frac{v_x}{i_{x1} + i_{x2} + i_{x3} + i_{x4}} = \frac{v_x}{(v_x/r_{ds4}) + (v_x/r_{ds2})} \tag{3.67}$$

which results in the simple relationship

$$r_{out} = r_{ds2} \parallel r_{ds4} \tag{3.68}$$

Therefore, at low frequencies the gain, A_v, is given by

$$A_v = g_{m1}(r_{ds2} \parallel r_{ds4}) \tag{3.69}$$

EXAMPLE 3.7

Consider the gain stage shown in Fig. 3.19, where $I_{bias} = 200\ \mu A$ and all transistors have $W/L = (100\ \mu m)/(1.6\ \mu m)$. Given that $\mu_n C_{ox} = 92\ \mu A/V^2$, $V_{tn} = 0.8$ V, and $r_{ds} = [8{,}000L\ (\mu m)]/[I_D\ (mA)]$, find the output impedance, r_{out}, and the gain from the differential input to the output, V_{out}.

Solution

To find the bias currents, we assume that I_{bias} splits evenly between the two sides of the differential circuit, resulting in

$$I_{D1} = I_{D2} = I_{D3} = I_{D4} = 100\ \mu A \tag{3.70}$$

Therefore, the transconductance of the input transistors is equal to

$$g_{m1} = g_{m2} = \sqrt{2\mu_n C_{ox}(W/L)(I_{bias}/2)} = 1.07\ mA/V \tag{3.71}$$

The output impedance of Q_2 and Q_4 is given by

$$r_{ds2} = r_{ds4} = \frac{8{,}000 \times 1.6}{0.1} = 128\ k\Omega \tag{3.72}$$

Thus, the gain for this stage is

$$A_v \equiv \frac{V_{out}}{V_{in}} = g_{m1}(r_{ds2} \parallel r_{ds4}) = 68.5\ V/V \tag{3.73}$$

3.9 BIPOLAR CURRENT MIRRORS

The most popular bipolar current mirrors are very similar to the MOS current mirrors. A simple bipolar current mirror is shown in Fig. 3.22. This current mirror has an out-

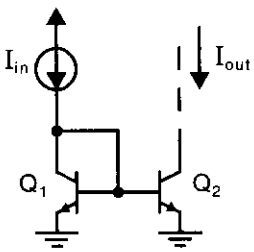


Fig. 3.22 A simple bipolar current mirror.

put current almost equal to its input current. In fact, its output current is slightly smaller than the input current, due to the finite base currents of Q_1 and Q_2. Taking these two base currents into account results in

$$I_{out} = \frac{1}{(1 + 2/\beta)} I_{in} \tag{3.74}$$

where β is the transistor current gain. For large β, this relation is approximately given by $I_{out} \cong (1 - 2/\beta)I_{in}$.

In one commonly used modification of the simple bipolar current mirror, an emitter-follower buffer, Q_3, is added to supply the base currents, as shown in Fig. 3.23. This additional transistor minimizes the errors due to finite base currents, resulting in $I_{out} \cong I_{in}(1 - 2/\beta^2)$. Such an arrangement is almost always used for current mirrors when lateral transistors[4] are used because of their low current gains (i.e., β's on the order of only 10–20). The output impedance of both of the previous current sources is equal to the output impedance of Q_2, which is r_{o2}.

In another often-used variation of the simple current mirror, emitter degeneration is added. This approach results in larger output impedances and also minimizes errors caused by mismatches between Q_1 and Q_2. An example of this current mirror is shown in Fig. 3.24. In an analysis similar to that given for the MOS current mirror with source degeneration, the output impedance is now found to be given by

$$r_{out} \cong r_{o2}(1 + g_{m2}R_e) \tag{3.75}$$

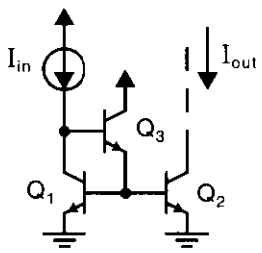


Fig. 3.23 A current mirror with fewer inaccuracies caused by finite base currents.

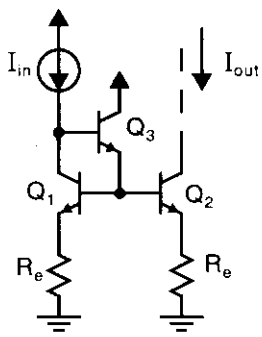


Fig. 3.24 A current mirror with emitter degeneration.

4. In many bipolar processes, pnp transistors are only available as lateral devices.

for $R_e << r_\pi$. Normally, this current mirror is designed so that the bias voltage across R_e, defined to be V_{Re}, is about 0.25 V. This implies that R_e is given by

$$R_e = \frac{V_{Re}}{I_{e2}} \cong \frac{V_{Re}}{I_{c2}} \tag{3.76}$$

Using the relationship $g_{m2} = I_{c2}/V_T$, where $V_T = kT/q \cong 26$ mV at 300 °K, gives

$$r_{out} \cong r_{o2}(1 + g_{m2}R_e) = r_{o2}\left(1 + \frac{V_{Re}}{V_T}\right) \cong 11 r_{o2} \tag{3.77}$$

assuming $V_{Re} = 0.25$ V. It should be mentioned here that the addition of these R_e resistors also minimizes the noise output current of this current mirror, generated by the base resistance thermal noise, which is often the major source of noise in bipolar wideband circuits.

To achieve still higher output impedances, one can use either a cascode or a Wilson current mirror, as shown in Fig. 3.25. The Wilson current mirror is preferred in bipolar realizations because the cascode mirror exhibits large errors, due to the fact that the base currents of all of the transistors are supplied by I_{in} only. As a result, the output current is smaller than the input current by a factor roughly equal to $1 - 4/\beta$. For the Wilson current mirror, the base currents of Q_3 and Q_4 are supplied by I_{in}, whereas the base currents of Q_1 and Q_2 come from I_{out}. It can be shown that the errors due to finite base currents are on the order of $2/\beta^2$ [Gray, 1993]. It can also be shown that both of these current mirrors have an output impedance on the order of [Gray, 1993]

$$r_{out} \cong \frac{\beta r_o}{2} \tag{3.78}$$

This concludes the review of commonly used current mirrors. Chapter 6 describes some improved MOS current mirrors that have become popular recently.

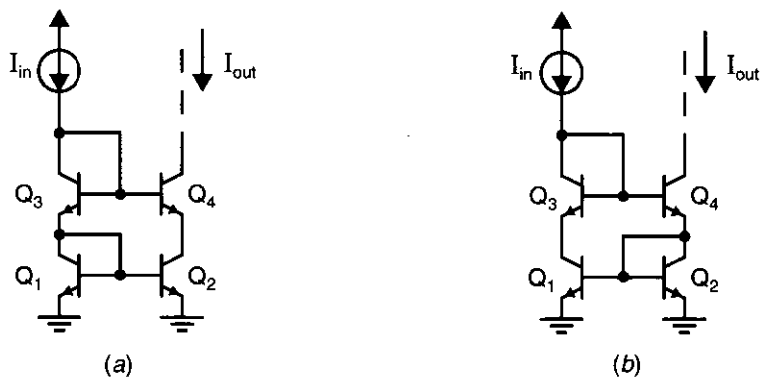


Fig. 3.25 High-output impedance (*a*) cascode and (*b*) Wilson current mirrors.

3.10 BIPOLAR GAIN STAGES

In this section, we look at two bipolar circuits—emitter followers and differential pairs. Although much of the small-signal analyses follow those of their MOS counterparts very closely, one difference is the finite base impedance of bipolar transistors. Here we shall see some simple rules for dealing with this finite base impedance during small-signal analysis. Because of its importance in translinear circuits, we also look at the *large-signal* behavior of the differential pair.

Emitter Follower

A bipolar emitter-follower is very similar to a MOS source follower, except that its input impedance is not infinite and its gain is normally much closer to unity. The analysis of its low-frequency response provides a good illustration of the use of the bipolar T model, and also of some relatively general principles for coming up with quick estimates of input and output impedances of bipolar circuits.

A bipolar emitter follower with a resistive load is shown in Fig. 3.26. Its small-signal model at low frequencies, where the bipolar T model has been used, is shown in Fig. 3.27(*a*). In this small-signal model, the device base resistance r_b has been ignored, since it can be easily taken into account after the fact by simply making R_S slightly larger (i.e., equal to $R_S + r_b$). Notice that R_E and r_o are in parallel. This allows a slightly simplified small-signal model to be analyzed, as shown in Fig. 3.27(*b*), where $R'_E = R_E \parallel r_o$. The analysis of the gain of this circuit is done in two steps. First, the input impedance looking into the base of the transistor, R_b, is found, allowing us to calculate the gain from the input to the base. Second, the gain from the base to the output is found, which allows us to derive the overall gain. The current in the emitter, i_e, is given by

$$i_e = i_b(\beta + 1) \tag{3.79}$$

Therefore,

$$v_b = i_e(r_e + R'_E) = i_b(\beta + 1)(r_e + R'_E) \tag{3.80}$$

This gives

$$R_b = \frac{v_b}{i_b} = (\beta + 1)(r_e + R'_E) \tag{3.81}$$

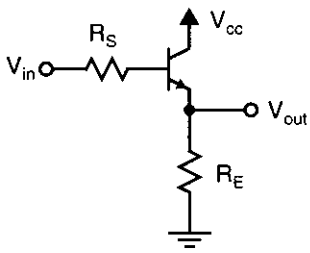


Fig. 3.26 A bipolar emitter follower.

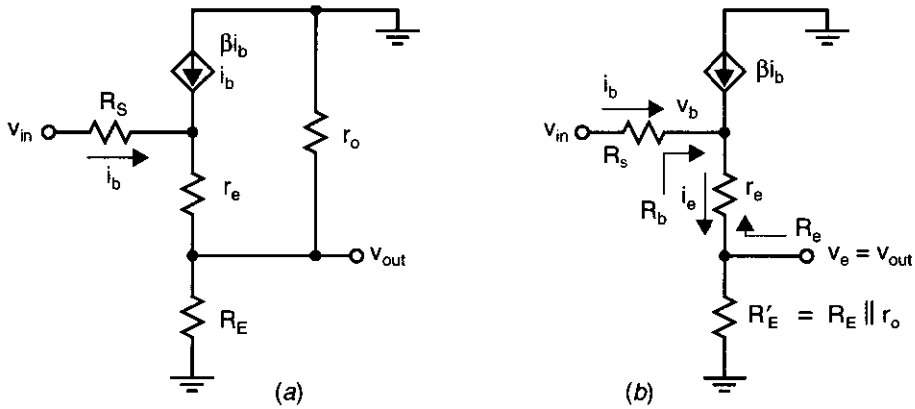


Fig. 3.27 (*a*) The small-signal model for the bipolar emitter follower and (*b*) a simplified model.

Alternatively, noting that

$$(\beta + 1)r_e = (\beta + 1)\frac{\alpha}{g_m} = (\beta + 1)\frac{\beta/(\beta + 1)}{g_m} = \frac{\beta}{g_m} = r_\pi \tag{3.82}$$

gives

$$R_b = r_\pi + (\beta + 1)\, R'_E \tag{3.83}$$

Equations (3.81) and (3.83) illustrate a principle that is generally applicable for bipolar circuits: *At low frequencies, resistances in series with the emitter appear* $\beta + 1$ *times larger when seen looking into the base or, equivalently, when they are reflected into the base.* Continuing, using the resistor-divider formula, we now have

$$\frac{v_b}{v_{in}} = \frac{R_b}{R_b + R_S} = \frac{(\beta + 1)(r_e + R'_E)}{(\beta + 1)(r_e + R'_E) + R_S} \tag{3.84}$$

The gain from the base to the emitter, which is the output, is easily found from Fig. 3.26(*b*), again using the resistor-divider formula, to be given by

$$\frac{v_{out}}{v_b} = \frac{R'_E}{R'_E + r_e} \cong \frac{R'_E}{R'_E + 1/g_m} \tag{3.85}$$

Using (3.84) and (3.85), the overall gain is now given by

$$\frac{v_{out}}{v_{in}} = \frac{v_b}{v_{in}}\frac{v_{out}}{v_b} = \left(\frac{(\beta + 1)(r_e + R'_E)}{(\beta + 1)(r_e + R'_E) + R_S}\right)\left(\frac{R'_E}{R'_E + 1/g_m}\right) \tag{3.86}$$

With a little practice, transfer functions such as that given by (3.86) can be written by simply inspecting the actual circuit, without actually analyzing the small-signal model.

It is also interesting to find the output impedance of the emitter follower at low frequencies excluding R'_E, which is equal to the impedance seen looking into the emitter. The small-signal model for this analysis is shown in Fig. 3.28, where the

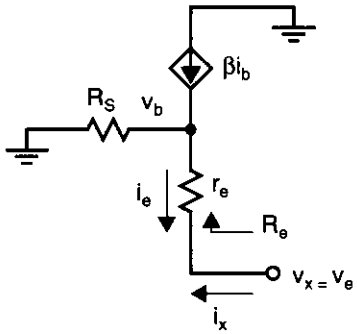


Fig. 3.28 The small-signal model for finding the output impedance of an emitter follower.

input source has been set to 0 and the impedance $R_e = v_x/i_x$ is to be found. First note that

$$i_b = \frac{i_e}{\beta + 1} = \frac{-i_x}{\beta + 1} \tag{3.87}$$

since $i_x = -i_e$. Therefore, we have

$$\begin{aligned} v_x &= v_b + i_x r_e \\ &= -i_b R_S + i_x r_e \\ &= i_x \frac{R_S}{\beta + 1} + i_x r_e \end{aligned} \tag{3.88}$$

This gives the impedance seen looking into the emitter as

$$R_e = \frac{v_x}{i_x} = \frac{R_S}{\beta + 1} + r_e = \frac{R_S}{\beta + 1} + \frac{r_\pi}{\beta + 1} \tag{3.89}$$

This is an example of a general principle for bipolar circuits: *Resistances in series with the base are divided by* $\beta + 1$ *when they are seen looking into the emitter, or, equivalently, are reflected to the emitter.* The total output impedance of the emitter-follower is now simply R_e in parallel with R'_E.

Bipolar Differential Pair—Large-Signal

A bipolar differential pair is shown in Fig. 3.29. This circuit's large-signal behavior can be analyzed by first recalling the exponential relationship for a bipolar transistor,

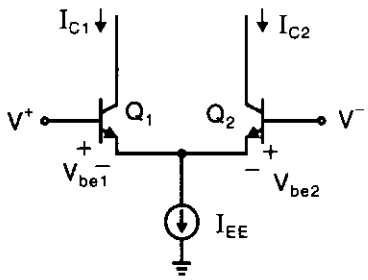


Fig. 3.29 A bipolar differential pair.

$$I_C = I_S e^{(V_{BE}/V_T)} \tag{3.90}$$

which can be used to find the base-emitter voltages

$$V_{BE1} = V_T \ln\left(\frac{I_{C1}}{I_{S1}}\right) \tag{3.91}$$

$$V_{BE2} = V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right) \tag{3.92}$$

Now, writing an equation for the sum of voltages around the loop of input and base-emitter voltages, we have

$$V^+ - V_{BE1} + V_{BE2} - V^- = 0 \tag{3.93}$$

Combining (3.91), (3.92), and (3.93), and assuming $I_{S1} = I_{S2}$, we can write

$$\frac{I_{C1}}{I_{C2}} = e^{(V^+ - V^-)/V_T} = e^{V_{id}/V_T} \tag{3.94}$$

where V_{id} is defined as the difference between V^+ and V^-. In addition, we can write

$$\alpha I_{EE} = I_{C1} + I_{C2} \tag{3.95}$$

where α is defined to be $\beta/(\beta + 1)$ and is due to some currents flowing through the base terminals. Finally, combining (3.94) and (3.95), we have

$$I_{C1} = \frac{\alpha I_{EE}}{1 + e^{-V_{id}/V_T}} \tag{3.96}$$

$$I_{C2} = \frac{\alpha I_{EE}}{1 + e^{V_{id}/V_T}} \tag{3.97}$$

A plot of these two currents with respect to V_{id} is shown in Fig. 3.30, where we note that the currents are split equally at $V_{id} = 0$ and saturate near I_{EE} or 0 for differential input voltages approaching 4 V_T (around 100 mV). Note that the relations shown

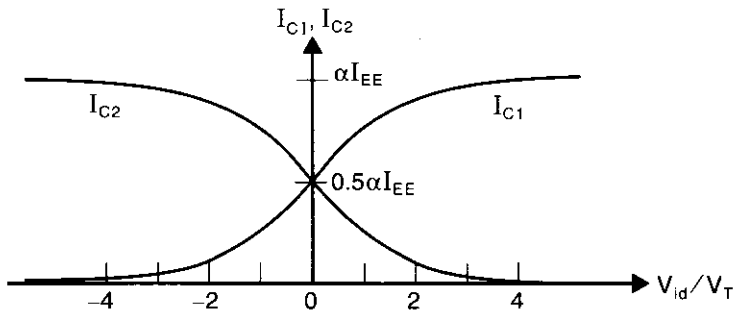


Fig. 3.30 Collector currents for a bipolar differential pair.

in (3.96) and (3.97) are hyperbolic tangent functions, and thus this current–voltage relationship for a bipolar differential is commonly referred to as the *tanh relationship.*

Bipolar Differential Pair—Small-Signal

The small-signal model for the bipolar differential pair of Fig. 3.29 is shown in Fig. 3.31. Once again, defining $v_{id} = v^+ - v^-$, we have

$$i_{c1} = \alpha i_{e1} = \frac{\alpha v_{id}}{r_{e1} + r_{e2}} = \frac{\alpha v_{id}}{(\alpha/g_{m1}) + (\alpha/g_{m2})} \tag{3.98}$$

In the case where both transistors have the same bias currents through them (i.e., the nominal differential voltage is 0, resulting in $g_{m1} = g_{m2}$), we find the same result as for a MOS differential pair,

$$i_{c1} = \frac{g_{m1}}{2} v_{id} \tag{3.99}$$

A similar result holds for i_{c2}.

Finally, the input impedance of this differential pair can be found using the impedance-scaling rule just discussed. Specifically, consider the small-signal model shown in Fig. 3.32, where v^- is grounded and we wish to find the impedance r_{id} looking into the base of Q_1. The impedance, r_2, seen looking into the emitter of Q_2 is simply r_{e2},

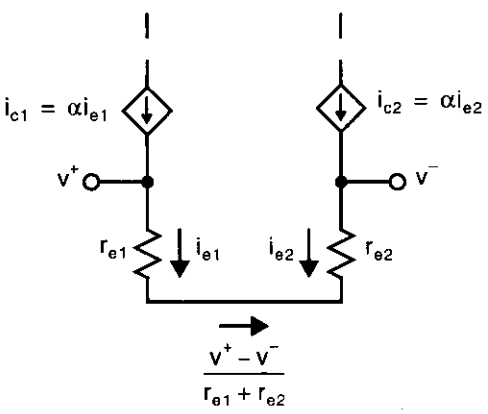


Fig. 3.31 The small-signal model of a bipolar differential pair.

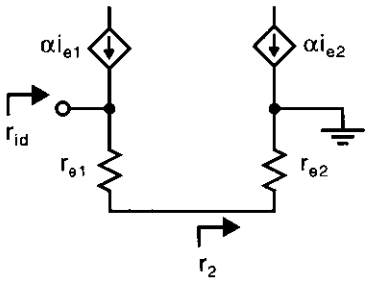


Fig. 3.32 Finding the input impedance of a bipolar differential pair.

since the base is grounded. Thus, the total emitter resistance to the ground of Q_1 is $r_{e1} + r_{e2}$, and using the impedance reflection rule, the impedance, r_{id}, seen looking into the base of Q_1 is equal to $\beta + 1$ times the emitter resistance, or, equivalently,

$$r_{id} = (\beta + 1)(r_{e1} + r_{e2}) \tag{3.100}$$

3.11 FREQUENCY RESPONSE

In this section, we look at the frequency response of many of the previous circuits. Although the precise calculation of frequency responses is most often left to computer simulations, there is much insight that can be obtained by finding the dominant frequency effects in integrated circuits.

Common-Source Amplifier

A small-signal equivalent circuit for high-frequency analysis of the common-source amplifier of Fig. 3.4 is shown in Fig. 3.33. Here, C_{gs1} is the gate-to-source capacitance of Q_1, whereas C_{gd1} is the gate-to-drain capacitance of Q_1. Note that we have assumed that the output capacitance of the input source can be ignored. The capacitance C_2 is made up of the parallel connection of the drain-to-bulk capacitances of Q_1 and Q_2 together with the load capacitance, C_L. Usually, C_L dominates.

To analyze the circuit at high frequencies, one can use nodal analysis. At node v_1, we add all of the currents leaving the node and set the sum equal to zero, to obtain

$$v_1(G_{in} + sC_{gs1} + sC_{gd1}) - v_{in}G_{in} - v_{out}sC_{gd1} = 0 \tag{3.101}$$

where $G_{in} = 1/R_{in}$. Also, at the output node, we have

$$v_{out}(G_2 + sC_{gd1} + sC_2) - v_1 sC_{gd1} + g_{m1}v_1 = 0 \tag{3.102}$$

where $v_1 = v_{gs1}$.

Solving (3.101) and (3.102) (somewhat tediously), we have

$$\frac{v_{out}}{v_{in}} = \frac{-g_{m1}R_2\left(1 - s\dfrac{C_{gd1}}{g_{m1}}\right)}{1 + sa + s^2b} \tag{3.103}$$

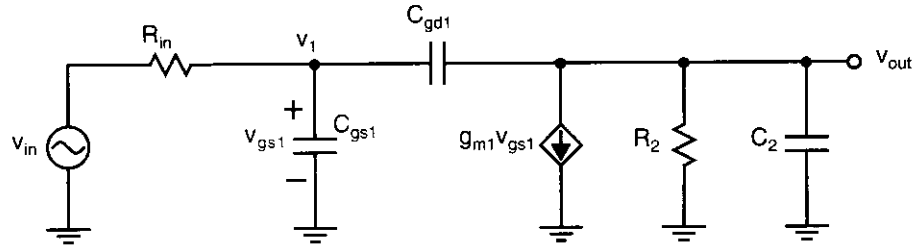


Fig. 3.33 A small-signal model for high-frequency analysis of the common-source amplifier.

where

$$a = R_{in}[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2) \tag{3.104}$$

and

$$b = R_{in}R_2(C_{gd1}C_{gs1} + C_{gs1}C_2 + C_{gd1}C_2) \tag{3.105}$$

At frequencies where the gain has started to decrease but is still much greater than unity, the first-order term in the numerator, $-s(C_{gd1}/g_{m1})$, and the second-order term in the denominator, s^2b, can be ignored. For this case, we have

$$A(s) = \frac{v_{out}}{v_{in}} \cong \frac{-g_{m1}R_2}{1 + s\{R_{in}[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)\}} \tag{3.106}$$

The low-frequency gain is as expected, $-g_{m1}R_2$. Also, setting $s = j\omega_{-3\,dB}$ and solving for

$$|A(j\omega_{-3\,dB})| = \frac{1}{\sqrt{2}} \tag{3.107}$$

gives

$$\omega_{-3\,dB} \cong \frac{1}{R_{in}[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)} \tag{3.108}$$

As an aside, it is of interest to note that the result for the –3-dB frequency is the same result that one would obtain if the zero-value time-constant analysis technique were used [Gray, 1993]. In this technique, one calculates a time constant for each capacitor by assuming all other capacitors are zero, replacing the capacitor in question with a voltage source, and then calculating the resistance seen by that capacitor by taking the ratio of the voltage source to the current flowing from the voltage source. The time constant seen by the capacitor is then simply the capacitor multiplied by the resistance seen by that capacitor. The –3-dB frequency for the complete circuit is then 1 divided by the sum of the individual capacitor time constants. For the common-source amplifier, the resistance seen by C_{gs1} is the input source impedance R_{in}, the resistance seen by C_{gd1} is $R_{in}(1 + g_{m1}R_2) + R_2$, and the resistance seen by C_2 is R_2.

Often, unless $R_{in} << R_2$, the first term in the denominator of (3.108) dominates, and we have

$$\omega_{-3\,dB} \cong \frac{1}{R_{in}[C_{gs1} + C_{gd1}(1 + A)]} \tag{3.109}$$

where $A = g_{m1}R_2$ is the magnitude of the low-frequency gain. The term $C_{gd1}(1 + A)$ is often called the Miller capacitance because it is the equivalent capacitance obtained when one uses the Miller approximation [Sedra, 1991]. Because the size of C_{gd1} is effectively multiplied by one plus the gain of the amplifier, C_{gd1} must be small.

At higher frequencies, when the gain is not much greater than unity, the second pole and the zero should be considered. The frequency for the second pole can be

found by assuming that the poles are real and widely separated and that the denominator can therefore be expressed as

$$D(s) = \left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right) \cong 1 + \frac{s}{\omega_{p1}} + \frac{s^2}{\omega_{p1}\omega_{p2}} \tag{3.110}$$

The coefficients of (3.110) can then be equated to the coefficients of the denominator of (3.103). This analysis is almost identical to that used later in Section 5.2 for analyzing the two-stage amplifier for its frequency response, so it will not be given here. However, the equation for the approximate frequency of the second pole of the denominator is simply given as

$$\omega_{p2} \cong \frac{g_{m1}C_{gd1}}{C_{gs1}C_{gd1} + C_{gs1}C_2 + C_{gd1}C_2} \tag{3.111}$$

It should be mentioned that the Miller approximation results in a very different and incorrect approximation for the frequency of the second pole.

EXAMPLE 3.8

Use the same parameters as in Example 3.2 along with the following: R_{in} = 180 kΩ, C_L = 0.3 pF, C_{gs1} = 0.2 pF, C_{gd1} = 0.015 pF, C_{db1} = 20 fF, and C_{db2} = 36 fF. Estimate the –3-dB frequency of the common-source amplifier in Fig. 3.4.

Solution

We have

$$R_2 = r_{ds1} \parallel r_{ds2} = 77\ \text{k}\Omega \tag{3.112}$$

and

$$C_2 = C_L + C_{db1} + C_{db2} = 0.36\ \text{pF} \tag{3.113}$$

The time constant due to R_{in}, namely, $R_{in}[C_{gs1} + C_{gd1}(1 + A)]$, is now equal to 0.26 μs. The time constant due to R_2, namely, $R_2(C_{gd1} + C_2)$, is equal to 0.03 μs. The –3-dB frequency (in hertz) is equal to

$$\begin{aligned} f_{-3\,dB} &\cong \left[\frac{1}{2\pi}\right][R_{in}[C_{gs1} + C_{gd1}(1 + g_{m1}R_2)] + R_2(C_{gd1} + C_2)]^{-1} \\ &= 550\ \text{kHz} \end{aligned} \tag{3.114}$$

Source-Follower Amplifier

Before proceeding, it should be stated that the high-frequency analysis of source-follower amplifiers is somewhat involved. We will show that these types of amplifiers can have complex poles, and thus a designer should be careful that the circuit

does not exhibit too much overshoot and ringing. We also show a compensation circuit that results in only real axis poles and therefore no overshoot and ringing. However, because the material is complex, we suggest that only advanced readers cover this subsection in detail.

The frequency response of the source follower can be found by modelling the source as a Norton equivalent circuit and adding a load capacitance, as Fig. 3.34 shows. The small-signal model used for this circuit, which includes the parasitic capacitances, is shown in Fig. 3.35. Capacitor C_s includes both the load capacitor, C_L, and the parasitic capacitor, C_{sb1}. Similar to what was done at low frequencies, r_{ds1}, r_{ds2}, and the voltage-controlled current source modelling the body-effect current source can be modelled by a single resistor. This model allows us to analyze the simplified small-signal model shown in Fig. 3.36, where again $R_{s1} = r_{ds1} \| r_{ds2} \| (1)/g_{s1}$ and the input capacitance is given by $C'_{in} = C_{in} + C_{gd1}$.

Nodal analysis is possible, but it is very complicated for this example. The analysis proceeds in four steps. First, the gain from v_{g1} to v_{out} is found. Second, the admittance, Y_g, looking into the gate of Q_1, but not taking into account C_{gd1}, is found. Third, the gain from i_{in} to v_{g1} is found. Finally, the overall gain from v_{in} to v_{out} is found and the results are interpreted.

At node v_{out}, we set the sum of the currents that leave the node to zero, so we have

$$v_{out}(sC_s + sC_{gs1} + G_{s1}) - v_{g1}sC_{gs1} - g_{m1}(v_{g1} - v_{out}) = 0 \tag{3.115}$$

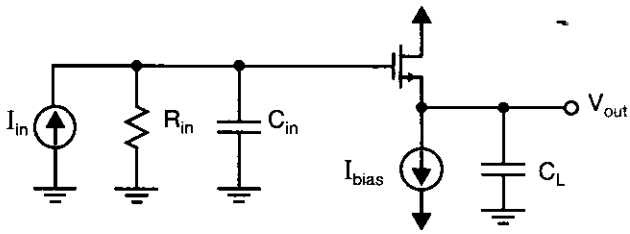


Fig. 3.34 The configuration used to analyze the frequency response of the source follower.

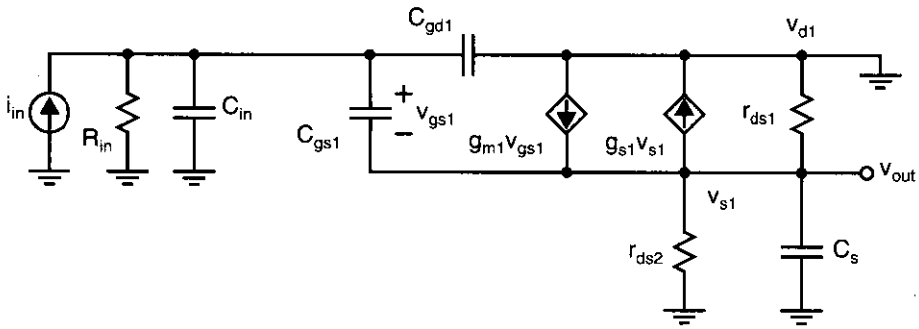


Fig. 3.35 An equivalent small-signal model for the source follower.

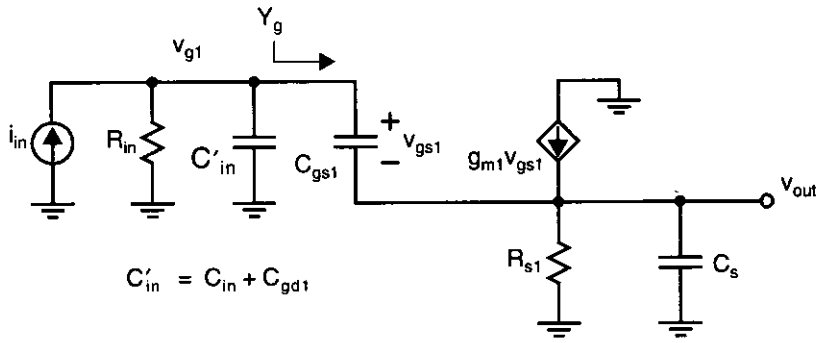


Fig. 3.36 A simplified equivalent small-signal model for the source follower.

Solving for v_{out}/v_{g1}, we have

$$\frac{v_{out}}{v_{g1}} = \frac{sC_{gs1} + g_{m1}}{s(C_{gs1} + C_s) + g_{m1} + G_{s1}} \tag{3.116}$$

The next step is to calculate the admittance, Y_g, looking into the gate of Q_1, but not taking into account the current entering C_{gd1}. The input current is given by

$$i_{g1} = (v_{g1} - v_{out})sC_{gs1} \tag{3.117}$$

Using (3.116) to eliminate v_{out} in (3.117), and solving for $Y_g = i_{g1}/v_{g1}$, we have

$$Y_g = \frac{i_{g1}}{v_{g1}} = \frac{sC_{gs1}(sC_s + G_{s1})}{s(C_{gs1} + C_s) + g_{m1} + G_{s1}} \tag{3.118}$$

We can write an equation relating the input current, i_{in}, to the gate voltage, v_{g1}, as

$$i_{in} = v_{g1}(sC'_{in} + G_{in} + Y_g) \tag{3.119}$$

Substituting (3.118) into (3.119) and rearranging gives

$$\frac{v_{g1}}{i_{in}} = \frac{s(C_{gs1} + C_s) + g_{m1} + G_{s1}}{a + sb + s^2c} \tag{3.120}$$

where

$$\begin{aligned} a &= G_{in}(g_{m1} + G_{s1}) \\ b &= G_{in}(C_{gs1} + C_s) + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1} \\ c &= C_{gs1}C_s + C'_{in}(C_{gs1} + C_s) \end{aligned} \tag{3.121}$$

Using (3.116) and (3.120), we then have

$$A(s) = \frac{v_{out}}{i_{in}} = \frac{sC_{gs1} + g_{m1}}{a + sb + s^2c} \tag{3.122}$$

Thus, we see that the transfer function is second order. Specifically, it has two poles (roots of the denominator) that may be either real or complex conjugate. If they are

complex conjugate, then the step response of the circuit will exhibit overshoot and possibly ringing. This potential problem is a disadvantage when using source followers.

To determine if the transfer function will exhibit ringing, (3.122) can be written in the form

$$A(s) = A(0)\frac{N(s)}{1 + \frac{s}{\omega_0 Q} + \frac{s^2}{\omega_0^2}} \tag{3.123}$$

where ω_0 and Q can be found by equating the coefficients of (3.123) to the coefficients of (3.122). Here, parameter ω_0 is called the *pole frequency* and parameter Q is called the *Q factor*[5] [Sedra, 1991]. It is well known that if $Q < \sqrt{1/2} \cong 0.707$, then the magnitude of the transfer function has its maximum at dc and no peaking will occur (assuming the zero is at a very high frequency and therefore has negligible effect). Furthermore, for $Q = \sqrt{1/2}$, the –3-dB frequency is equal to ω_0. When the time-domain response is investigated, restrictions on the Q factor can also be found to guarantee no peaking for a step input. Specifically, to have no peaking in the step response, both poles must be real; this is equivalent to the requirement that $Q \leq 0.5$. In the case where $Q > 0.5$, the percentage overshoot of the output voltage can be shown to be given by

$$\%\ \text{overshoot} = 100e^{-\pi/\sqrt{4Q^2-1}} \tag{3.124}$$

For the source follower, equating the coefficients of (3.123) to the coefficients of (3.122) and solving for ω_0 and Q results in

$$\omega_0 = \sqrt{\frac{G_{in}(g_{m1} + G_{s1})}{C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)}} \tag{3.125}$$

$$Q = \frac{\sqrt{G_{in}(g_{m1} + G_{s1})[C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)]}}{G_{in}C_s + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1}} \tag{3.126}$$

If Q is greater than 0.5, the poles will be complex conjugate, and the circuit will exhibit overshoot. Although this Q equation is rather complex, note that if C_s, C'_{in}, or both become large (i.e., if the load, the input capacitor, or both become large), then Q becomes small, and no overshoot will occur (although the circuit will be slow). And when C'_{in} and G_{s1} become small (G_{s1} becomes small when the transistor's source is connected to the substrate, which eliminates the body effect), then the circuit will have a large Q (i.e., large ringing) when G_{in} becomes small and $C_s \cong C_{gs1}$. *In summary, source follower (and emitter follower) circuits can exhibit large amounts of overshoot and ringing under certain conditions.* Fortunately, the parasitic capacitances and output impedances in practical microcircuits typically result in only moderate overshoot for worst-case conditions.

5. The Q factor is 1/2 times the inverse of the damping factor. The damping factor is an alternative method of indicating the pole locations in second-order transfer functions.

Finally, note also that the numerator zero of the transfer function lies on the negative real axis at a frequency given by

$$\omega_z = \frac{-g_{m1}}{C_{gs1}} \tag{3.127}$$

and is typically at a much higher frequency than ω_0.

EXAMPLE 3.9

Use the parameters in Example 3.3 and assume that $R_{in} = 180\ k\Omega$, $C_L = 10$ pF, $C_{gs1} = 0.2$ pF, $C_{gd1} = 15$ fF, $C_{sb1} = 40$ fF, and $C_{in} = 30$ fF. Find ω_0, Q, and the frequency of the zero for the source follower in Fig. 3.34.

Solution

From Example 3.3, we have $g_{m1} = 1.06$ mA/V, $r_{ds1} = 128\ k\Omega$, $r_{ds2} = 128\ k\Omega$, and $g_{s1} = 0.16$ mA/V. Thus, we have

$$C'_{in} = C_{in} + C_{gd1} = 45 \text{ fF} \tag{3.128}$$

$$G_{s1} = g_{s1} + g_{ds1} + g_{ds2} = 0.176 \text{ mA/V} \tag{3.129}$$

$$C_s = C_L + C_{sb1} = 10.04 \text{ pF} \tag{3.130}$$

and so we can find ω_0 as

$$\omega_0 = \sqrt{\frac{G_{in}(g_{m1} + G_{s1})}{C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)}}$$

$$= 5.24 \times 10^7 \text{ rad/s} = 2\pi \times 8.34 \text{ MHz} \tag{3.131}$$

$$Q = \frac{\sqrt{G_{in}(g_{m1} + G_{s1})[C_{gs1}C_s + C'_{in}(C_{gs1} + C_s)]}}{G_{in}C_s + C'_{in}(g_{m1} + G_{s1}) + C_{gs1}G_{s1}} \tag{3.132}$$

$$= 0.8$$

This results in an overshoot for a step input given by

$$\%\text{ overshoot} = 100e^{-\pi/\sqrt{4Q^2-1}} = 8\% \tag{3.133}$$

The zero frequency is found using (3.127) to be 844 MHz, and thus it can almost certainly be ignored.

A bipolar-transistor emitter follower is very similar to a CMOS source follower, and therefore its high-frequency analysis is not included here. However, note that the problem of complex-conjugate poles can be more severe for bipolar emitter followers.

For both CMOS source followers and bipolar emitter followers, when complex-conjugate poles occur, they can be eliminated by adding a compensation network. To see this, note that (3.118) can be rewritten as

$$Y_g = sC_2 + \frac{1}{-R_1 - \frac{1}{sC_1}} \tag{3.134}$$

where

$$\begin{aligned} C_1 &= \frac{C_{gs1}(C_s g_{m1} - C_{gs1}G_{s1})}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} \cong \frac{g_{m1}C_{gs1}C_s}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} \\ R_1 &= \frac{(C_{gs1} + C_s)^2}{C_{gs1}(C_s g_{m1} - C_{gs1}G_{s1})} \cong \frac{(C_{gs1} + C_s)^2}{C_{gs1}C_s g_{m1}} \\ C_2 &= \frac{C_{gs1}C_s}{C_{gs1} + C_s} \end{aligned} \tag{3.135}$$

and the approximations result from the fact that typically $C_s > C_{gs1}$ and $g_{m1} > G_{s1}$. This is the same admittance as the circuit shown in Fig. 3.37. Thus, the input admittance is the same as a capacitor in parallel with a series combination of a negative capacitor and a negative resistor. If a third network consisting of a capacitor of size C_1 and a resistor of size R_1, in series, is connected to the gate of the source follower, as shown in Fig. 3.38, then the negative elements are cancelled. The resulting input admittance is then simply C_2, as given in (3.135). In this case, (3.120) becomes

$$\frac{v_{g1}}{i_{in}} = \frac{1}{G_{in} + s\left(C'_{in} + \frac{C_{gs1}C_s}{C_{gs1} + C_s}\right)} \tag{3.136}$$

and (3.122) becomes

$$A(s) = \frac{v_{out}}{i_{in}} = R_{in}\left(\frac{g_{m1}}{g_{m1} + G_{s1}}\right)\frac{\left(1 + s\frac{C_{gs1}}{g_{m1}}\right)}{\left(1 + \frac{s}{p_1}\right)\left(1 + \frac{s}{p_2}\right)} \tag{3.137}$$

where

$$p_1 = \frac{G_{in}}{C'_{in} + \frac{C_{gs1}C_L}{C_{gs1} + C_L}} \cong \frac{G_{in}}{C'_{in} + C_{gs1}} \tag{3.138}$$

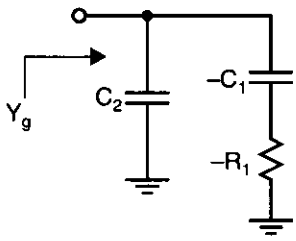


Fig. 3.37 A circuit that has the same admittance as the input impedance looking into the gate of a source follower (and ignoring C_{gd}).

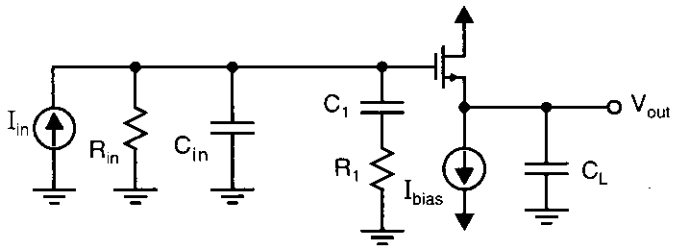


Fig. 3.38 Addition of a compensation network (C_1 and R_1) to compensate for the negative components of the admittance looking into the gate of the source follower.

$$p_2 = \frac{g_{m1} + G_{s1}}{C_{gs} + C_L} \cong \frac{g_{m1} + G_{s1}}{C_L} \tag{3.139}$$

The approximation is accurate when $C_s >> C_{gs1}$. Regardless of the approximation, the poles are now guaranteed real, and no overshoot will occur.

Therefore, when designing source followers (or emitter followers), the recommended procedure is to check to see if the poles are complex by using either (3.126) or a SPICE transient analysis to look for overshoot. When the poles are complex, then increase either C_{in}, C_s, or both, or, alternatively, add the compensation network, as shown Fig. 3.38.

EXAMPLE 3.10

Using the same parameters as in Example 3.9, find the compensation network and the resulting first and second poles of the source follower in Fig. 3.34.

Solution

Using (3.135), we have

$$C_1 \cong \frac{g_{m1}C_{gs1}C_s}{(g_{m1} + G_{s1})(C_{gs1} + C_s)} = 0.170 \text{ pF} \tag{3.140}$$

and

$$R_1 \cong \frac{(C_{gs1} + C_s)^2}{C_{gs1}C_s g_{m1}} \cong 49.3 \text{ k}\Omega \tag{3.141}$$

The capacitor is a reasonable value to be realized on chip. The resistor can be realized by a MOS transistor biased in the triode (i.e., linear) region. Assuming the compensation network is used, the poles of the transfer function then become

$$p_1 \cong \frac{G_{in}}{C_{gs1} + C'_{in}} = 2\pi \times 3.61 \text{ MHz} \tag{3.142}$$

and

$$p_2 = \frac{g_{m1} + G_{s1}}{C_{gs} + C_L} = 2\pi \times 19.3 \text{ MHz} \tag{3.143}$$

The speed penalty paid for using the compensation network is quite high, because the pole frequency without compensation was around 8 MHz whereas here the dominant pole is at 3.6 MHz.

Finally, it should be mentioned that if the source-follower buffer is intended to be used in an opamp (and thus feedback will be placed around the buffer), and if the resonant frequency of the source follower is substantially greater than the unity-gain frequency of the amplifier, then the overshoot can be tolerated, and no compensation network is necessary.

Common-Gate Amplifier

The frequency response of the common-gate stage is usually superior to that of the common-source stage due to the low impedance, r_{in}, at the source node, assuming G_L is not considerably smaller than g_{ds1}. Analysis of the frequency response of the common-gate stage is left as an exercise for the reader.

High-Output Impedance Mirrors

Both the Wilson and the cascode current mirrors introduce high-frequency poles into the signal transfer function. The approximately equivalent time constant of these poles is C_{gs}/g_m. The proof of this statement can been found by doing a high-frequency, small-signal analysis (where the capacitances of the small-signal models are included in the analysis). This analysis is left as an exercise for the reader.

Cascode Gain Stage

The exact high-frequency analysis of a cascode gain stage (Fig. 3.16) is usually left to simulation on a computer; however, an approximate analysis is not too complicated. At high frequencies, the time constant due to the output node almost always dominates since the impedance is so large at that node. The total capacitance at the output node, C_{out}, is the parallel combination of $C_{gd2} + C_{db2}$, the load capacitance C_L, and the output capacitance of the bias current source, C_{bias} (normally, C_L is the major contributor). Assuming that the time constant at the output stage dominates, the –3-dB frequency is approximately equal to the inverse of the time constant. In other words, we have

$$\omega_{-3\text{ dB}} \cong \frac{1}{R_{out}C_L} \cong \frac{2g_{ds}^2}{g_m C_L} \tag{3.144}$$

A more accurate (though still not exact) estimate may be found using the zero-value time-constant analysis method of [Gray, 1993]. The advantage of this estimation technique is that it gives some insight into the relative importance of each capacitor in determining the overall –3-dB frequency. The small-signal model being analyzed is shown in Fig. 3.39, where

$$\begin{aligned} C_{s2} &= C_{db1} + C_{sb2} + C_{gs2} \\ C_{d2} &= C_{gd2} + C_{db2} + C_L + C_{bias} \end{aligned} \tag{3.145}$$

In the zero-value time-constant analysis, all independent sources are set to zero (i.e., here, V_{in} is set to 0 V) and each capacitor is considered in turn with all other capacitors set to zero. The corresponding time constant is found and labeled τ_{Ci}. Then the –3-dB frequency, $\omega_{-3\,dB}$, is estimated to be 1 divided by the sum of all the time constants.

In this circuit, the first time constant found is the one that corresponds to C_{gs1}, and it is labelled τ_{Cgs1}. The resistance seen by C_{gs1} is R_{in}, and therefore,

$$\tau_{Cgs1} = C_{gs1}R_{in} \tag{3.146}$$

The calculation of the time constant corresponding to C_{gd1} is more involved, so formal methods are used in its calculation. C_{gd1} is replaced by a voltage source, V_x. Next, the resistance seen by C_{gd1} is found by calculating the ratio of V_x to i_x (the current leaving V_x). The final time constant is then given by this resistance multiplied by C_{gd1}. The small-signal model for this analysis is shown in Fig. 3.40(*a*), where the resistance R_{d1} is the parallel combination of r_{ds1} and the impedance seen looking into the source of Q_2 (the cascode transistor) at low frequencies. The circuit of Fig. 3.40(*a*) can be redrawn as the equivalent circuit shown in Fig. 3.40(*b*). In this transformation, we changed the ground node (which node is called ground is arbitrary in an analysis) and the direction of the voltage-controlled current source. Notice that the circuit of Fig. 3.40(*b*) is essentially the same as the circuit of Fig. 3.13, which was used for finding the output impedance of a source-degenerated current source. The analysis here is essentially the same as that used in Fig. 3.13, but will be included here for clarity. We have

$$v_y = i_x R_{in} \tag{3.147}$$

Also,

$$i_x = (v_x - v_y)G_{d1} - g_{m1}v_y \tag{3.148}$$

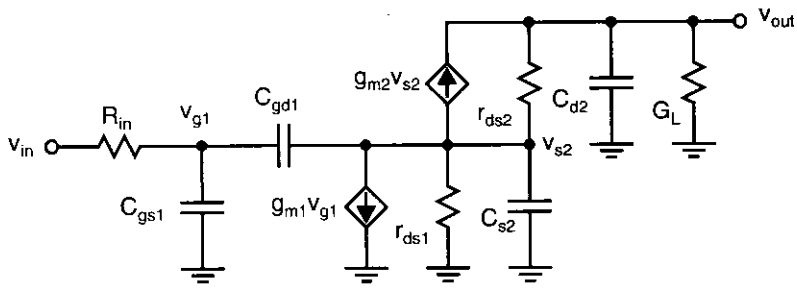


Fig. 3.39 The small-signal model of the cascode gain stage.

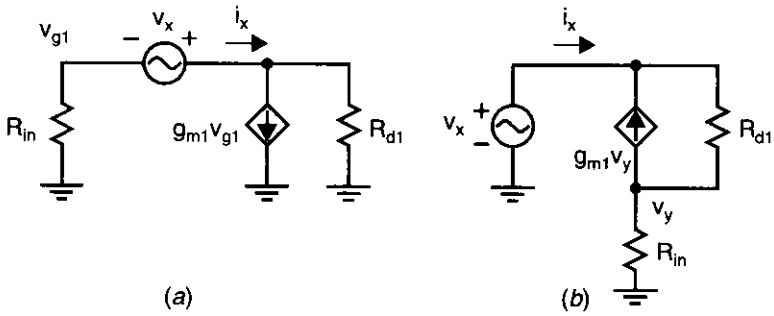


Fig. 3.40 Two equivalent small-signal models for calculating the resistance seen by C_{gd1}.

Substituting (3.147) into (3.148) and solving for v_x/i_x gives

$$r_{Cgd1} = \frac{v_x}{i_x} = R_{d1}[1 + R_{in}(G_{d1} + g_{m1})] \tag{3.149}$$

The admittance looking into the source of the cascode transistor, Q_2, was found previously in (3.52). Namely,

$$Y_{s2} \cong g_{ds} \tag{3.150}$$

The impedance, R_{d1}, is the parallel combination of this admittance and r_{ds1}. Therefore, we have

$$R_{d1} \cong \frac{r_{ds}}{2} \tag{3.151}$$

Substituting this result into (3.149), we have

$$\begin{aligned} r_{Cgd1} &\cong \frac{r_{ds}}{2}[1 + R_{in}(2g_{ds} + g_{m1})] \\ &\cong \frac{r_{ds}}{2}(1 + g_{m1}R_{in}) \end{aligned} \tag{3.152}$$

and therefore

$$\tau_{Cgd1} \cong C_{gd1}\frac{r_{ds}}{2}(1 + g_{m1}R_{in}) \tag{3.153}$$

If R_{in} is large, say, on the order of a transistor output impedance, r_{ds}, then this time constant is approximately given by

$$\tau_{Cgd1} \cong C_{gd1}\frac{g_m r_{ds}^2}{2} \tag{3.154}$$

This time constant can be almost as large as the corresponding time constant for a common-source amplifier—a fact not well known.

The resistance seen by capacitor C_{s2} is r_{ds1} in parallel with the impedance seen looking into the source of Q_2, which, from (3.52), is approximately r_{ds}. Thus, we have

$$\tau_{Cs2} \cong C_{s2}\frac{r_{ds}}{2} \tag{3.155}$$

The resistance seen by C_{d2} is the output impedance of the cascode amplifier, which is approximately given by $(g_m r^2_{ds})/2$ from (3.50). Thus, the time constant due to C_{d2} is given by

$$\tau_{Cd2} \cong C_{d2}\frac{g_m r_{ds}^2}{2} \tag{3.156}$$

Note that this time constant has the same form as (3.154), but normally C_{d2} will be much larger than C_{gd1} (because C_L is often large), making τ_{Cd2} dominate.

The sum of the time constants is then given by

$$\begin{aligned}\tau_{total} &\cong \tau_{Cgs1} + \tau_{Cgd1} + \tau_{Cs2} + \tau_{Cd2}\\ &\cong C_{gs1}R_{in} + C_{gd1}\frac{g_m r_{ds}^2}{2} + C_{s2}\frac{r_{ds}}{2} + C_{d2}\frac{g_m r_{ds}^2}{2}\end{aligned} \tag{3.157}$$

The –3-dB frequency, $\omega_{-3\,dB}$, is estimated to be $1/\tau_{total}$.

EXAMPLE 3.11

Assume that for both the input transistor and the cascode transistor, g_m = 1 mA/V, r_{ds} = 100 kΩ, R_{in} = 180 kΩ, C_L = 5 pF, C_{gs} = 0.2 pF, C_{gd} = 15 fF, C_{sb} = 40 fF, C_{db} = 20 fF, and C_{bias} = 20 fF. Estimate the –3-dB frequency of the cascode amplifier of Fig. 3.16(*a*).

Solution

The time constants associated with each capacitor are readily evaluated using (3.157). First, note that

$$\begin{aligned}C_{s2} &= C_{db1} + C_{sb2} + C_{gs2} = 0.26\text{ pF}\\ C_{d2} &= C_{gd2} + C_{db2} + C_L + C_{bias} = 5.055\text{ pF}\end{aligned} \tag{3.158}$$

We have

$$\begin{aligned}\tau_{Cgs1} &= C_{gs1}R_{in} = 36\text{ ns}\\ \tau_{Cgd1} &= C_{gd1}\frac{g_m r_{ds}^2}{2} = 75\text{ ns}\\ \tau_{Cs2} &= C_{s2}\frac{r_{ds}}{2} = 13\text{ ns}\\ \tau_{Cd2} &= C_{d2}\frac{g_m r_{ds}^2}{2} = 25.3\ \mu\text{s}\end{aligned} \tag{3.159}$$

As expected, the time constant at the output node dominates, and the second most important time constant is that due to C_{gd1}, although its effect on the –3-dB frequency is negligible. Therefore, the –3-dB frequency is accurately given by $\omega_{-3\,dB} \cong 1/\tau_{Cd2} = 2\pi \times 6.3\,\text{kHz}$.

Before leaving this section, some comments should be made about the high-frequency operation of cascode-gain stages. As we just saw, typically one pole dominates, and thus we can reasonably model the amplifier gain as

$$A(s) = \frac{A_v}{1 + s/\omega_{-3\,dB}} \tag{3.160}$$

Thus, at frequencies substantially larger than $\omega_{-3\,dB}$, which is usually the frequency band of operation, the gain is approximately given by

$$A(s) \cong \frac{A_v}{s/\omega_{-3\,dB}} \cong -\frac{g_{m1}}{sC_L} \tag{3.161}$$

using (3.54) and (3.144). It should also be noted that the approximations of (3.144) and (3.161) are quite good, unless either the source impedance or source capacitance is very large. In addition, at frequencies much larger than the –3-dB frequency, the admittance at the source of Q_2 can be found by using (3.51), where G_L is replaced by $G_L + sC_L$. Such a substitution results in

$$\begin{aligned} Y_{in2} &= \frac{g_{m2}}{1 + \dfrac{g_{ds2}}{G_L + sC_L}} \\ &= g_{m2}\left(\frac{G_L + sC_L}{g_{ds2} + G_L + sC_L}\right) \\ &\cong g_{m2}\left(\frac{G_L + sC_L}{g_{ds2} + sC_L}\right) \end{aligned} \tag{3.162}$$

At frequencies where $\omega >> 1/(r_{ds}C_L)$, the terms in s dominate, and $Y_{in2} \cong g_{m2}$. The approximate time constant due to the node at the source of Q_2 is then simply given by the total capacitance at that node divided by g_{m2}. The total capacitance at the source of Q_2 is C_{gs2} in parallel with C_{db1} in parallel with C_{gd1}. Because this capacitance is not excessively large and the impedance at the node, $1/g_{m2}$, is small, the time constant at that node can usually be neglected. However, in amplifiers with a small source impedance, this node would still be the primary factor determining the second pole of a cascode amplifier. It is easy to derive an upper bound on the time constant at the source of Q_2. Almost always, C_{db1} in parallel with C_{gd1} is less than C_{gs2}. Therefore, the total capacitance at the source of Q_2 is equal to KC_{gs2}, where K is between 1 and 2 (usually closer to 1). Using

$$g_{m2} = \mu_p C_{ox}\left(\frac{W}{L}\right)_2 V_{eff2} \tag{3.163}$$

for the folded-cascode amplifier (for the telescopic-cascode amplifier, substitute μ_n for μ_p), and using

$$C_{S2} = KC_{gs2} = K\frac{2}{3}(WL)_2 C_{ox} \tag{3.164}$$

gives the approximate frequency of the second pole (ignoring the time constants at other nondominant nodes):

$$\omega_{p2} \cong \frac{1}{\tau_{S2}} = \frac{g_{m2}}{C_{S2}} = \frac{3\mu_p V_{eff2}}{2KL_2^2} > \frac{3\mu_p V_{eff2}}{4L_2^2} \tag{3.165}$$

This equation is an upper limit on the unity-gain frequency of any amplifier that uses a cascode gain stage. Note that (3.165) is relatively independent of that actual design once V_{eff2} is chosen, and V_{eff2} is usually determined by maximum signal-handling requirements. Also note that ω_{p2} has a very strong dependence on the channel length.

EXAMPLE 3.12

Estimate the lower bound on the frequency of the second pole of a folded-cascode amplifier for a 0.8-μm technology, where a typical value of 0.25 V is chosen for V_{eff2}.

Solution

Normally, the minimum length of a cascode transistor in an analog circuit might be 25 to 50 percent of the minimum length of transistors used in digital circuits. Therefore, assuming $L_2 = 1.5 \cdot 0.8\ \mu m = 1.2\ \mu m$, and using $\mu_p = 0.02\ m^2/V \cdot s$, and $V_{eff2} = 0.25$ V, we have $\omega_{p2} > 1.7 \times 10^9$ rad $= 2\pi \cdot 276$ MHz. For a telescopic-cascode amplifier, the upper bound would be 690 MHz. In most practical opamp designs, the unity-gain frequency of a typical design might be limited to around one-half the frequency of the lower bound second pole. In this example, the typical unity-gain frequencies would be 138 MHz and 345 MHz for the folded-cascode and telescopic-cascode amplifiers, respectively.

Finally, the time constant at the gate of the input transistor can be important when the source resistance is large, though not as important as in a common-source stage, since a cascode gain-stage does not suffer as much from a *Miller effect*. In other words, at high frequencies, the effective impedance at the source of Q_2 has decreased to $1/g_{m2}$, and there is not much gain from the gate of Q_1 to the source of Q_2. Recall that in a common-source amplifier, the effective size of C_{gd1} was magnified by the gain of the common-source amplifier.

3.12 SPICE SIMULATION EXAMPLES

In this section, circuit diagrams, netlists, and simulation results for selected examples in this chapter are presented. Each of the netlists contains the circuit elements and commands used in the simulations, but does not contain transistor model parameters.

Simulation of Example 3.2

The circuit for Example 3.2 is shown in Fig. 3.41 where a 5-V power supply is used. The bias voltage, V_{in}, has to be set such that both Q_1 and Q_2 are in the active region, and approximately 2.5 V appears across the drain-source nodes for each of them.

```
NETLIST:
vdd    1    0                    dc 5
ibias  2    0                    dc 100u
m3     2    2    1    1          pmos w=100u l=1.6u
m2     3    2    1    1          pmos w=100u l=1.6u
m1     3    4    0    0          nmos w=100u l=1.6u
vin    4    0                    dc 0.849 ac 1

.op
.ac dec 10 1k 10000Meg
.print vdb(3)
.option post
.lib "../cmos_models" library
.end
```

The frequency plot for this simulation is shown in Fig. 3.42. Note that the gain is 36 dB, which corresponds to approximately 63 V/V.

Simulation of Example 3.3

The circuit for Example 3.3 is shown in Fig. 3.43, where it should be noted that the body of Q_1 is tied to ground, rather than to its source.

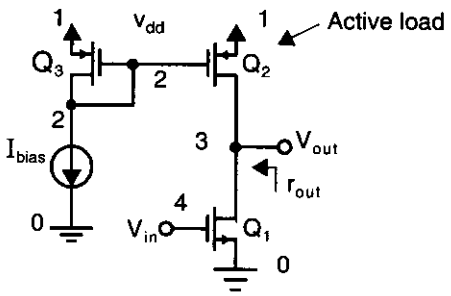


Fig. 3.41 Circuit for Example 3.2.

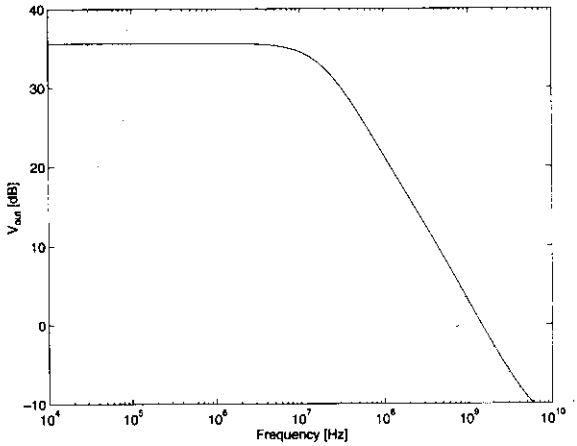


Fig. 3.42 Frequency simulation result for the common-source amplifier.

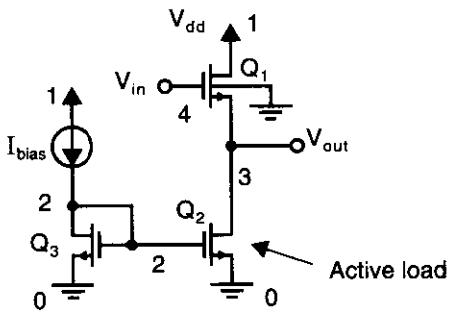


Fig. 3.43 Circuit for Example 3.3.

```
NETLIST:
vdd     1   0                   dc 5
ibias   1   2                   dc 100u
m3      2   2   0   0           nmos w=100u l=1.6u
m2      3   2   0   0           nmos w=100u l=1.6u
m1      1   4   3   0           nmos w=100u l=1.6u
vin     4   0                   dc 2 ac 1

.op
.ac dec 10 1k 1000Meg
.print vdb(3)
.option post
.lib "../cmos_models" library
.end
```

The frequency plot for this simulation is shown in Fig. 3.44. The gain is –1.3 dB, which corresponds to 0.86 V/V.

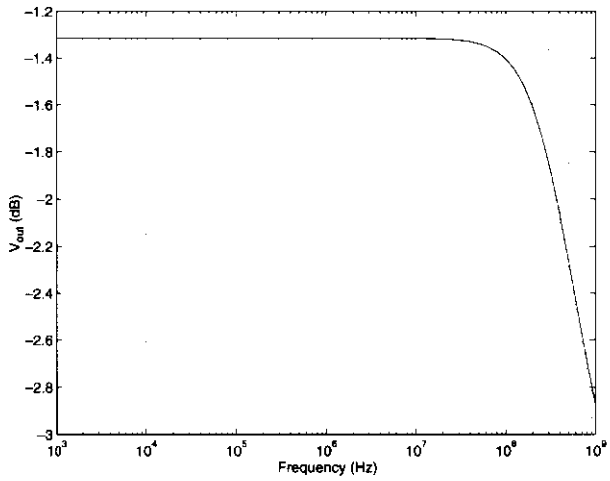


Fig. 3.44 Frequency plot for the source-follower circuit.

Simulation of Example 3.8

The circuit for Example 3.8 is shown in Fig. 3.45.

```
NETLIST:
vdd     1   0                   dc 5
ibias   2   0                   dc 100u
m3      2   2   1   1           pmos w=100u l=1.6u
m2      3   2   1   1           pmos w=100u l=1.6u
m1      3   4   0   0           nmos w=100u l=1.6u
rin     5   4                   180k
vin     5   0                   dc 0.849 ac 1
cl      3   0                   0.3p

.op
.ac dec 20 1k 100Meg
.print vdb(3)
```

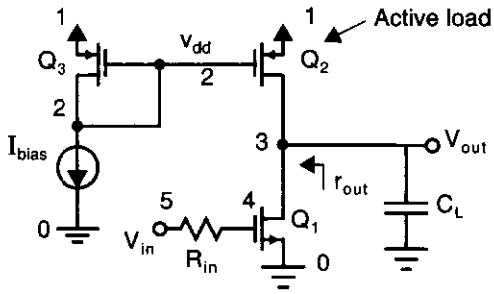


Fig. 3.45 Circuit for Example 3.8.

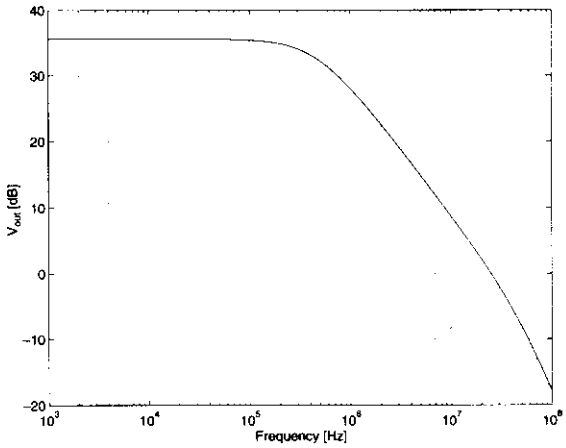


Fig. 3.46 Estimate of the –3-dB frequency for a common-source amplifier.

```
.option post
.lib "../cmos_models" library
.end
```

The frequency plot for this simulation is shown in Fig. 3.46, where we see that the –3-dB frequency occurs around 460 kHz.

Simulation of Example 3.9

The circuit for Example 3.9 is shown in Fig. 3.47.

```
NETLIST:
vdd     1   0     dc 5
vss     2   0     dc -5
ibias   3   2     dc 100u
rin     4   0     180k
```

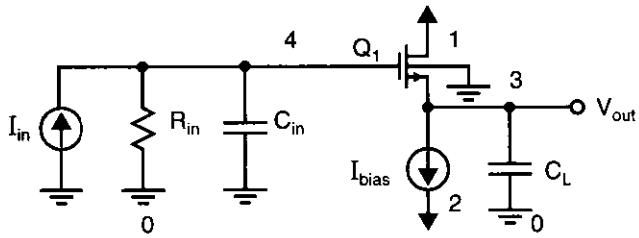


Fig. 3.47 Circuit for Example 3.9.

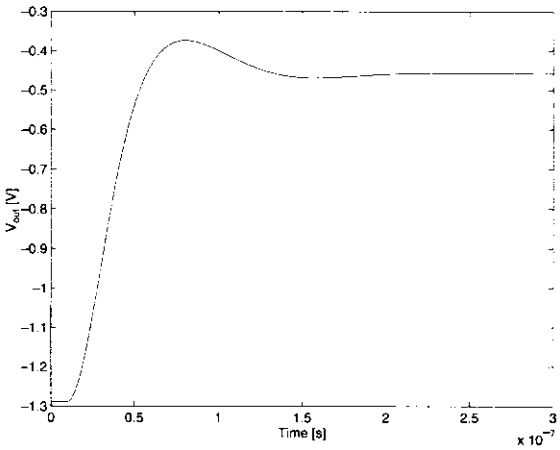


Fig. 3.48 The step response of a source follower, showing a 10-percent overshoot.

```
cin     4   0                   30f
cl      3   0                   10p
m1      1   4     3     2       nmos w=100u l=1.6u
iin     4   0                   pulse (0 -5u 10n 0 0)

.op
.option post
.tran 0.5n 300n
.print v(3)
.lib "../cmos_models" library
.end
```

The step response of the source follower is shown in Fig. 3.48. The overshoot here is about 10 percent.

Simulation of Example 3.10

The circuit for Example 3.10 is shown in Fig. 3.49. The compensation network that eliminates the complex conjugate poles is included. The simulation is used to obtain the frequency plot from which the first and second poles of the resulting source follower are determined.

```
NETLIST:
vdd     1   0           dc 5
vss     2   0           dc -5
ibias   3   2           dc 100u
```

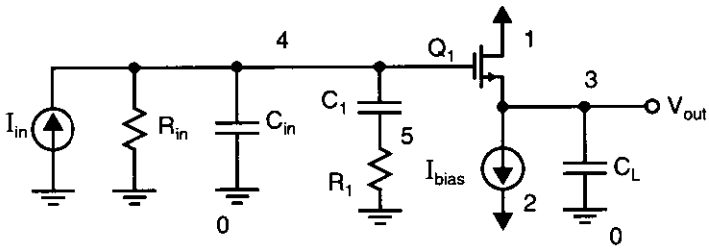


Fig. 3.49 Circuit for Example 3.10.

```
rin     4   0                   180k
cin     4   0                   30f
cl      3   0                   10p
m1      1   4   3     2         nmos w=100u l=1.6u
iin     4   0                   dc 0 ac 1
c1      4   5                   0.17p
r1      5   0                   49.3k

.op
.print vdb(3)
.lib "../cmos_models" library
.end
```

The frequency plot of this source follower is shown in Fig. 3.50. The first pole occurs around 3.6 MHz, whereas the second pole occurs around 16 MHz.

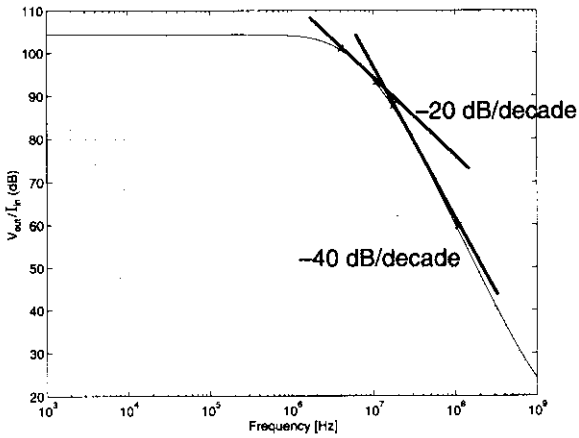


Fig. 3.50 The bode plot of a source follower with a compensation network.

Simulation of Example 3.11

The circuit for Example 3.11 is shown in Fig. 3.51. Notice that Q_3 through Q_6 form a cascode current mirror to produce I_{bias}. The width and length of the p-channel transistors are chosen such that their g_m and r_{ds} are matched to those of the n-channel transistors.

```
NETLIST:
vdd     1   0                   dc 5
ibias   6   0                   dc 100u
m4      6   6   7   1           pmos w=390u l=2u
m5      7   7   1   1           pmos w=390u l=2u
m6      8   7   1   1           pmos w=390u l=2u
m3      2   6   8   1           pmos w=390u l=2u
m2      2   3   4   0           nmos w=100u l=1.6u
m1      4   5   0   0           nmos w=100u l=1.6u
cl      2   0                   5p
vbias   3   0                   dc 2.5
vin     5   0                   dc 0.8425 ac 1

.op
.ac dec 10 0.1 1000Meg
.print vdb(2)
.option post
.lib "../cmos_models" library
.end
```

The frequency plot of this cascode amplifier is shown in Fig. 3.52. The gain at dc is 80 dB (i.e., 10,000 V/V) and has a –3-dB frequency around 2 kHz.

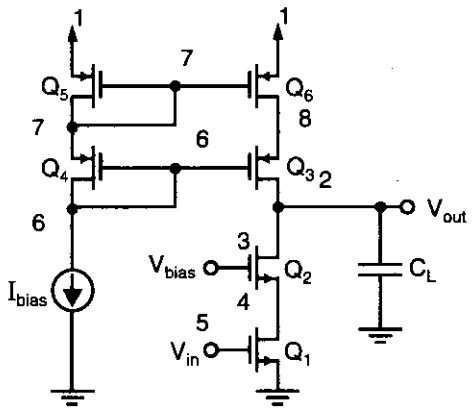


Fig. 3.51 Circuit for Example 3.11.

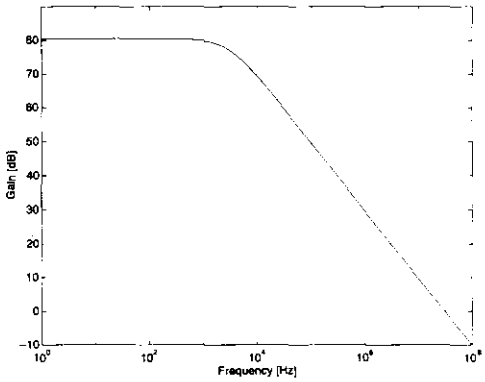


Fig. 3.52 The frequency plot of the cascode amplifier.

3.13 REFERENCES

P. R. Gray and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits*, 3rd ed. John Wiley & Sons, New York, 1993.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits*, 3rd ed. Holt, Rinehart & Winston, New York, 1991.

3.14 PROBLEMS

Unless otherwise stated, assume the following:

- npn bipolar transistors:
  $\beta = 100$
  $V_A = 80$ V
  $\tau_b = 13$ ps
  $\tau_s = 4$ ns
  $r_b = 330\ \Omega$
- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8$ V
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}\ (\Omega) = 8{,}000\,L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9$ V
  $\gamma = 0.8\ V^{1/2}$

r_{ds} (Ω) = 12,000L (μm)/I_D (mA) in active region
$C_j = 4.5 \times 10^{-4}$ pF/(μm)^2
$C_{j\text{-}sw} = 2.5 \times 10^{-4}$ pF/μm
$C_{ox} = 1.9 \times 10^{-3}$ pF/(μm)^2
$C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}$ pF/μm

3.1 Consider the current mirror shown in Fig. 3.1, where I_{in} = 80 μA, transistor Q_1 has W/L = 100 μm/1.6 μm, and transistor Q_2 has W/L = 25 μm/1.6 μm. Find the nominal output current as well as the output impedance, r_{out}. Also, find the minimum output voltage such that both transistors remain in the active region.

3.2 For the common-source amplifier in Fig. 3.4, derive the relationship between the bias current, I_{bias}, and the dc gain. Assume all transistor sizes are the same.

3.3 For the common-source amplifier in Fig. 3.4, assume there is a load capacitor, C_L, that dominates the frequency response. Derive the –3-dB frequency of the amplifier. Find the relationship between the –3-dB frequency and the bias current, I_{bias}. Assume all transistor sizes are the same.

3.4 In the high-frequency analysis of a common-source amplifier, show that the transfer function given in equation (3.103) can be derived from nodal equations (3.101) and (3.102).

3.5 Find the –3-dB frequency of the common-source amplifier in Fig. 3.4, assuming that all transistors have W/L = 75 μm/1.6 μm and that I_{bias} = 75 μA. The first step is to estimate the areas and peripheries of the junctions based on simple layout rules. In this step, assume each transistor is composed of three gate stripes. Based on this geometry, the next step is to estimate the parasitic capacitances of the small-signal model. In this step, assume the dc bias voltage of the output node is halfway between ground and the 5-V power-supply voltage.

3.6 Derive the low-frequency output impedance of the source follower shown in Fig. 3.6.

3.7 Derive an equation for the impedance looking into the source of a source follower. Use a small-signal model similar to that in Fig. 3.8, but do not take into account C_s and R_{s1}.

3.8 Repeat Example 3.9, but assume C_L = 0.5 pF when (*a*) the source is not connected to the substrate, and (*b*) the source is connected to the substrate.

3.9 Repeat Example 3.10, but assume C_L = 0.5 pF when (*a*) the source is not connected to the substrate, and (*b*) the source is connected to the substrate.

3.10 Assume that the common-gate amplifier of Fig. 3.9 has a bias current of 0.1 mA and that all transistors have a W/L of 100 μm/1.6 μm. Also assume the following:
$r_{ds\text{-}n}$ (Ω) = 8,000L (μm)/I_D (mA)
$r_{ds\text{-}p}$ (Ω) = 12,000L (μm)/I_D (mA)

R_{in} = 180 kΩ
C_L = 1 pF
C_{gs} = 0.2 pF
C_{gd} = 15 fF
C_{sb} = 40 fF
C_{db} = 20 fF
And assume the input source has a 30-fF output capacitance. Find the –3-dB frequency of the amplifier.

3.11 Derive the output impedance of the current mirror shown in Fig. P3.11, where a diode-connected transistor has been included in series with the source of the output transistor.

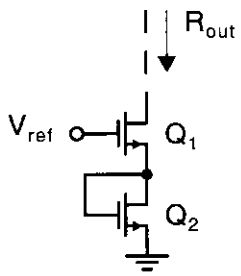


Fig. P3.11

3.12 A MOS n-channel cascode current mirror has a bias current of 0.1 mA and transistor sizes given by W/L = 50 μm/1.6 μm. What is the minimum output voltage allowable across the current mirror without any transistors entering the triode region?

3.13 Using small-signal analysis, find the output impedance of a MOS cascode current mirror. Include in your analysis the voltage-dependent current source that models the body effect.

3.14 Using small-signal analysis, find the equation for the output impedance, r_{out}, of a MOS Wilson current mirror, and show that it is approximately given by

$$r_{out} \cong r_{ds4}\left(\frac{g_{m1} r_{ds1}}{2}\right)$$

Ignore the body effect. Remember to include the effect of the output impedance of the input current source.

3.15 Repeat Problem 3.14, but include the voltage-dependent current source that models the body effect.

3.16 Repeat Example 3.11, in which a cascode amplifier is analyzed to find its –3-dB frequency, but assume C_L = 0.5 pF.

3.17 Derive an expression for the frequency at which the magnitude of the gain of a cascode amplifier has decreased to unity. What is this value when the parameters of Example 3.11 are used, except that C_L = 2 pF?

3.18 Derive the current gain, I_{out}/I_{in}, for the bipolar current mirror shown in Fig. 3.22, and show that, for $\beta >> 1$, this gain is $1 - 2/\beta$. (Neglect finite output impedances.)

3.19 Derive the current gain, I_{out}/I_{in}, for the bipolar current mirror shown in Fig. 3.23, and show that, for $\beta >> 1$, this gain is $1 - 2/\beta^2$. (Neglect finite output impedances.)

3.20 For the bipolar cascode current mirror shown in Fig. 3.25(*a*), derive the current gain, I_{out}/I_{in}, and show that, for large β, this gain is given by

$$\frac{I_{out}}{I_{in}} \approx 1 - \frac{4}{\beta}$$

3.21 For the bipolar Wilson current mirror shown in Fig. 3.25(*b*), show that the current gain, I_{out}/I_{in}, is given by

$$\frac{I_{out}}{I_{in}} = 1 - \frac{2}{\beta^2 + 2\beta + 2}$$

3.22 Derive the output impedance of the simple bipolar current mirror shown in Fig. P3.22. Take the finite output impedance, r_o, into account for both transistors. What is the required value for R, and what is the output impedance if we want the transistors to be biased at 0.2 mA ?

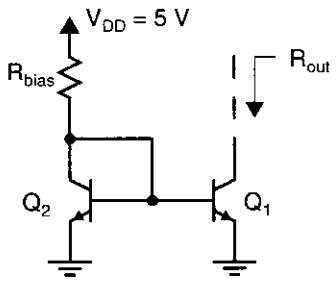


Fig. P3.22

3.23 For a common-source amplifier similar to that in Fig. 3.4 but realized using bipolar transistors, derive the relationship between the bias current, I_{bias}, and the dc gain.

3.24 Derive the –3-dB frequency of a bipolar common-emitter amplifier.

3.25 Derive the low-frequency gain and output impedance of a bipolar emitter follower that is biased with a current mirror. What are these values for a 0.5-mA bias current?

3.26 Derive the low-frequency input impedance and the gain of a bipolar common-base amplifier.

3.27 For the circuit shown in Fig. P3.27, show that the output impedance, r_{out}, is approximately given by

$$r_{out} \cong r_{o2}(1 + g_{m2}(R_e \parallel r_\pi))$$

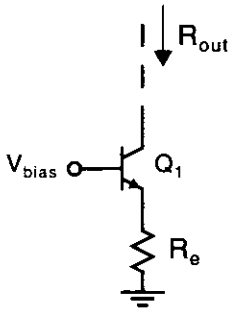


Fig. P3.27

3.28 Taking into account the finite current gain of bipolar transistors, but ignoring the finite output impedance, find the large-signal current gains of cascode and Wilson current mirrors.

3.29 Assuming $\beta >> 1$ and $g_m r_o >> \beta$, show that the output impedance of a bipolar Wilson current mirror is approximately given by

$$r_{out} \cong \frac{\beta r_o}{2}$$

3.30 Repeat Problem 3.29 for a bipolar cascode current mirror.

3.31 For the differential-input stage in Fig. 3.19, assume that $I_{bias} = 0.1\text{mA}$, all transistors have $W/L = 100\ \mu\text{m}/1.6\ \mu\text{m}$, and the load capacitance is 100 pF. Find the dc gain and the –3-dB frequency.

CHAPTER

4 Noise Analysis and Modelling

To develop good analog circuit design techniques, a basic understanding of noise sources and analysis is required. Another motivation to study noise analysis is to learn basic concepts of random signals for a proper understanding of oversampling converters. The purpose of this chapter is to present some fundamentals of noise analysis followed by an introduction to electronic noise sources and circuit analysis.

It should be mentioned here that this chapter deals with *inherent noise* as opposed to *interference noise*. Interference noise is a result of unwanted interaction between the circuit and the outside world, or between different parts of the circuit itself. This type of noise may or may not appear as random signals. Examples are power supply noise on ground wires (such as a 60-Hz hum) or electromagnetic interference between wires. Interference noise can be significantly reduced by careful circuit wiring or layout. In contrast, inherent noise refers to random noise signals that can be reduced but never eliminated since this noise is due to fundamental properties of the circuits. Some examples of inherent noise are thermal, shot, and flicker noise. Inherent noise is only moderately affected by circuit wiring or layout, such as using multiple base contacts to change the resistance value of the base of a transistor. However, inherent noise can be significantly reduced through proper circuit design, such as changing the circuit structure or increasing the power consumption.

The outline of this chapter is as follows: First, a time-domain analysis of noise signals (or other random signals) is presented. Here, basic concepts such as rms value, signal-to-noise ratio, and noise summation are presented. Next, a frequency-domain analysis of noise is presented. As with deterministic signals, analysis in the frequency domain results in more powerful analysis tools than does analysis that remains strictly in the time domain. Noise models for circuit elements are then presented, and finally, two circuit noise analyses are performed to give the reader some experience in such analysis.

4.1 TIME-DOMAIN ANALYSIS

Since inherent noise signals are random in nature, we define here some basic tools to effectively deal with random signals. Specifically, in this section, we define the following terms in the time domain: rms value, SNR, dBm, and noise summation.

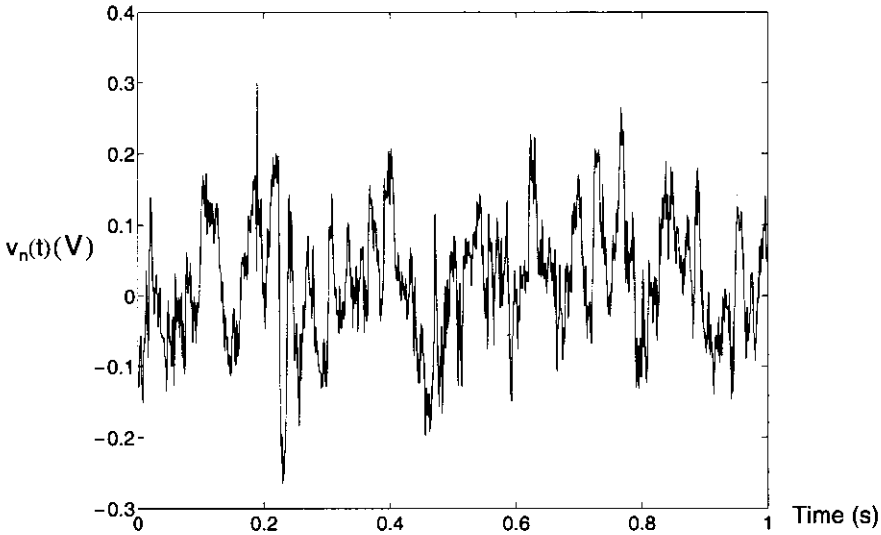


Fig. 4.1 Example of a voltage noise signal (time domain).

An example of a random noise signal in the time domain is shown in Fig. 4.1. Although this signal here is a voltage signal, it could just as easily be current noise or some other quantity. It should be noted that this noise signal appears to have an average value of zero. In fact, *throughout this chapter we will assume all noise signals have a mean value of zero*, which simplifies many of the definitions and is also valid in most physical systems.

Rms Value

Consider a noise voltage, $v_n(t)$, such as that shown in Fig. 4.1, or a noise current, $i_n(t)$. The *rms*, or *root mean square*, voltage value is defined[1] as

$$V_{n(rms)} \equiv \left[\frac{1}{T}\int_0^T v_n^2(t)\, dt\right]^{1/2} \tag{4.1}$$

where T is a suitable averaging time interval. Typically, a longer T gives a more accurate rms measurement. Similarly, the rms current value is defined as

$$I_{n(rms)} \equiv \left[\frac{1}{T}\int_0^T i_n^2(t)\, dt\right]^{1/2} \tag{4.2}$$

The benefit in knowing the rms value of a signal is that it indicates the *normalized noise power* of the signal. Specifically, if the random signal $v_n(t)$ is applied to a 1-Ω resistor, the average power dissipated, P_{diss}, in watts, equals the normalized noise

1. For those more rigorously inclined, we assume throughout this chapter that random signals are ergodic, implying their ensemble averages can be approximated by their time averages.

power and is given by

$$P_{diss} = \frac{V_{n(rms)}^2}{1\ \Omega} = V_{n(rms)}^2 \tag{4.3}$$

This relationship implies that the power dissipated by a resistor is the same whether a random signal or a dc level of k volts (rms) is applied across it. For example, a noise signal with an rms value of 1 mV (rms) dissipates the same power across a resistor as a dc voltage of 1 mV.

Similarly, for a noise current, $i_n(t)$, applied to a 1-Ω resistor,

$$P_{diss} = I_{n(rms)}^2 \times 1\ \Omega = I_{n(rms)}^2 \tag{4.4}$$

As a result, the square of the rms values, $V_{n(rms)}^2$ and $I_{n(rms)}^2$, are sometimes referred to as the normalized noise powers of these two signals.

SNR

The *signal-to-noise ratio (SNR)* value (in dB) of a signal node in a system is defined as

$$\text{SNR} \equiv 10 \log \left[\frac{\text{signal power}}{\text{noise power}} \right] \tag{4.5}$$

Thus, assuming a node in a circuit consists of a signal, $v_x(t)$, that has a normalized signal power of $V_{x(rms)}^2$ and a normalized noise power of $V_{n(rms)}^2$, the SNR is given by

$$\text{SNR} = 10 \log \left[\frac{V_{x(rms)}^2}{V_{n(rms)}^2} \right] = 20 \log \left[\frac{V_{x(rms)}}{V_{n(rms)}} \right] \tag{4.6}$$

Clearly, when the mean-squared values of the noise and signal are the same, then SNR = 0 dB.

Units of dBm

Although dB units relate the relative ratio of two power levels, it is often useful to know a signal's power in dB on an absolute scale. One common measure is that of dBm, where all power levels are referenced by 1 mW. In other words, a 1-mW signal corresponds to 0 dBm, whereas a 1-μW signal corresponds to −30 dBm. When voltage levels are measured, it is also common to reference the voltage level to the equivalent power dissipated if the voltage is applied across either a 50-Ω or a 75-Ω resistor.

EXAMPLE 4.1

Find the rms voltage of a 0-dBm signal referenced to a 50-Ω resistor. What is the level in dBm of a 2-volt rms signal?

Solution

A 0-dBm signal referenced to a 50-Ω resistor implies that the rms voltage level equals

$$V_{(rms)} = \sqrt{(50\ \Omega) \times 1\ \text{mW}} = 0.2236 \tag{4.7}$$

Thus, a 2-volt rms signal corresponds to

$$20 \log\left(\frac{2.0}{0.2236}\right) = 19\ \text{dBm} \tag{4.8}$$

and would dissipate 80 mW across a 50-Ω resistor.

Note that the measured voltage may never be physically applied across any 50-Ω resistor. The measured voltage is referenced only to power levels that would occur if the voltage were applied. Similar results are obtained if the power is referenced to a 75-Ω resistor.

Noise Summation

Consider the case of two noise sources added together, as shown in Fig. 4.2. If the rms values of each individual noise source are known, what can be said about the rms value of the combined signal? We answer this question as follows. Define $v_{no}(t)$ as

$$v_{no}(t) = v_{n1}(t) + v_{n2}(t) \tag{4.9}$$

where $v_{n1}(t)$ and $v_{n2}(t)$ are two noise sources with known rms values $V_{n1(rms)}$ and $V_{n2(rms)}$, respectively. Then we can write

$$V^2_{no(rms)} = \frac{1}{T}\int_0^T [v_{n1}(t) + v_{n2}(t)]^2\, dt \tag{4.10}$$

which, when expanded, gives,

$$V^2_{no(rms)} = V^2_{n1(rms)} + V^2_{n2(rms)} + \frac{2}{T}\int_0^T v_{n1}(t) v_{n2}(t)\, dt \tag{4.11}$$

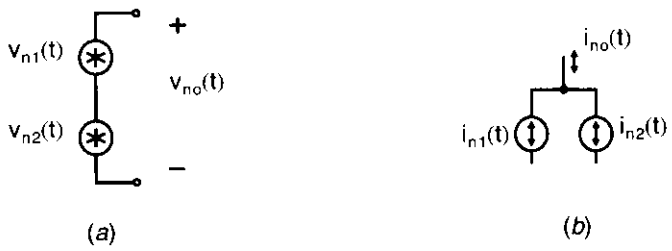


Fig. 4.2 Combining two noise sources, (*a*) voltage, and (*b*) current.

Note that the first two terms in the right-hand side of (4.11) are the individual mean-squared values of the noise sources. The last term shows the correlation between the two signal sources, $v_{n1}(t)$ and $v_{n2}(t)$. An alternate way to write (4.11) that better indicates the effects of signal correlation, is to define a correlation coefficient, C, as

$$C \equiv \frac{\frac{1}{T}\int_0^T v_{n1}(t)v_{n2}(t)\,dt}{V_{n1(rms)}V_{n2(rms)}} \tag{4.12}$$

With this definition, (4.11) can also be written as

$$V_{no(rms)}^2 = V_{n1(rms)}^2 + V_{n2(rms)}^2 + 2CV_{n1(rms)}V_{n2(rms)} \tag{4.13}$$

It can be shown that the correlation coefficient always satisfies the condition $-1 \le C \le 1$. Also, a value of $C = \pm 1$ implies the two signals are fully correlated, whereas $C = 0$ indicates the signals are uncorrelated. Values in between imply the signals are partially correlated. Fortunately, we have little reason to analyze partially correlated signals since different inherent noise sources are typically uncorrelated.

In the case of two uncorrelated signals, the mean-squared value of their sum is given by

$$V_{no(rms)}^2 = V_{n1(rms)}^2 + V_{n2(rms)}^2 \tag{4.14}$$

This relationship indicates that two rms values add as though they were vectors at right angles to each other (i.e., orthogonal) when signals are uncorrelated.

It is of interest to contrast this uncorrelated case with that for fully correlated signals. An example of two fully correlated (though deterministic) sources are two sinusoidal signals that have the same frequency and a phase of 0 or 180 degrees with each other. In this fully correlated case, the mean-squared value of their sum is given by

$$V_{no(rms)}^2 = [V_{n1(rms)} \pm V_{n2(rms)}]^2 \tag{4.15}$$

where the sign is determined by whether the signals are in or out of phase with each other. Note that, in this case where the signals are fully correlated, the rms values add linearly (similar to aligned vectors).

EXAMPLE 4.2

Given two uncorrelated noise sources that have $V_{n1(rms)} = 10\ \mu V$ and $V_{n2(rms)} = 5\ \mu V$, find their total output rms value when combined. If we are required to maintain the total rms value at $10\ \mu V$, how much should $V_{n1(rms)}$ be reduced while $V_{n2(rms)}$ remains unchanged?

Solution

Using (4.14) results in

$$V_{no(rms)}^2 = (10^2 + 5^2) = 125(\mu V)^2 \tag{4.16}$$

which results in $V_{no(rms)} = 11.2\ \mu V$.

To maintain $V_{no(rms)} = 10\ \mu V$ and $V_{n2(rms)} = 5\ \mu V$, we have

$$10^2 = V_{n1(rms)}^2 + 5^2 \tag{4.17}$$

which results in $V_{n1(rms)} = 8.7\ \mu V$. Therefore, reducing $V_{n1(rms)}$ by 13 percent is equivalent to eliminating $V_{n2(rms)}$ altogether!

The above example has an important moral. *To reduce overall noise, concentrate on large noise signals.*

4.2 FREQUENCY-DOMAIN ANALYSIS

As with deterministic signals, the frequency-domain techniques are useful for dealing with random signals such as noise. This section presents frequency-domain techniques for dealing with noise signals and other random signals. It should be noted that units of hertz (Hz) (rather than radians/second) are used throughout this chapter since, historically, such units have been commonly used in the measurement of continuous-time spectral densities.

Noise Spectral Density

Although periodic signals (such as a sinusoid) have power at distinct frequency locations, random signals have their power spread out over the frequency spectrum. For example, if the time-domain signal shown in Fig. 4.1 is applied to a spectrum analyzer, the resulting spectrum might look like that shown in Fig. 4.3(*a*). Note here that although the horizontal scale is the usual frequency axis, the vertical scale is in units of microvolts-squared/hertz. In other words, the vertical axis is a measure of the normalized noise power (mean-squared value) over a 1-Hz bandwidth at each frequency point. For example, the measurement at 100 Hz in Fig. 4.3(*a*) indicates that the normalized power between 99.5 Hz and 100.5 Hz is 10 $(\mu V)^2$.

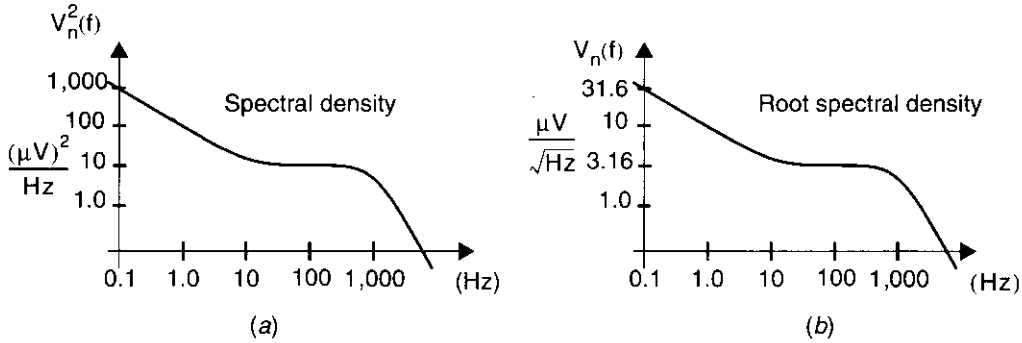


Fig. 4.3 Example of voltage spectral density (frequency domain), for (*a*) spectral density, and (*b*) root spectral density.

Thus, we define the noise spectral density, $V_n^2(f)$ (or, in the case of current, $I_n^2(f)$), as the average normalized noise power over a 1-Hz bandwidth. The units of $V_n^2(f)$ are volts-squared/hertz, whereas those of $I_n^2(f)$ are amps-squared/hertz. Also, $V_n^2(f)$ is a positive real-valued function.

It should be emphasized here that the *mean-squared value of a random noise signal at a single precise frequency is zero.* In other words, the mean-squared value of the signal shown in Fig. 4.3(*a*) measured at 100 Hz is directly proportional to the bandwidth of the bandpass filter used for the measurement. In a laboratory spectrum analyzer, the bandwidth of the bandpass filter is determined by the resolution-bandwidth control. Thus, as the resolution bandwidth goes to zero, the mean-squared value also becomes zero.[2] Conversely, as the resolution bandwidth increases, so does the measured mean-squared value. In either case, the measured mean-squared value should be normalized to the value that would be obtained for a bandwidth of 1 Hz when the noise spectral density is stated in units of $V^2/(Hz)$. For example, if the signal corresponding to Fig. 4.3(*a*) were measured at around 0.1 Hz using a resolution bandwidth of 1 mHz, the mean-squared value measured would be 1 $(\mu V)^2$, which, when scaled to a 1-Hz bandwidth, equals 1,000 $(\mu V^2)/Hz$.

An intuitive explanation of how a random noise signal is measured using a spectrum analyzer is as follows. A random noise signal has a frequency that is continually changing through a broad continuum of frequencies. A spectrum analyzer is sensitive to only a narrow frequency range that is the passband of its filter, and it measures the mean-squared value in that range. The filter of the spectrum analyzer reacts with a time constant approximately given by

$$\tau \approx \frac{1}{\pi W} \tag{4.18}$$

where W is the bandwidth of the filter. For some of the time, the random signal has the same frequency as the spectrum analyzer's filter. Thus, the spectrum analyzer effectively measures what percentage of time the random signal is in the frequency range of its filter. The narrower the bandwidth of the filter, the less percentage of time the signal is within its frequency range, and therefore the smaller is the spectrum analyzer's output. However, if the signal is not a random signal wandering in and out of the frequency range of the spectrum analyzer's filter, but is a deterministic signal at the center frequency of the filter, then the spectrum analyzer's reading is independent of the filter's bandwidth.

It is often convenient to plot the square root of the noise spectral density when we deal with filtered noise. Taking a square root results in $V_n(f)$, as shown in Fig. 4.3(*b*). We will refer to $V_n(f)$ as the *root spectral density which is expressed in units of volts/root-hertz* (i.e., $V/\sqrt{Hz}$). In the case of current noise, the resulting units are *amps/root-hertz*. Note that the horizontal axis remains unchanged although there is a root-hertz factor in the vertical axis.

Since the spectral density measures the mean-squared value over a 1-Hz bandwidth, one can obtain the total mean-squared value by integrating the spectral density

2. Such a result would not occur when a 100-Hz sinusoidal waveform is measured.

over the entire frequency spectrum. Thus, the rms value of a noise signal can also be obtained in the frequency domain using the following relationship:

$$V^2_{n(rms)} = \int_0^\infty V^2_n(f)\,df \tag{4.19}$$

and similarly for current noise,

$$I^2_{n(rms)} = \int_0^\infty I^2_n(f)\,df \tag{4.20}$$

Finally, it should be mentioned here that the spectral density function, $V^2_n(f)$, is the Fourier transform of the autocorrelation function of the time-domain signal, $v_n(t)$. This relationship is known as the *Wiener–Khinchin theorem.* It should also be noted that the relationship just shown defines a one-sided spectral density function since the noise is integrated only over positive frequencies as opposed to both negative and positive frequencies, again primarily for historical reasons. A two-sided definition results in the spectral density being divided by two since, for real-valued signals, the spectral density is the same for positive and negative frequencies, and the total mean-squared value remains unchanged.

EXAMPLE 4.3

What mean-squared value would be measured on the signal shown in Fig. 4.3 at 100 Hz when a resolution bandwidth of 30 Hz is used on a spectrum analyzer? Answer the same question for a 0.1-Hz resolution bandwidth.

Solution

Since the portion of spectral density function is flat at about 100 Hz, the measured value should simply be proportional to the bandwidth. Since the noise spectral density is $10\ (\mu V)^2/Hz$ at 100 Hz, the output of a 30-Hz filter is $30\ Hz \times 10\ (\mu V)^2 = 300\ (\mu V)^2$ (or an rms value of $\sqrt{300}\ \mu V$).

For a 0.1-Hz bandwidth, the measured value would be 10 times smaller than for a 1-Hz bandwidth, resulting in a value of $1\ (\mu V)^2$ (or an rms value of $1\ \mu V$).

White Noise

One common type of noise is *white noise.* A noise signal is said to be white if its spectral density is constant over a given frequency. In other words, a white noise signal would have a flat spectral density, as shown in Fig. 4.4, where $V_n(f)$ is given by

$$V_n(f) = V_{nw} \tag{4.21}$$

and V_{nw} is a constant value.

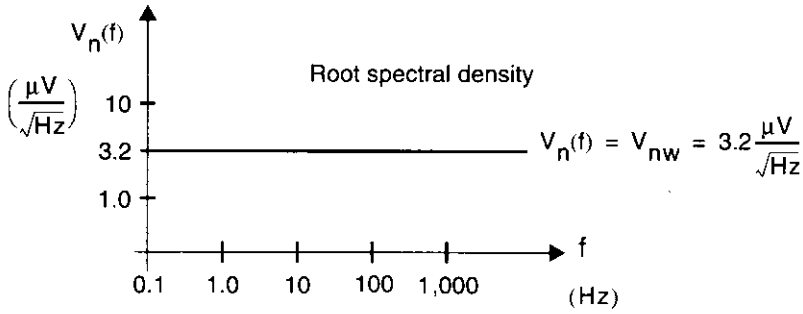


Fig. 4.4 An example of a white noise signal.

1/f, or Flicker, Noise

Another common noise shape is that of *1/f, or flicker,* noise.[3] The spectral density, $V_n^2(f)$, of 1/f noise is approximated by

$$V_n^2(f) = \frac{k_v^2}{f} \tag{4.22}$$

where k_v is a constant. Thus, the *spectral density is inversely proportional to frequency,* and hence the term "1/f noise." In terms of root spectral density, 1/f noise is given by

$$V_n(f) = \frac{k_v}{\sqrt{f}} \tag{4.23}$$

Note that it is inversely proportional to $\sqrt{f}$ (rather than f). An example of a signal having both 1/f and white noise is shown in Fig. 4.5. Note that the 1/f noise falls off at a rate of −10 dB/decade since it is inversely proportional to $\sqrt{f}$. The intersection of the 1/f and white noise curves is often referred to as the *1/f noise corner* (it occurs at 10 Hz in Fig. 4.5).

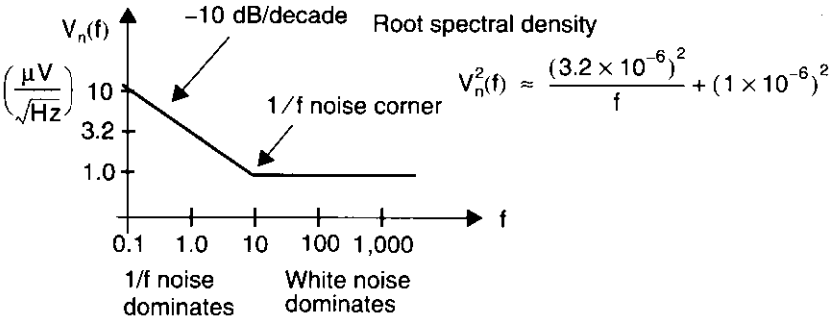


Fig. 4.5 A noise signal that has both 1/f and white noise.

3. 1/f, or flicker, noise is also referred to as *pink noise* since it has a large low-frequency content.

Filtered Noise

Consider the case of a noise signal, $V_{ni}(f)$, being filtered by the transfer function $A(s)$, as shown in Fig. 4.6. Here, $A(s)$ represents a linear transfer function as a result of some circuit amplification, filtering, or both. The following relationship between the input and output signals can be derived using the definition of the spectral density.

$$V_{no}^2(f) = |A(j2\pi f)|^2 V_{ni}^2(f) \tag{4.24}$$

The term $2\pi f$ arises here since, for physical frequencies, we replace s with $j\omega = j2\pi f$. Note that the output spectral density is a function only of the magnitude of the transfer function, and not its phase. As a result of (4.24), the total output mean-squared value is given by

$$V_{no(rms)}^2 = \int_0^\infty |A(j2\pi f)|^2 V_{ni}^2(f)\, df \tag{4.25}$$

If we wish to work with root spectral densities, we can take the square root of both sides of (4.24) resulting in

$$V_{no}(f) = |A(j2\pi f)| V_{ni}(f) \tag{4.26}$$

The relationship in (4.26) should make intuitive sense since it indicates that the transfer function simply shapes the root spectral density of the input noise signal. It is important to note here that the root spectral density is simply shaped by $|A(j2\pi f)|$, whereas the spectral density is shaped by $|A(j2\pi f)|^2$ (as seen by (4.24)). Hence, we see the benefit in dealing with the root spectral density rather than the spectral density. Specifically, straightforward transfer function analysis is applied when using root spectral densities, whereas squared terms are required to deal with spectral densities.

It is also of interest to consider the case of multiple uncorrelated noise sources that are each filtered and summed together. For example, consider the system shown in Fig. 4.7 in which three filtered, uncorrelated noise sources combine to form the total output noise, $V_{no}(f)$. In this case, one can show that if the input random signals are uncorrelated, the filter outputs are also uncorrelated. As a result, the output spectral density is given by

$$V_{no}^2(f) = \sum_{i=1,2,3} |A_i(j2\pi f)|^2 V_{ni}^2(f) \tag{4.27}$$

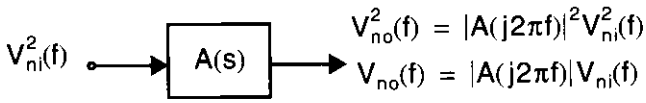
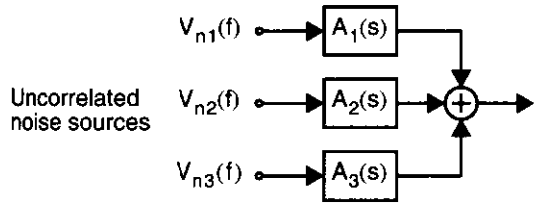


Fig. 4.6 Applying a transfer function (i.e., filter) to a noise signal.



$$V_{no}(f) = \left(\sum_{i=1,2,3} |A_i(j2\pi f)|^2 V_{ni}^2(f) \right)^{1/2}$$

Fig. 4.7 Filtered uncorrelated noise sources contributing to total output noise.

EXAMPLE 4.4

Consider a noise signal, $V_{ni}(f)$, that has a white root spectral density of $20\ \text{nV}/\sqrt{\text{Hz}}$, as shown in Fig. 4.8(*a*). Find the total noise rms value between dc and 100 kHz. What is the total noise rms value if it is filtered by the RC filter shown in Fig. 4.8(*b*), where it is assumed the RC filter is noise free?

Solution

For the noise mean-square value from dc to 100 kHz of $V_{ni}(f)$, we have

$$V_{ni(rms)}^2 = \int_0^{10^5} 20^2\, df = 4 \times 10^7 (\text{nV})^2 \tag{4.28}$$

resulting in an rms value of $V_{ni(rms)} = 6.3\ \mu\text{V rms}$. Note that, for this simple case, one could also obtain the rms value by multiplying $20\ \text{nV}/\sqrt{\text{Hz}}$ by the square root of the frequency span, or $\sqrt{100\ \text{kHz}}$, resulting in 6.3 μV rms.

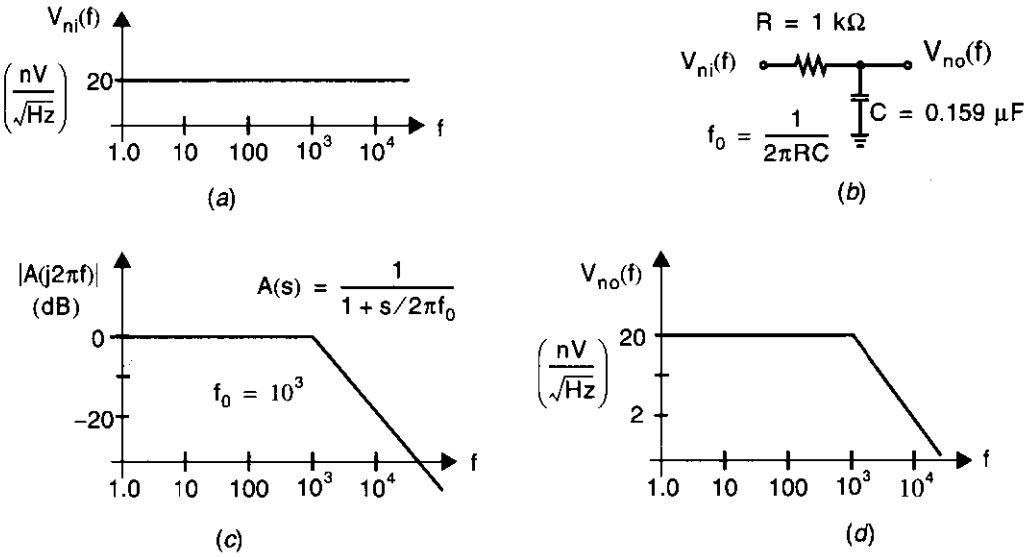


Fig. 4.8 (*a*) Spectral density for $V_{ni}(f)$. (*b*) RC filter to shape noise. (*c*) RC filter frequency response. (*d*) Spectral density for $V_{no}(f)$.

For the filtered signal, $V_{no}(f)$, we find that the RC filter has the frequency response shown in Fig. 4.8(*c*). Therefore, we can multiply the root spectral density of $V_{ni}(f)$ with the frequency response, $|A(j2\pi f)|$, to obtain the root spectral density of $V_{no}(f)$, as shown in Fig. 4.8(*d*). Mathematically, the output root spectral density is given by

$$V_{no}(f) = \frac{20 \times 10^{-9}}{\sqrt{1 + \left(\frac{f}{f_0}\right)^2}} \tag{4.29}$$

where $f_0 = 10^3$. Thus, the noise mean-squared value of $V_{no}(f)$ between dc and 100 kHz is given by

$$\begin{aligned} V_{no(rms)}^2 &= \int_0^{10^5} \frac{20^2}{1 + \left(\frac{f}{f_0}\right)^2} df = 20^2 f_0 \arctan(f/f_0) \Big|_0^{10^5} \\ &= 6.24 \times 10^5 (nV)^2 \end{aligned} \tag{4.30}$$

which results in an rms value of $V_{no(rms)} = 0.79\ \mu V$ rms. Note that the noise rms value of $V_{no}(f)$ is almost $1/10$ that of $V_{ni}(f)$ since high-frequency noise above 1 kHz was filtered. The lesson here is that you should not design circuits for larger bandwidths than your signal requires, otherwise noise performance suffers.

Noise Bandwidth

We just saw that the spectral density function is determined by the noise power within each 1-Hz bandwidth. Thus, in theory, one could measure the spectral density function by filtering a noise signal with a brick-wall bandpass filter having a 1-Hz bandwidth. The term brick wall implies here that the 1-Hz bandwidth of the filter is passed with a gain of one, whereas all other frequencies are entirely eliminated. However, practical filters can only approach a brick-wall response as their complexity (i.e., filter order) is increased. Thus, for lower-order filters with a 1-Hz passband, more noise power is passed than what is simply in a 1-Hz bandwidth. To account for the fact that practical filters have more gradual stopband characteristics, the term *noise bandwidth* is defined. *The noise bandwidth of a given filter is equal to the frequency span of a brick-wall filter that has the same output noise rms value that the given filter has when white noise is applied to both filters (peak gains are the same for the given and brick-wall filters).* In other words, given a filter response with peak gain A_0, the noise bandwidth is the width of a rectangular filter that has the same area and peak gain, A_0, as the original filter.

For example, consider a first-order, low-pass response with a 3-dB bandwidth of f_0, as shown in Fig. 4.9(*a*). Such a response would occur from the RC filter shown in Fig. 4.8(*b*) with $f_0 = (1/2\pi RC)$. The transfer function of $A(s)$ is given by

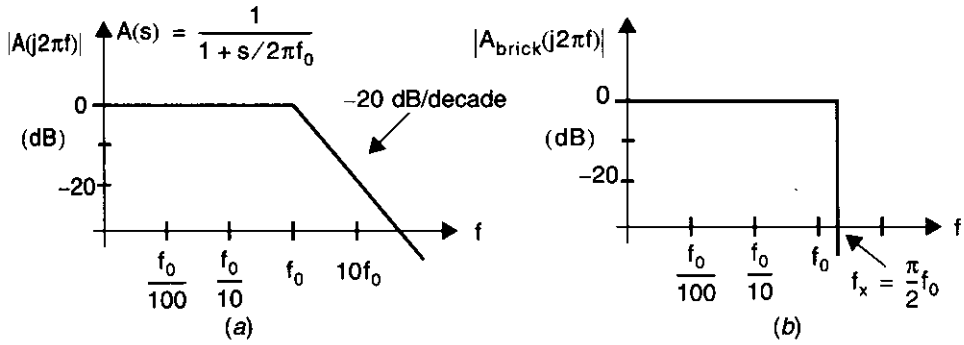


Fig. 4.9 (*a*) A first-order, low-pass response, and (*b*) a brick-wall filter that has the same peak gain and area as the first-order response.

$$A(s) = \frac{1}{1 + \dfrac{s}{2\pi f_0}} \tag{4.31}$$

This results in the magnitude response of $A(s)$ being equal to

$$|A(jf)| = \left(\frac{1}{1 + \left(\dfrac{f}{f_0}\right)^2}\right)^{1/2} \tag{4.32}$$

An input signal, $V_{ni}(f)$, is a white noise source given by

$$V_{ni}(f) = V_{nw} \tag{4.33}$$

where V_{nw} is a constant. The total output noise rms value of $V_{no}(f)$ is equal to

$$V^2_{no(rms)} = \int_0^\infty \frac{V^2_{nw}}{1 + \left(\dfrac{f}{f_0}\right)^2}\, df = V^2_{nw} f_0 \arctan\left(\frac{f}{f_0}\right)\Bigg|_0^\infty = \frac{V^2_{nw}\pi f_0}{2} \tag{4.34}$$

If this same input signal, $V_{ni}(f)$, is applied to the filter shown in Fig. 4.9(*b*), then the total output noise rms value equals,

$$V^2_{brick(rms)} = \int_0^{f_x} V^2_{nw}\, df = V^2_{nw} f_x \tag{4.35}$$

Finally, equating the two output noise rms values, $V_{no(rms)} = V_{brick(rms)}$, results in

$$f_x = \frac{\pi f_0}{2} \tag{4.36}$$

Thus, the noise bandwidth of a first-order, low-pass filter with a 3-dB bandwidth of f_0 equals $\pi(f_0/2)$. Note that, for the common case in which a first-order circuit is realized by a capacitor, C, and the resistance seen by that capacitor, R_{eq}, then

$$f_0 = \frac{1}{2\pi R_{eq} C} \tag{4.37}$$

and the noise bandwidth is given by

$$f_x = \frac{1}{4R_{eq}C} \tag{4.38}$$

The advantage of knowing the noise bandwidth of a filter is that, when white noise is applied to the filter input, the total output noise mean-squared value is easily calculated by multiplying the spectral density by the noise bandwidth. Specifically, in the first-order case just described, the total output noise mean-squared value, $V^2_{no(rms)}$, is equal to

$$V^2_{no(rms)} = V^2_{nw} f_x = V^2_{nw}\left(\frac{\pi}{2}\right) f_0 \tag{4.39}$$

Similar results for noise-bandwidth relationships can be obtained for higher-order and bandpass filters.

Piecewise Integration of Noise

Although simulation and computer techniques can analyze circuits or systems quite precisely (depending on their modelling accuracy), it is often convenient to make approximations that are useful in the early design stages. Such approximations are particularly useful in noise analysis, where large inaccuracies occur naturally due to such things as unknown parasitic effects and incomplete noise models. One such approximation is the estimation of total noise by integrating over frequency with the assumption that piecewise-linear Bode diagrams are exact. With such an approximation, integration formulas become much simpler when one needs only to integrate under linear functions and add together the resulting portions. The following example demonstrates this approach.

EXAMPLE 4.5

Consider an input noise signal, $V_{ni}(f)$, being applied to the amplifier $A(s)$, as shown in Fig. 4.10. Find the output noise rms value of $V_{no}(f)$ above 1 Hz.

Solution

As shown, the output root spectral density, $V_{no}(f)$, is determined by the addition of the Bode diagrams for $V_{ni}(f)$ and $A(s)$. To perform piecewise integration of $V_{no}(f)$, the frequency range is broken into four regions, N_1 through N_4, as shown.

For the region N_1, the total mean-square noise is given by

$$N_1^2 = \int_1^{100} \frac{200^2}{f}\, df = 200^2 \ln(f)\Big|_1^{100} = 1.84 \times 10^5 (nV)^2 \tag{4.40}$$

In the region N_2, we have

$$N_2^2 = \int_{100}^{10^3} 20^2\, df = 20^2 f\Big|_{100}^{10^3} = 3.6 \times 10^5 (nV)^2 \tag{4.41}$$

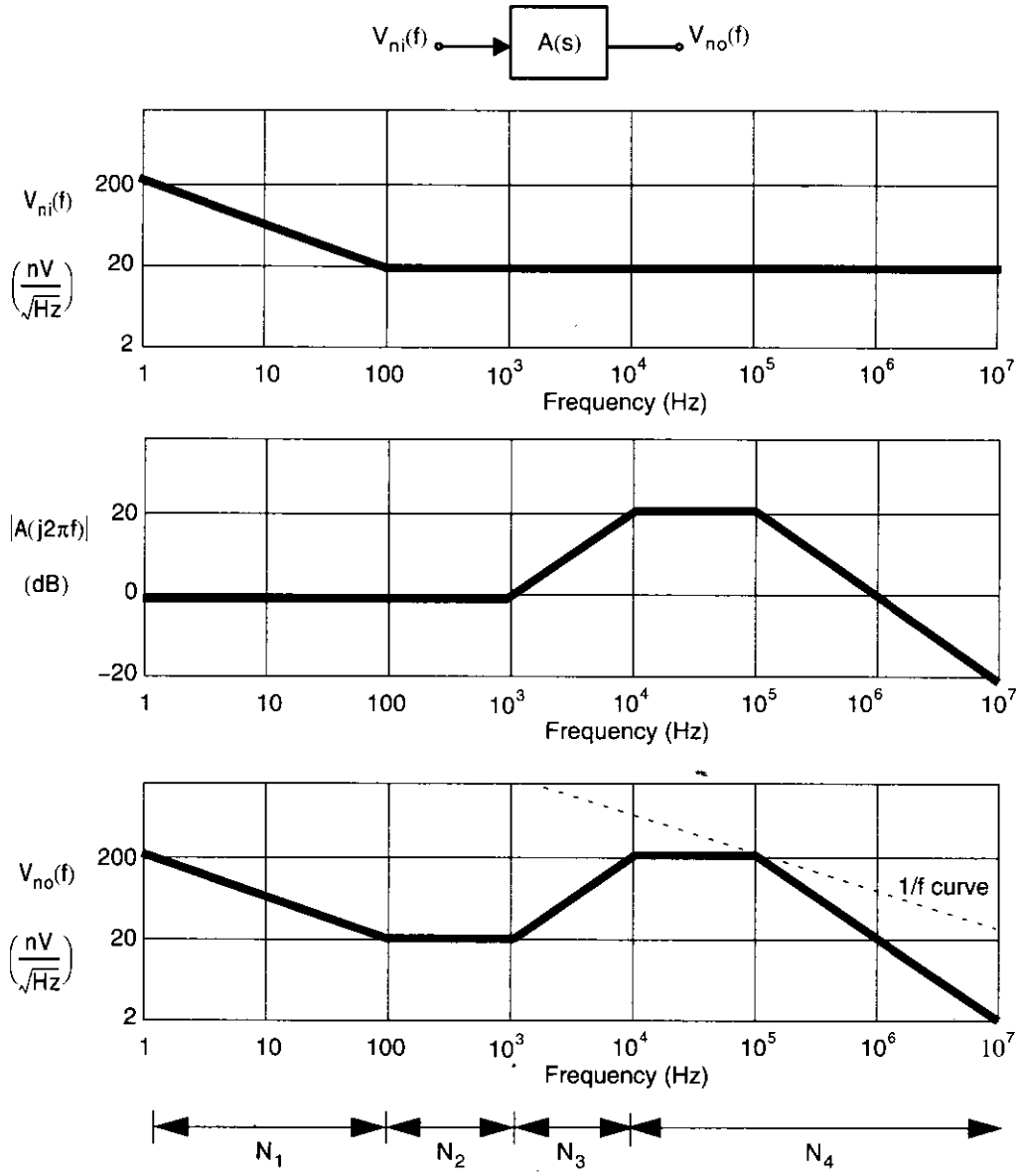


Fig. 4.10 Root spectral densities and amplifier curve example. $V_{no}(f)$ is the output noise that results from applying an input signal with noise $V_{ni}(f)$ to the amplifier $A(s)$.

Region N_3 ramps up rather than down, resulting in

$$N_3^2 = \int_{10^3}^{10^4} \frac{20^2 f^2}{(10^3)^2}\, df = \left(\frac{20}{10^3}\right)^2 \left[\frac{1}{3} f^3 \Big|_{10^3}^{10^4}\right] = 1.33 \times 10^8 (nV)^2 \tag{4.42}$$

Finally, for region N_4, we can use the noise bandwidth result of a first-order, low-pass response and simply remove the noise portion resulting from under

10^4 Hz. Specifically, we have,

$$N_4^2 = \int_{10^4}^{\infty} \frac{200^2}{1+\left(\frac{f}{10^5}\right)^2}\,df = \int_0^{\infty} \frac{200^2}{1+\left(\frac{f}{10^5}\right)^2}\,df - \int_0^{10^4} 200^2\,df \tag{4.43}$$

$$= 200^2\left(\frac{\pi}{2}\right)10^5 - (200^2)(10^4) = 5.88\times 10^9 (\text{nV})^2$$

Thus, the total output noise can be estimated to be

$$V_{no(rms)} = (N_1^2 + N_2^2 + N_3^2 + N_4^2)^{1/2} = 77.5\ \mu\text{V rms} \tag{4.44}$$

An interesting point to note here is that in the preceding example, $N_4 = 76.7\ \mu\text{V rms}$ is quite close to the total noise value of $77.5\ \mu\text{V rms}$. Thus, in practice, there is little need to find the noise contributions in the regions N_1, N_2, and N_3. Such an observation leads us to the 1/f noise tangent principle.

1/f Noise Tangent Principle

The *1/f noise tangent principle* is as follows: To determine the frequency region or regions that contribute the dominant noise, *lower a 1/f noise line until it touches the spectral density curve—the total noise can be approximated by the noise in the vicinity of the 1/f line* [Kennedy, 1988]. For example, lowering a 1/f line toward the root spectral density of $V_{no}(f)$ in Fig. 4.10 indicates that the noise around 10^5 dominates. The reason this simple rule works is that a curve proportional to $1/x$ results in equal power over each decade of frequency. Therefore, by lowering this constant power/frequency curve, the largest power contribution will touch it first. However, because the integration of $1/x$ approaches infinity if either the upper bound is infinity or the lower bound is zero, one must be careful in cases where the spectral density curve runs parallel to a 1/f tangent line for an appreciable frequency span. For example, consider once again Fig. 4.10, where the region N_1 runs parallel to the 1/f tangent line. However, in this example, region N_1 does not contribute much noise since the noise was only integrated above 1 Hz. If a much lower frequency bound is used, this region can also contribute appreciable noise power.

4.3 NOISE MODELS FOR CIRCUIT ELEMENTS

There are three main fundamental noise mechanisms—thermal, shot, and flicker. In this section, we discuss noise models for popular circuit elements where all three mechanisms occur. However, first we briefly describe these noise phenomena.

Thermal noise is due to the thermal excitation of charge carriers in a conductor. This noise has a white spectral density and is proportional to absolute temperature. It

is not dependent on bias conditions (dc bias current) and it occurs in all resistors (including conductors) above absolute zero. Thus, thermal noise places fundamental limits on the dynamic range achievable in electronic circuits. It should be mentioned here that thermal noise is also referred to as Johnson or Nyquist noise since it was first observed by J. B. Johnson [Johnson, 1928] and analyzed using the second law of thermodynamics by H. Nyquist [Nyquist, 1928].

Shot noise was first studied by W. Schottky using vacuum-tube diodes [Schottky, 1918], but shot noise also occurs in pn junctions. This noise occurs because the dc bias current is not continuous and smooth but instead is a result of pulses of current caused by the individual flow of carriers. As such, shot noise is dependent on the dc bias current. It can also be modelled as a white noise source. Shot noise is also typically larger than thermal noise and is sometimes used to create white noise generators.

Flicker noise is the least understood of the three noise phenomena. It is found in all active devices as well as in carbon resistors,[4] but it occurs only when a dc current is flowing. Flicker noise usually arises due to traps in the semiconductor, where carriers that would normally constitute dc current flow are held for some time period and then released. Flicker noise is also commonly referred to as 1/f noise since it is well modelled as having a $1/f^{\alpha}$ spectral density, where α is between 0.8 and 1.3. Although both bipolar and MOSFET transistors have flicker noise, it is a significant noise source in MOS transistors, whereas it can often be ignored in bipolar transistors.

Resistors

The major source of noise in resistors is thermal noise. As just discussed, it appears as white noise and can be modelled as a voltage source, $V_R(f)$, in series with a noiseless resistor. With such an approach, the spectral density function, $V_R^2(f)$, is found to be given by

$$V_R^2(f) = 4kTR \tag{4.45}$$

where k is Boltzmann's constant (1.38×10^{-23} JK^{-1}), T is the temperature in Kelvins, and R is the resistance size.

An alternate way to write (4.45) is to note that a 1-kΩ resistor exhibits a root spectral density of 4.06 nV/$\sqrt{\text{Hz}}$ in thermal noise at room temperature (300 °K). Since the root spectral density is proportional to the square root of the resistance, we can also write

$$V_R(f) = \sqrt{\frac{R}{1k}} \times 4.06 \text{ nV}/\sqrt{\text{Hz}} \quad \text{for } 27\ ^\circ\text{C} \tag{4.46}$$

Note that, to reduce the thermal noise due to resistors, one must either lower the temperature or use lower resistance values. The fact that lower resistance values cause less thermal noise becomes much more apparent when we look at kT/C noise in capacitors later in this section.

4. Carbon resistors are not used in integrated-circuit design but are available as discrete elements.

An alternate model can be derived by finding the Norton equivalent circuit. Specifically, the series voltage noise source, $V_R(f)$, can be replaced with a parallel current noise source, $I_R(f)$, given by

$$I_R^2(f) = \frac{V_R^2(f)}{R^2} = \frac{4kT}{R} \tag{4.47}$$

Both resistor models are summarized in Fig. 4.11.

Diodes

Shot noise is typically the dominant noise in diodes and can be modelled with a current source in parallel with the small-signal resistance of the diode, as Fig. 4.11 shows. The spectral density function of the current source is found to be given by

$$I_d^2(f) = 2qI_D \tag{4.48}$$

where q is one electronic charge (1.6×10^{-19} C) and I_D is the dc bias current flowing through the diode. The small-signal resistance of the diode, r_d, is given by the usual relationship,

$$r_d = \frac{kT}{qI_D} \tag{4.49}$$

Note that the Thévenin equivalent circuit can also be used, as shown in Fig. 4.11. It should be noted here that the small-signal resistance, r_d, is used for modelling and is not a physical resistor; hence, r_d does not contribute any thermal noise.

Bipolar Transistors

The noise in bipolar transistors is due to the shot noise of both the collector and base currents, the flicker noise of the base current, and the thermal noise of the base resistance. A common practice is to combine all these noise sources into two equivalent noise sources at the base of the transistor, as shown in Fig. 4.11. Here, the equivalent input voltage noise, $V_i(f)$, is given by

$$V_i^2(f) = 4kT\left(r_b + \frac{1}{2g_m}\right) \tag{4.50}$$

where the r_b term is due to the thermal noise of the base resistance and the g_m term is due to collector-current shot noise referred back to the input. The equivalent input current noise, $I_i(f)$, equals

$$I_i^2(f) = 2q\left(I_B + \frac{KI_B}{f} + \frac{I_C}{|\beta(f)|^2}\right) \tag{4.51}$$

where the $2qI_B$ term is a result of base-current shot noise, the KI_B/f term models 1/f noise (K is a constant dependent on device properties), and the I_C term is the input-referred collector-current shot noise (it is often ignored).

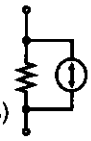
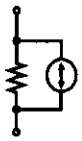
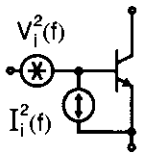
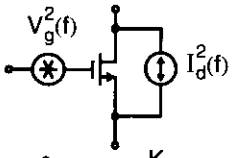
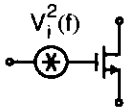
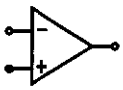
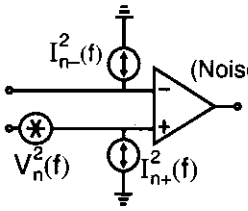
Element	Noise Models	
Resistor	R (Noiseless) $V_R^2(f) = 4kTR$	R (Noiseless)  $I_R^2(f) = \frac{4kT}{R}$
Diode (Forward biased)	$r_d = \frac{kT}{qI_D}$ (Noiseless) $V_d^2(f) = 2kTr_d$	$r_d = \frac{kT}{qI_D}$  $I_d^2(f) = 2qI_D$ (Noiseless)
BJT (Active region)	 $V_i^2(f)$ (Noiseless) $I_i^2(f)$	$V_i^2(f) = 4kT\left(r_b + \frac{1}{2g_m}\right)$ $I_i^2(f) = 2q\left(I_B + \frac{KI_B}{f} + \frac{I_C}{\lvert\beta(f)\rvert^2}\right)$
MOSFET (Active region)	 $V_g^2(f)$ $I_d^2(f)$ $V_g^2(f) = \frac{K}{WLC_{ox}f}$ $I_d^2(f) = 4kT\left(\frac{2}{3}\right)g_m$	 $V_i^2(f)$ (Noiseless) $V_i^2(f) = 4kT\left(\frac{2}{3}\right)\frac{1}{g_m} + \frac{K}{WLC_{ox}f}$ Simplified model for low and moderate frequencies
Opamp 	 $I_{n-}^2(f)$ (Noiseless) $V_n^2(f)$ $I_{n+}^2(f)$	$V_n(f)$, $I_{n-}(f)$, $I_{n+}(f)$ — Values depend on opamp — Typically, all uncorrelated

Fig. 4.11 Circuit elements and their noise models. Note that capacitors and inductors do not generate noise.

A couple of comments here are that the noise of r_b typically dominates in $V_i(f)$, and the base-current shot noise often dominates in the input-current noise, $I_i^2(f)$. Thus, the equivalent voltage and current noise are not derived from the same noise sources. As a result, it is common practice to assume that the input voltage and current noise sources in a bipolar transistor are uncorrelated.[5]

MOSFETS

The dominant noise sources for active MOSFET transistors are flicker and thermal noise, as shown in Fig. 4.11. The flicker noise is modelled as a voltage source in series with the gate of value

$$V_g^2(f) = \frac{K}{WLC_{ox}f} \tag{4.52}$$

where the constant K is dependent on device characteristics and can vary widely for different devices in the same process. The variables W, L, and C_{ox} represent the transistor's width, length, and gate capacitance per unit area, respectively. *An important point to note here is that the 1/f noise is inversely proportional to the transistor area,* WL. In other words, larger devices have less 1/f noise. 1/f noise is extremely important in MOSFET circuits, because it typically dominates at low frequencies unless switching-circuit techniques are used to reduce its effect. Also, typically p-channel transistors have less noise than their n-channel counterparts since their majority carriers (holes) are less likely to be trapped.

The derivation of the thermal noise term is straightforward and is due to the resistive channel of a MOS transistor in the active region. If the transistor was in triode, the thermal noise current in the drain due to the channel resistance would simply be given by $I_d^2(f) = (4kT)/r_{ds}$, where r_{ds} is the channel resistance. However, when the transistor is in the active region, the channel cannot be considered homogeneous, and thus, the total noise is found by integrating over small portions of the channel. Such an integration results in the noise current in the drain being given by

$$I_d^2(f) = 4kT\left(\frac{2}{3}\right)g_m \tag{4.53}$$

for the case $V_{DS} = V_{GS} - V_T$.

Often, noise analyses are done just by including this noise source between the transistor drain and source. Sometimes, however, analysis may be simplified if it is replaced by an equivalent input noise source. To find the equivalent noise voltage that would cause this drain current, we note that the drain current is equal to the gate voltage times the transconductance of the device, or, mathematically, $I_d(f) = g_m V_i(f)$.

5. An exception to this is high-frequency designs in which the collector shot noise becomes more important because, at high frequencies, β becomes small. In this case, the input current noise source is partially correlated with the input voltage noise source. If neither noise source dominates, then the correct analysis is more difficult and beyond the scope of this text. Fortunately, this case is not often encountered in practice.

Thus, dividing (4.53) by g_m^2 results in the simplified MOSFET model, also shown in Fig. 4.11, where there is now only one voltage noise source. However, one should be aware that this simplified model assumes the gate current is zero. Although this assumption is valid at low and moderate frequencies, an appreciable amount of current would flow through the gate-source capacitance, C_{gs}, at higher frequencies. In summary, although most noise analysis can be performed using the simplified model, if in doubt, one should use the model with the thermal noise placed as a current source in parallel with the drain-source channel.

Finally, it should be noted that no gate leakage noise terms have been included in this noise model since, in modern process, the gate leakage is so small that its noise contribution is rarely significant.

EXAMPLE 4.6

A large MOS transistor consists of ten individual transistors connected in parallel. Considering 1/f noise only, what is the equivalent input voltage noise spectral density for the ten transistors compared to that of a single transistor?

Solution

From (4.52), the 1/f noise can be modelled as a current source going from the drain to the source with a noise spectral density of

$$I_d^2(f) = \frac{Kg_m^2}{WLC_{ox}f} \tag{4.54}$$

where g_m is the transconductance of a single transistor. When ten transistors are connected in parallel, the drain-current noise spectral density is ten times larger, so we have

$$I_{d=10}^2(f) = \frac{10Kg_m^2}{WLC_{ox}f} \tag{4.55}$$

When this noise is referred back to the input of the equivalent transistor, we have

$$V_{i=10}^2(f) = \frac{10Kg_m^2}{WLC_{ox}fg_{m=10}^2} = \frac{K}{10WLC_{ox}f} \tag{4.56}$$

since the transconductance of ten transistors in parallel, $g_{m=10}$, is equal to $10\ g_m$. Thus, the drain current noise spectral density of the equivalent transistor is ten times larger, but the input voltage noise spectral density is ten times smaller. This result is expected because the input voltage noise source due to 1/f noise is inversely proportional to the equivalent transistor area, which in this case is $10\ WL$.

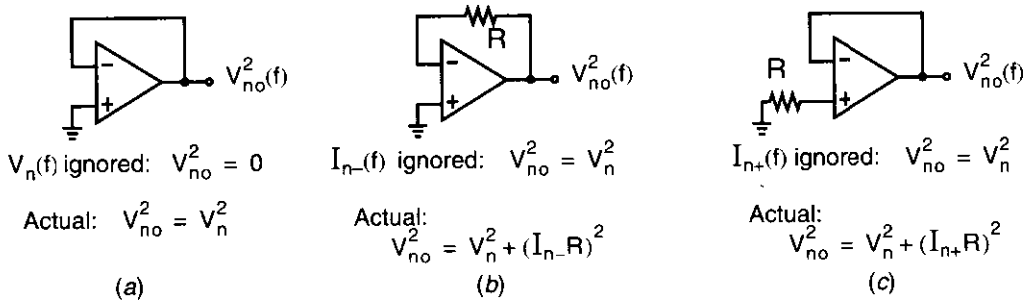


Fig. 4.12 Opamp circuits showing the need for three noise sources in an opamp noise model. Assume the resistance, R, is noiseless. Also, notation is simplified from $V_n(f)$ to V_n, and so on.

Opamps

Noise in opamps is modelled using three uncorrelated input-referred noise sources, as shown in Fig. 4.11. With an opamp that has a MOSFET input stage, the current noises can often be ignored at low frequencies since their values are small. However, for bipolar input stages, all three noise sources are typically required, as shown in Fig. 4.12. In Fig. 4.12(*a*), if $V_n(f)$ is not included in the model, a unity-gain buffer with no resistors indicates that the circuit output is noiseless. In fact, the voltage noise source, $V_n(f)$, dominates. If $I_{n-}(f)$ is not included in an opamp model, the circuit shown in Fig. 4.12(*b*) indicates that the output noise voltage equals $V_n(f)$. Here, the current noise may dominate if the resistance, R, is large. A similar conclusion can be drawn with $I_{n+}(f)$, as shown in Fig. 4.12(*c*).

Capacitors and Inductors

Capacitors and inductors do not generate any noise. However, they do accumulate noise generated by other noise sources. Here, we will see that the capacitor noise mean-squared value equals kT/C when it is connected to an arbitrary resistor value.

Consider a capacitance, C, in parallel with a resistor of arbitrary size, R, as shown in Fig. 4.13(*a*). The equivalent circuit for noise analysis is shown in Fig. 4.13(*b*). To determine the total noise mean-squared value across the capacitor, we note that $V_{no}(f)$ is simply a first-order, low-pass, filtered signal with $V_R(f)$ as the

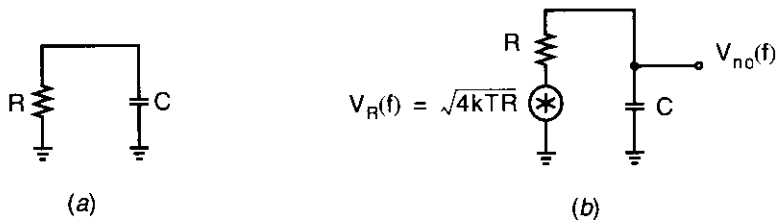


Fig. 4.13 (*a*) Capacitor, C, in parallel with a resistor, and (*b*) equivalent noise model circuit.

input. Therefore, we recognize that the noise bandwidth is given by $(\pi/2)f_0$ (see Section 4.2, Eq. (4.36), and since the input has a white spectral density, the total output mean-squared value is calculated as

$$V^2_{no(rms)} = V^2_R(f)\left(\frac{\pi}{2}\right)f_0 = (4kTR)\left(\frac{\pi}{2}\right)\left(\frac{1}{2\pi RC}\right)$$
$$V^2_{no(rms)} = \frac{kT}{C} \tag{4.57}$$

In other words, the rms voltage value across a capacitor is equal to $\sqrt{kT/C}$, regardless of the resistance seen across it. Such a result is due to fact that small resistances have less noise spectral density but result in a wide bandwidth, compared to large resistances, which have reduced bandwidth but larger noise spectral density.

Finally, it should be stated that this noise property for capacitors gives a fundamental limit on the minimum noise level across a capacitor.[6] Thus, to lower the noise level, either the temperature must be lowered or the capacitance value must be increased.

EXAMPLE 4.7

At a room temperature of 300 °K, what capacitor size is needed to achieve a 96-dB dynamic range in an analog circuit with maximum signal levels of 1 V rms?

Solution

The value of noise that can be tolerated here is 96 dB down from 1 V rms, which is

$$V_{n(rms)} = \frac{1V}{10^{96/20}} = 15.8\ \mu V\ \text{rms} \tag{4.58}$$

Using (4.57), we have

$$C = \frac{kT}{V^2_{n(rms)}} = 16.6\ \text{pF} \tag{4.59}$$

Thus, the minimum capacitor size that can be used (without oversampling) is 16.6 pF. The use of this minimum capacitor size determines the maximum resistance size allowed to achieve a given time constant.

Finally, it should be mentioned that the equivalent noise current mean-squared value in an inductor of value L is given by (see Problem 4.9)

$$I^2_{no(rms)} = \frac{kT}{L} \tag{4.60}$$

6. Some feedback circuits can make the noise smaller but signal levels are also typically smaller (see Problem 4.8).

Sampled Signal Noise

In many cases, one should obtain a sampled value of an analog voltage. For example, switched-capacitor circuits make extensive use of sampling, and sample-and-hold circuits are commonly used in analog-to-digital and digital-to-analog conversion.

Consider a basic sample-and-hold circuit, as shown in Fig. 4.14. When ϕ_{clk} drops, the transistor turns off and, in an ideal noiseless world, the input voltage signal at that instance would be held on capacitance C. However, when thermal noise is present, the resistance when the transistor is switched on causes the capacitance voltage noise to be equal to $\sqrt{kT/C}$. Thus, when the switch is turned off, the noise as well as the desired signal is held on C. As a result, a fundamental limit occurs for sampled signals using a capacitance C—an rms noise voltage of $\sqrt{kT/C}$.

It should be noted that this noise voltage will not depend on the sampling rate and is independent from sample to sample. This fact suggests a method to reduce the noise level of a signal measurement. Specifically, in the case where v_{in} is a dc (or low-frequency) signal, taking only one sample results in a noise voltage of $\sqrt{kT/C}$. However, if many samples are taken (say, 1,000) and all samples are averaged, the averaged value will have a reduced noise level. The reason this technique improves the measurement accuracy is that, when individual sampled values are summed together, their signal values add linearly, whereas their noise values add as the root of the sum of squares. This technique is known as oversampling and will be discussed at length with respect to oversampling converters in Chapter 14.

4.4 NOISE ANALYSIS EXAMPLES

In this section, a variety of circuits are analyzed from a noise perspective. Although some useful design techniques for reducing noise are presented, the main purpose of this section is to give the reader some experience in analyzing circuits from a noise perspective.

Opamp Example

Consider an inverting amplifier and its equivalent noise model, as shown in Fig. 4.15. Here, $V_n(f)$, $I_{n-}(f)$, and $I_{n+}(f)$ represent the opamp's equivalent input noise, and the remaining noise sources are resistor thermal noise sources. Note that current noise sources are used in the models for R_1 and R_f, whereas a voltage noise source is used for R_2. As we will see, these choices of noise sources simplify the circuit analysis.

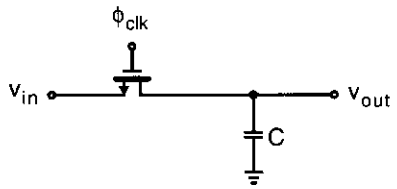


Fig. 4.14 A sample-and-hold circuit.

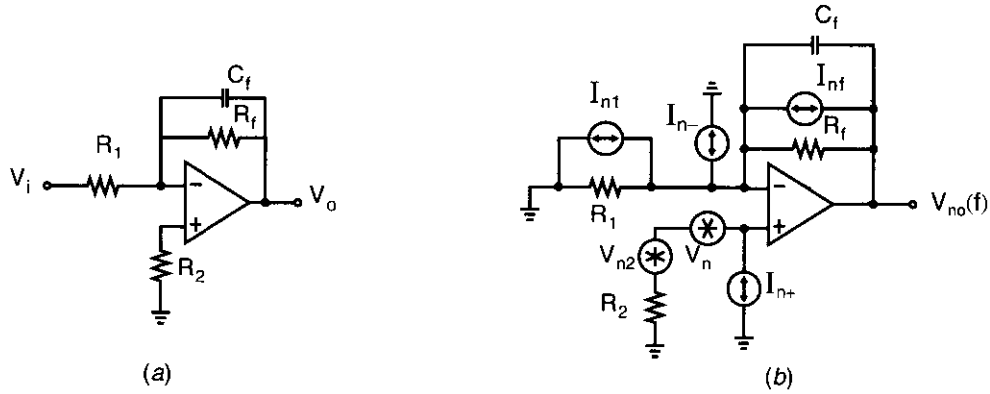


Fig. 4.15 (*a*) Low-pass filter, and (*b*) equivalent noise model.

First, using superposition and assuming all noise sources are uncorrelated, consider the output noise voltage, $V_{no1}^2(f)$, due only to I_{n1}, I_{nf}, and I_{n-}. These three noise currents add together, and their total current sum is fed into the parallel combination of C_f and R_f. Note that no current passes through R_1 since the voltage across it is zero due to the virtual ground at the negative opamp terminal (assuming a high-gain opamp). Thus, the output noise mean-squared value due to these three noise currents is given by

$$V_{no1}^2(f) = [I_{n1}^2(f) + I_{nf}^2(f) + I_{n-}^2(f)]\left|\frac{R_f}{1 + j2\pi f R_f C_f}\right|^2 \tag{4.61}$$

This equation indicates that this part of the output noise value equals the sum of the noise currents multiplied by R_f^2. This noise portion is then shaped by a low-pass filter with a 3-dB frequency equal to $f_0 = 1/(2\pi R_f C_f)$.

Using superposition further, the output mean-squared value, $V_{no2}^2(f)$, due to the three noise sources at the positive opamp terminal can be found as follows: By converting I_{n+} to a voltage source (by multiplying it by R_2), we see that the three noise voltages are summed and are applied to the positive opamp terminal. Since the gain from the positive opamp terminal to the output signal is easily found, the output noise mean-squared value due to these three noise sources is given by

$$V_{no2}^2(f) = [I_{n+}^2(f)R_2^2 + V_{n2}^2(f) + V_n^2(f)]\left|1 + \frac{R_f/R_1}{1 + j2\pi f C_f R_f}\right|^2 \tag{4.62}$$

This equation indicates that this part of the output noise mean-squared value equals the sum of the noise voltages, and this noise portion is then shaped by the shown transfer function. For this transfer function, if $R_f << R_1$, then its gain approximately equals unity for all frequencies. Thus, for an ideal opamp, the noise would exist up to infinite frequency, resulting in an infinite amount of mean-squared volts. However, for practical circuits, the gain drops off above the unity-gain frequency of the opamp, and thus the noise is effectively low-pass filtered. In the case where $R_f >> R_1$, the low-frequency gain is roughly R_f/R_1, and its 3-dB frequency is the same as in the

case of the noise sources at the negative opamp terminal (i.e., $f_0 = 1/(2\pi R_f C_f)$). However, in this case, the gain only decreases to unity and then remains at that level. Treating the Bode plot as exact and noting that, above f_0, the gain drops off at –20 dB/decade, this transfer function reaches unity around $f_1 = (R_f/R_1)f_0$. Thus, one should also include the opamp's positive input noise (with a gain of one) integrated over the region between f_1 and the unity-gain frequency of the opamp.

Finally, the total output noise mean-squared value is simply the sum

$$V_{no}^2(f) = V_{no1}^2(f) + V_{no2}^2(f) \tag{4.63}$$

or, if rms values are found,

$$V_{no(rms)}^2 = V_{no1(rms)}^2 + V_{no2(rms)}^2 \tag{4.64}$$

EXAMPLE 4.8

Estimate the total output noise rms value for a 10-kHz low-pass filter, as shown in Fig. 4.15, when $C_f = 160$ pF, $R_f = 100$ k, $R_1 = 10$ k, and $R_2 = 9.1$ k. Also, find the SNR for an input signal equal to 100 mV rms. Assume that the noise voltage of the opamp is given by $V_n(f) = 20\ \text{nV}/\sqrt{\text{Hz}}$, both its noise currents are $I_n(f) = 0.6\ \text{pA}/\sqrt{\text{Hz}}$, and that its unity-gain frequency equals 5 MHz.

Solution

Assuming the device is at room temperature, the resistor noise sources are equal to

$$I_{nf} = 0.406\ \text{pA}/\sqrt{\text{Hz}} \tag{4.65}$$

$$I_{n1} = 1.28\ \text{pA}/\sqrt{\text{Hz}} \tag{4.66}$$

$$V_{n2} = 12.2\ \text{nV}/\sqrt{\text{Hz}} \tag{4.67}$$

The low-frequency value of $V_{no1}^2(f)$ is found by letting $f = 0$ in (4.61).

$$\begin{aligned} V_{no1}^2(0) &= [I_{n1}^2(0) + I_{nf}^2(0) + I_{n-}^2(0)]R_f^2 \\ &= (0.406^2 + 1.28^2 + 0.6^2)(1 \times 10^{-12})^2(100\,\text{k})^2 \\ &= (147\ \text{nV}/\sqrt{\text{Hz}})^2 \end{aligned} \tag{4.68}$$

Since (4.61) also indicates that this noise is low-pass filtered, the rms output noise value due to these three sources can be found using the concept of a noise equivalent bandwidth. Specifically, we multiply the spectral density by $(\pi f_0)/2$, where $f_0 = 1/(2\pi R_f C_f)$. Thus,

$$\begin{aligned} V_{no1(rms)}^2 &= (147\ \text{nV}/\sqrt{\text{Hz}})^2 \times \frac{1}{4(100\ \text{k}\Omega)(160\ \text{pF})} \\ &= (18.4\ \mu\text{V})^2 \end{aligned} \tag{4.69}$$

To estimate the output noise due to the sources at the positive opamp terminal, we find $V_{no2}^2(0)$ to be given by

$$\begin{aligned} V_{no2}^2(0) &= [I_{n+}^2(f)R_2^2 + V_{n2}^2(f) + V_n^2(f)](1 + R_f/R_1)^2 \\ &= (24.1\ \text{nV}/\sqrt{\text{Hz}})^2 \times 11^2 \\ &= (265\ \text{nV}/\sqrt{\text{Hz}})^2 \end{aligned} \tag{4.70}$$

This noise is also low-pass filtered at f_0 until $f_1 = (R_f/R_1)f_0$, where the noise gain reaches unity and it remains until $f_t = 5$ MHz (i.e, the unity-gain frequency of the opamp). Thus, breaking this noise into two portions, and using (4.38) to calculate the first portion, we have

$$\begin{aligned} V_{no2(rms)}^2 &= (265 \times 10^{-9})^2\left(\frac{1}{4R_fC_f}\right) + (24.1 \times 10^{-9})^2\left(\frac{\pi}{2}\right)(f_t - f_1) \\ &= (74.6\ \mu\text{V})^2 \end{aligned} \tag{4.71}$$

Thus, the total output noise is estimated to be

$$V_{no(rms)} = \sqrt{V_{no1(rms)}^2 + V_{no2(rms)}^2} = 77\ \mu\text{V rms} \tag{4.72}$$

It should be noted here that the major source of noise at low frequencies is due to the opamp's voltage noise, $V_n(f)$.

To obtain the SNR for this circuit with a 100-mV rms input level, one can find the output signal mean-squared value and relate that to the output noise value. Alternatively, one can find the equivalent input noise value by dividing the output noise value by the circuit's gain and then relate the input signal and noise mean-squared values. Taking the first approach results in an output signal level of 1 V rms, which gives an SNR of

$$\text{SNR} = 20\log\left(\frac{1\ \text{V}}{77\ \mu\text{V}}\right) = 82\ \text{dB} \tag{4.73}$$

Note here that using a lower-speed opamp would have reduced the total output noise, as would choosing an opamp with a lower noise voltage. Also note that R_2 contributes to the output noise through both its thermal noise and the noise current of the opamp's positive input. Since its only purpose is to improve the dc offset performance, R_2 should be eliminated in a low-noise circuit (assuming dc offset can be tolerated).

Bipolar Common-Emitter Example

In this example, we consider a bipolar common-emitter amplifier, as shown in Fig. 4.16. Here, we wish to find the optimum bias current to minimize the equivalent input noise of this amplifier (i.e., maximize the signal-to-noise ratio of the amplifier).

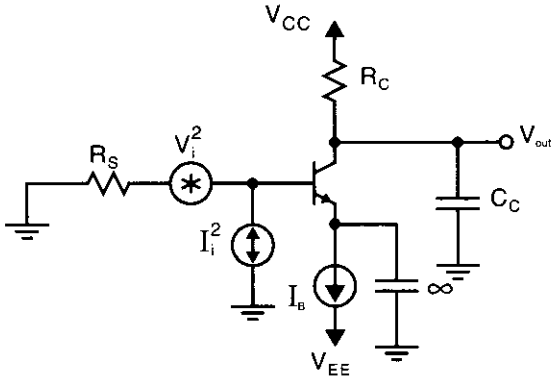


Fig. 4.16 A bipolar common-emitter amplifier.

We should state that it is assumed here that the collector-current shot noise dominates in the input voltage noise source and the base-current shot noise dominates in the input current noise source. Therefore, the input voltage noise due to the transistor alone is given by

$$V_i^2(f) = 2kT/g_m \tag{4.74}$$

and the input current noise due to the transistor alone is given by

$$I_i^2(f) = 2qI_B \tag{4.75}$$

These assumptions are reasonable since usually $1/f$ noise is not important for bipolar transistors, and the collector shot noise is usually not a major component of the input current noise source for wideband examples. Also, although the base resistance has temporarily been ignored, it will be taken into account shortly by simply modifying the size of the source resistance.

The first step is to replace the input current noise source by an input voltage noise source. First, note that the gain from the base to the collector (ignoring the transistor output impedance, r_o) is equal to $g_m R_C$, and the impedance looking into the base is r_π. Therefore, the output noise due to $I_i^2(f)$ is given by

$$V_{oi}^2(f) = I_i^2(f)[(R_S \parallel r_\pi)g_m R_C]^2 \tag{4.76}$$

Now the gain from the input voltage noise source to the output is given by

$$A_V = \frac{V_o}{V_i} = \frac{r_\pi}{r_\pi + R_S} g_m R_C \tag{4.77}$$

To represent the output noise due to the input current source by an equivalent input voltage source, we have

$$V_{ieq}^2(f) = \frac{V_{oi}^2(f)}{A_V^2} = \left(\frac{\dfrac{r_\pi R_S}{(r_\pi + R_S)} g_m R_C}{\dfrac{r_\pi}{(r_\pi + R_S)} g_m R_C} \right)^2 I_i^2(f) = I_i^2(f) R_S^2 \tag{4.78}$$

We can now replace all noise sources by a single voltage noise source, which includes the noise due to the input voltage noise source and the input current noise source of the transistor, as well as the noise of the source resistor. We have

$$V^2_{i=total}(f) = V^2_i(f) + V^2_{ieq}(f) + 4kTR_S \tag{4.79}$$

The first two terms represent the noise of the transistor and the third term represents the noise of the source resistance. Using (4.74), (4.75), and (4.78), we have

$$V^2_{i=total}(f) = \frac{2kT}{g_m} + 2qI_BR_S^2 + 4kTR_S \tag{4.80}$$

Substituting $g_m = qI_C/(kT)$ and $I_B = I_C/\beta$, we find

$$V^2_{i=total}(f) = \frac{2(kT)^2}{qI_C} + \frac{2qI_CR_S^2}{\beta} + 4kTR_S \tag{4.81}$$

The noise of the transistor base resistor can now be included by replacing R_S with $R_S + r_b$ to obtain

$$V^2_{i=total}(f) = \frac{2(kT)^2}{qI_C} + \frac{2qI_C(R_S + r_b)^2}{\beta} + 4kT(R_S + r_b) \tag{4.82}$$

Alternatively, (4.82) can be expressed in terms of $g_m = qI_C/(kT)$ as in [Buchwald, 1995]

$$V^2_{i=total}(f) = 4kT\left[R_s + r_b + \frac{1}{2g_m} + \frac{g_m(R_s + r_b)^2}{2\beta}\right] \tag{4.83}$$

These two equations are good approximations for most applications of bipolar common-emitter amplifiers. Notice in (4.82), the first term models the transistor base-current shot noise and increases with increased bias current. However, the second term (which models the transistor collector shot noise) decreases with increasing collector bias current. Finally, the terms modelling the noise of the source resistor and the transistor base resistor are independent of the transistor bias current. To find the optimum bias current, we differentiate (4.82) with respect to I_C and set the result to zero. This yields the rather simple result of

$$I_{C=opt} = \frac{kT}{q}\frac{\sqrt{\beta}}{R_S + r_b} \tag{4.84}$$

or, equivalently,

$$g_{m=opt} = \frac{\sqrt{\beta}}{R_S + r_b} \tag{4.85}$$

EXAMPLE 4.9

Consider the common-emitter amplifier shown in Fig. 4.16, where $R_C = 10\ k\Omega$, $C_C = 1\ pF$, $R_S = 500\ \Omega$, $\beta = 100$, and the base resistance is given

by $r_b = 300\ \Omega$. At room temperature, (i.e. 300 °K), find the optimum bias current and the total equivalent input noise.

Solution

Using (4.84), we find

$$I_{C=opt} = 0.026 \times \frac{\sqrt{100}}{500 + 300} = 0.325 \text{ mA} \tag{4.86}$$

implying that

$$g_m = \frac{I_C}{V_T} = 12.5 \text{ mA/V} \tag{4.87}$$

To find the output noise spectral density, we now use (4.83) to find

$$V^2_{i=total}(f) = 4kT(500 + 300 + 40 + 40) = 1.46 \times 10^{-17} \text{ V}^2/\text{Hz} \tag{4.88}$$

Notice that the noise due to the source resistance dominates, even though the source resistance is moderately small. Even for no source resistance, the thermal noise due to the base resistance would still dominate. Thus, for a very low source resistance, the major way to improve the noise is to minimize the base resistance by using larger transistors (or to combine a number of parallel transistors).

Assuming the RC time constant at the collector dominates the frequency response, the noise bandwidth of the amplifier is given by

$$f_x = \frac{1}{4R_C C_C} = 25 \text{ MHz} \tag{4.89}$$

resulting in the total input-referred voltage noise given by

$$V_{ni(rms)} = [f_x V^2_{i=total}(f)]^{1/2} = 19.1\ \mu\text{V} \tag{4.90}$$

CMOS Example

In this example, we look at the input circuitry of a traditional two-stage CMOS opamp, as shown in Fig. 4.17. Note that each of the transistors have been modelled using an equivalent voltage noise source, as presented in Fig. 4.11. Voltage noise sources are used here since we will be addressing the low-frequency noise performance of this stage. It should be mentioned here that in the following derivations, we have assumed matching between transistor pair Q_1 and Q_2 as well as in pair Q_3 and Q_4.

We start by finding the gains from each noise source to the output node, V_{no}. The gains from V_{n1} and V_{n2} are the same as the gains from the input signals, resulting in

$$\left|\frac{V_{no}}{V_{n1}}\right| = \left|\frac{V_{no}}{V_{n2}}\right| = g_{m1} R_o \tag{4.91}$$

where R_o is the output impedance seen at V_{no}.

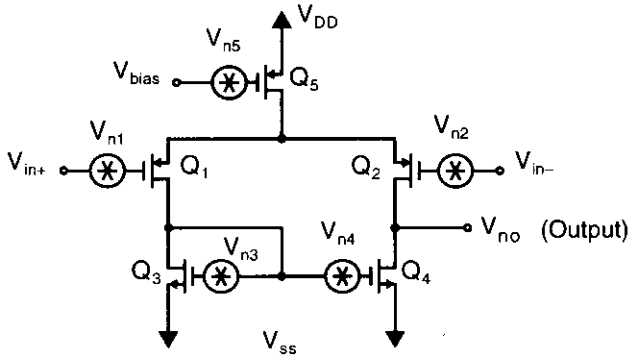


Fig. 4.17 A CMOS input stage for a traditional opamp with MOSFET noise sources shown.

Next, for V_{n3}, notice that the current through this voltage source must be zero because one side is connected to the gate of Q_3 only. Therefore, assuming all other sources are zero, the drain current is unaffected by V_{n3}. This implies that the gate voltage of Q_3 is also unaffected by V_{n3}. Therefore, in the small-signal model, V_{n3} is equal to v_{gs4}, and the gain from V_{n3} to the output is the same as the gain from V_{n4} to the output. Thus, we have

$$\left|\frac{V_{no}}{V_{n3}}\right| = \left|\frac{V_{no}}{V_{n4}}\right| = g_{m3}R_o \tag{4.92}$$

Finally, the noise gain from V_{n5} to the output can be found by noting that it modulates the bias current and the fact that, due to the symmetry in the circuit, the drain of Q_2 will track that of Q_1. As a result, the last gain factor is given by

$$\left|\frac{V_{no}}{V_{n5}}\right| = \frac{g_{m5}}{2g_{m3}} \tag{4.93}$$

Since this last gain factor is relatively small compared to the others, it will be ignored from here on.

Using the gain factors just shown, the output noise value is seen to be given by

$$V_{no}^2(f) = 2(g_{m1}R_o)^2V_{n1}^2(f) + 2(g_{m3}R_o)^2V_{n3}^2(f) \tag{4.94}$$

This output noise value can be related back to an equivalent input noise value, $V_{neq}(f)$, by dividing it by the gain, $g_{m1}R_o$, which results in

$$V_{neq}^2(f) = 2V_{n1}^2(f) + 2V_{n3}^2(f)\left(\frac{g_{m3}}{g_{m1}}\right)^2 \tag{4.95}$$

Thus, for the white noise portion of $V_{n1}(f)$ and $V_{n3}(f)$, we make the substitution

$$V_{ni}^2(f) = 4kT\left(\frac{2}{3}\right)\left(\frac{1}{g_{mi}}\right) \tag{4.96}$$

resulting in

$$V_{neq}^2(f) = \left(\frac{16}{3}\right)kT\left(\frac{1}{g_{m1}}\right) + \left(\frac{16}{3}\right)kT\left(\frac{g_{m3}}{g_{m1}}\right)^2\left(\frac{1}{g_{m3}}\right) \tag{4.97}$$

Assuming g_{m3}/g_{m1} is not far from unity, we see here that the two pairs of transistors contribute an approximately equal amount of noise, and this noise is inversely proportional to the transconductance of g_{m1}. *In other words,* g_{m1} *should be made as large as possible to minimize thermal noise contribution.*

However, to look at the effects of 1/f, or flicker, noise, which normally greatly dominates at low frequencies, we make the following substitution into (4.95),

$$g_{mi} = \sqrt{2\mu_i C_{ox}\left(\frac{W}{L}\right)_i I_{Di}} \tag{4.98}$$

resulting in

$$V_{ni}^2(f) = 2V_{n1}^2(f) + 2V_{n3}^2(f)\left[\frac{(W/L)_3\mu_n}{(W/L)_1\mu_p}\right] \tag{4.99}$$

Now, letting each of the noise sources have a spectral density given by

$$V_{ni}^2(f) = \frac{K_i}{W_i L_i C_{ox} f} \tag{4.100}$$

we have [Bertails, 1979]

$$V_{ni}^2(f) = \frac{2}{C_{ox} f}\left[\frac{K_1}{W_1 L_1} + \left(\frac{\mu_n}{\mu_p}\right)\left(\frac{K_3 L_1}{W_1 L_3^2}\right)\right] \tag{4.101}$$

Recall that the first term in (4.101) is due to the p-channel input transistors, Q_1 and Q_2, and the second term is due to the n-channel loads, Q_3 and Q_4. We note some points for 1/f noise here:

1. For $L_1 = L_3$, the noise of the n-channel loads dominate since $\mu_n > \mu_p$ and typically n-channel transistors have larger 1/f noise than p-channel transistors (i.e., $K_3 > K_1$).
2. Taking L_3 longer greatly helps due to the inverse squared relationship in the second term of (4.101). This limits the signal swings somewhat, but it may be a reasonable trade-off where low noise is important.
3. The input noise is independent of W_3, and therefore we can make it large to maximize signal swing at the output.
4. Taking W_1 wider also helps to minimize 1/f noise. (Recall that it helps white noise, as well.)
5. Taking L_1 longer increases the noise because the second term in (4.101) is dominant. Specifically, this decreases the input-referred noise of the p-channel drive transistors, which are not the dominant noise sources, but it also increases

the input-referred noise of the n-channel load transistors, which are the dominant noise sources!

Finally, we can integrate (4.101) from f_1 to f_2 to find the equivalent input noise value given by

$$V_{ni(rms)}^2 = 2\left[\frac{a_p}{W_1 L_1} + a_n\left(\frac{\mu_n}{\mu_p}\right)\left(\frac{L_1}{W_1 L_3^2}\right)\right] \tag{4.102}$$

where

$$a_i = \frac{K_i}{C_{ox}} \ln\left(\frac{f_2}{f_1}\right) \tag{4.103}$$

Fiber-Optic Preamp Example

The most popular means of detecting light from a fiber-optic cable is to use a transresistance amplifier, as shown in Fig. 4.18. The light from the fiber cable hits a photodetector, such as a reverse-biased diode. The light produces electron hole carriers in the depletion region of the reverse-biased diode, causing current to flow through the resistor, and therefore the output of the amplifier becomes positive. A popular choice for the first stage of the amplifier is to use a common-source amplifier with a resistor load. In low-noise applications, an active load would be too noisy. Assuming a CMOS transistor is used,[7] the preamp can be modelled as shown in Fig. 4.19. The photodetector is modelled as an input current source along with a parasitic capacitance, C_{in}. Also shown in Fig. 4.19 are the two major noise sources, namely the thermal current noise at the drain of Q_1 and the thermal noise from the feedback resistor, R_F. The 1/f noise of the transistor is ignored because it is assumed that the circuit is high speed, such that thermal noise dominates. The second (and perhaps subsequent) stage is modelled by an amplifier that has a gain of A_2. The noise due to this second stage is also ignored since the noise sources in the second stage are not amplified by the gain of the first stage.

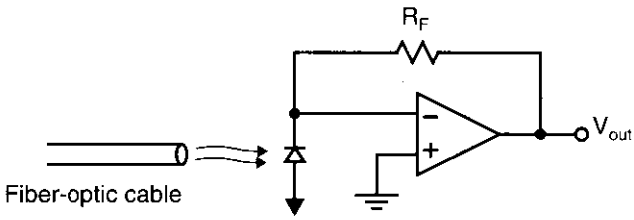


Fig. 4.18 A fiber-optic transresistance preamp.

7. Fiber-optic preamps are often realized using JFET transistors as well, but since their thermal noise model is identical to the noise model of the CMOS transistor, the analysis would be almost unchanged.

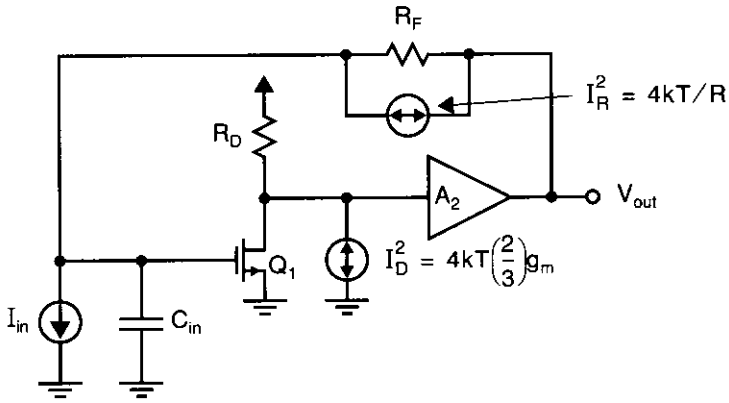


Fig. 4.19 A simplified model for a CMOS fiber-optic preamp.

A simplified small-signal model of this preamplifier, used for noise analysis, is shown in Fig. 4.20. The only parasitic capacitance considered in the transistor model is C_{gs}, since the gain of the first stage is only moderate, due to the resistor load, and therefore we can assume that C_{gd} and C_{db} can be ignored. The transfer function from i_{in} to the output is found by using nodal analysis to be

$$\frac{V_{out}(s)}{I_{in}(s)} = R_F\left(\frac{A_V}{1+A_V}\right)\frac{1}{1+s\left(\frac{R_F C_T}{1+A_V}\right)} \tag{4.104}$$

where $A_V = A_1 A_2$ is the total voltage gain of the preamp and $C_T = C_{gs} + C_{in}$. It is also found that this is the same transfer function from I_R^2 to the output. Thus, the noise current source I_R^2 can be replaced by an input current noise source having a spectral density of $4kT/R_F$. In a typical design, one would choose R_F as large as possible to limit this noise. However, this choice is constrained by the bandwidth requirements of passing the signal frequencies. From (4.104), we have the –3-dB frequency of the amplifier, given by

$$\omega_{-3\,dB} = \frac{1+A_V}{R_F C_T} \tag{4.105}$$

For a given amplifier gain and detector capacitance, R_F is chosen using (4.105) to place the –3-dB frequency as small as possible without substantially attenuating the signals. This makes the dominant node for determining stability the input node. The bandwidth of the second amplifier must be substantially greater than that of the input node to guarantee stability. Unfortunately, this constraint greatly amplifies the thermal noise due to input transistor Q_1, as we will see next.

The gain from the noise source I_D^2 to the output is found by using nodal analysis to be

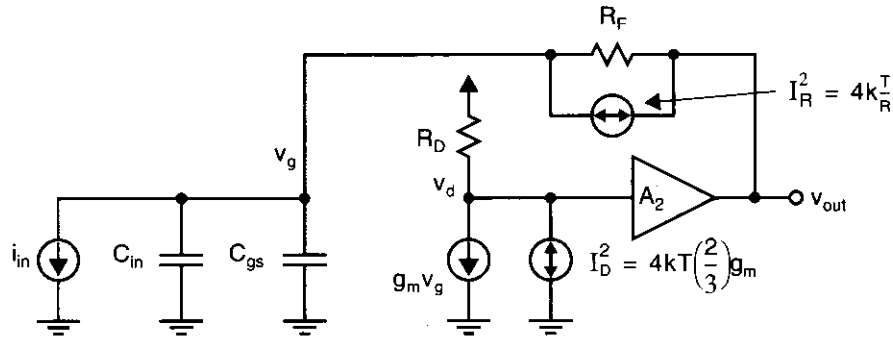


Fig. 4.20 A simplified small-signal model used for noise analysis.

$$\frac{V_{out}(s)}{I_D(s)} = \frac{1}{g_m}\frac{A_V}{1+A_V}\frac{1+sR_FC_T}{1+s\left[\frac{R_FC_T}{(1+A_V)}\right]} \tag{4.106}$$

At low frequencies, this gain is approximately given by $1/g_m$, whereas at high frequencies it is as much as $1 + A_V$ times greater. Thus, the high-frequency output noise due to Q_1 is much greater than the low-frequency noise due to Q_1. Furthermore, the only bandwidth limitation of this noise at high frequencies is due to the finite bandwidth of the second amplifier, A_2. From the discussion on compensation in Chapter 5, we know the second pole frequency of the amplifier must be almost four times greater than the closed-loop –3-dB frequency when lead compensation is not used.[8]

Continuing, in order to refer the noise due to Q_1 back to the input, we use (4.104) and (4.106) to obtain

$$I_{iD}^2(f) = I_D^2(f)\left|\frac{V_{out}(j\omega)}{I_D(jw)}\right|^2\left|\frac{I_{in}(j\omega)}{V_{out}(jw)}\right|^2 = I_D^2(f)\frac{1+\omega^2(R_FC_T)^2}{(g_mR_F)^2} \tag{4.107}$$

We use

$$I_D^2(f) = 4kT\left(\frac{2}{3}\right)g_m \tag{4.108}$$

and we add to this the input noise source that models the noise of the feedback resistor to obtain the total input-referred noise, given by

$$I_i^2(f) = \frac{4kT}{R_F} + 4kT\left(\frac{2}{3}\right)\frac{1+\omega^2(R_FC_T)^2}{g_mR_F^2} \tag{4.109}$$

Notice that the second term starts to quickly increase at a frequency of

8. Lead compensation could possibly be achieved by placing a capacitor in parallel with R_F. This is beyond the scope of this text.

$$\omega_z = \frac{1}{R_F C_T} = \frac{\omega_{-3\,dB}}{1 + A_V} \tag{4.110}$$

which is a relatively low frequency.

Continuing, normally $2/(3g_m R_F) << 1$, and we have

$$\begin{aligned} I_i^2(f) &= \frac{4kT}{R_F}\left(1 + \frac{2}{3g_m R_F}\right) + \frac{8kT}{3}\frac{\omega^2 C_T^2}{g_m} \\ &\cong \frac{4kT}{R_F} + \frac{8kT}{3}\frac{\omega^2 C_T^2}{g_m} \end{aligned} \tag{4.111}$$

Using the facts that

$$g_m = \mu_n C_{ox}\frac{W}{L}V_{eff} \tag{4.112}$$

and that

$$C_{gs} = \frac{2}{3}C_{ox}WL \tag{4.113}$$

we can write (4.111) as

$$I_i^2(f) \cong \frac{4kT}{R_F} + \frac{16}{9}kT\frac{L^2}{\mu_n V_{eff}}\omega^2\frac{(C_{gs} + C_{in})^2}{C_{gs}} \tag{4.114}$$

Normally, V_{eff} is taken as large as possible given power-supply-voltage and power-dissipation constraints. The only parameter left for the designer to choose is the width of the input transistor and, therefore, C_{gs}. It can be shown (by differentiating the second term of (4.114) with respect to C_{gs} and then setting the result to zero) that the second term of (4.114) is minimized by the choice $C_{gs} = C_{in}$, in which case we have

$$I_i^2(f) \cong 4kT\left(\frac{1}{R_F} + \frac{4}{9}\frac{L^2}{\mu_n V_{eff}}\omega^2 C_{in}\right) \tag{4.115}$$

4.5 REFERENCES

J. C. Bertails, "Low-Frequency Noise Considerations for MOS Amplifiers Design," *IEEE J. of Solid-State Circuits*, Vol. SC-14, No. 4, pp. 773–776, August, 1979.

A. Buchwald and K. Martin. *Integrated Fiber-Optic Receivers*. Kluwer Academic Publishers, Norwell, Massachusetts, 1995.

P. R. Gray and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, New York, 1993.

J. B. Johnson, *Phys. Rev.*, Vol. 32, pp. 97–109, 1928.

E. J. Kennedy, *Operational Amplifier Circuits: Theory and Applications*. Holt, Rinehart, & Winston, New Yo 1988.

A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, Reading, Massachusetts, 1989.

C. D. Motchenbacher and J. A. Connelly. *Low Noise Electronic System Design*. John Wiley & Sons, New York, 1993.

H. Nyquist, *Phys. Rev.*, Vol. 32, pp. 110–113, 1928.

W. Schottky, *Ann. Phys. (Leipzig)*, Vol. 57, pp. 541–567, 1918.

M. Steyaert, Z. Y. Chang, and W. Sansen. "Low-Noise Monolithic Amplifier Design: Bipolar versus CMOS," *Analog Integrated Circuits and Signal Processing 1*. Kluwer Academic Publishers, Boston, pp. 9–19, September 1991.

A. Van der Ziel, *Noise in Solid State Devices and Circuits*. John Wiley & Sons, New York, 1986.

4.6 PROBLEMS

Unless otherwise stated, assume dBm values are referenced to 50 Ω.

4.1 If a signal is measured to have $V_{(rms)}$ volts, what is the difference in db if it is expressed in dBm referenced to 50 Ω as opposed to being referenced to 75 Ω?

4.2 Consider the sum of two noise sources of values −20 dBm and −23 dBm. Find the total noise power in dBm for the cases in which the two noise sources are (*a*) uncorrelated, (*b*) $C = 0.3$, (*c*) $C = +1$, and (*d*) $C = -1$.

4.3 The output noise of a circuit is measured to be −40 dBm around 100 kHz when a resolution bandwidth of 30 Hz is used. What is the expected dBm measurement if a resolution bandwidth of 10 Hz is used? Find the root spectral density in $V/\sqrt{Hz}$.

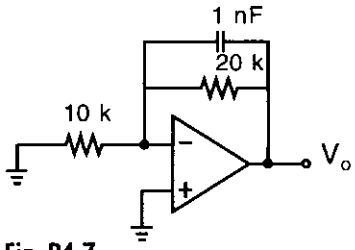
4.4 At 0.1 Hz, a low-frequency measurement has a noise value of −60 dBm when a resolution bandwidth of 1 MHz is used. Assuming 1/f noise dominates, what would be the expected noise value (in dBm) over the band from 1 MHz to 1 Hz?

4.5 Show that, when two resistors of values R_1 and R_2 are in series, their noise model is the same as a single resistance of value $R_1 + R_2$. Repeat the problem for parallel resistances.

4.6 Sketch the spectral density of voltage noise across a 100-pF capacitor when it is in parallel with a 1-kΩ resistor. Make another sketch for the same capacitor but with a 1-MΩ resistance in parallel. What can you say about the area under the curves of the two sketches?

4.7 Consider the circuit shown in Fig. P4.7 on p. 218, where the opamp has a unity-gain frequency of 1 MHz and equivalent input voltage and current noise sources of $V_n(f) = 20\ nV/\sqrt{Hz}$ and $I_n(f) = 10\ pA/\sqrt{Hz}$, respectively. Estimate the total rms noise voltage at V_o.

4.8 Assuming an ideal noiseless opamp, find the noise level at V_x and V_o for the opamp circuit shown in Fig. P4.8. By what factor is the noise value at V_o larger (or smaller) than $kT/1\ nF$? How do you account for this increase (or decrease)? Also explain why the noise value at V_x is smaller than $kT/1\ nF$.



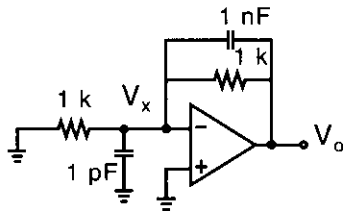
Opamp

$V_n(f) = 20\ nV/\sqrt{Hz}$

$I_{n+}(f) = I_{n-}(f) = 10\ pA/\sqrt{Hz}$

$f_t = 1\ MHz$

Fig. P4.7



Ideal noiseless opamp

Fig. P4.8

4.9 Consider an inductor of value L and an arbitrary resistor in parallel, as shown in Fig. P4.9. Show that the current noise, $i_{no}(t)$, has a noise value given by

$$I^2_{no(rms)} = \frac{kT}{L}$$

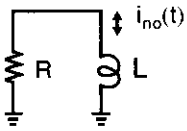


Fig. P4.9

4.10 The two circuits shown in Fig. P4.10 realize a first-order, low-pass filter. Assume the opamps are ideal and noiseless.

(*a*) Show that the two circuits have the same input-output transfer function.

(*b*) Estimate the total output noise for circuit I, using only dominant noise sources.

(*c*) Repeat (*b*) for circuit II.

4.11 Modify circuit I in Fig. P4.10 such that the new circuit has the same transfer function but uses an 80-pF capacitor instead of an 80-nF capacitor. If the opamp is ideal and noiseless, what is the new total output noise?

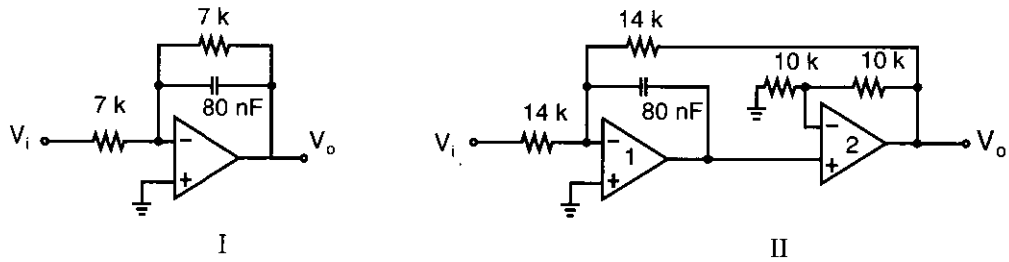


Fig. P4.10

4.12 Using the 1/f tangent principle, estimate the total noise above 0.1 Hz for the spectral density shown in Fig. 4.3.

4.13 Consider a bandpass amplifier that has equivalent input noise root spectral density and amplifier response, as shown in Fig. P4.13. Sketch the root spectral density for the output signal. Estimate the total output rms noise value by applying the 1/f tangent principle.

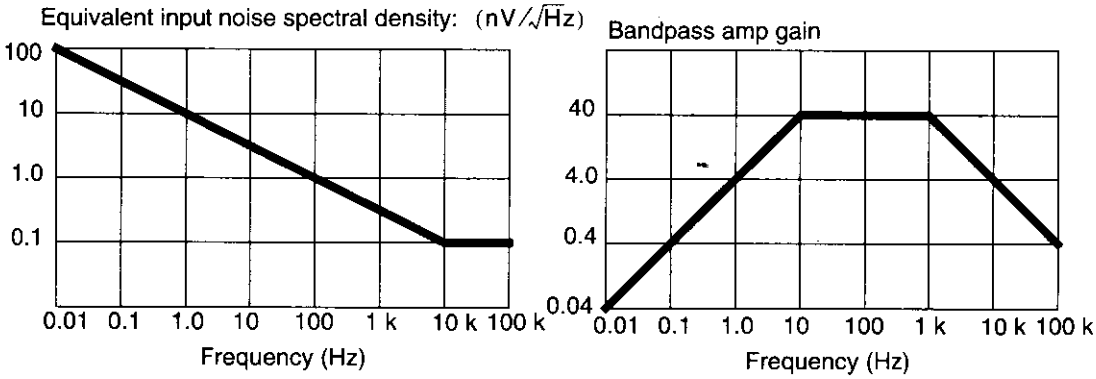


Fig. P4.13

4.14 Consider the noise root spectral density of the signal shown in Fig. P4.14. Find the total rms noise value using a graphical approach for 0.01 to ∞ Hz. Compare your result with that obtained when using the 1/f tangent principle.

4.15 We saw on page 193, Eq. (4.36), that the noise bandwidth of a first-order, low-pass filter is $(\pi/2)f_0$. Show that the noise bandwidth of a second-order, low-pass filter given by

$$A(s) = \frac{A_o}{\left(1 + \frac{s}{(2\pi f_0)}\right)^2}$$

is equal to $(\pi/4)f_o$.

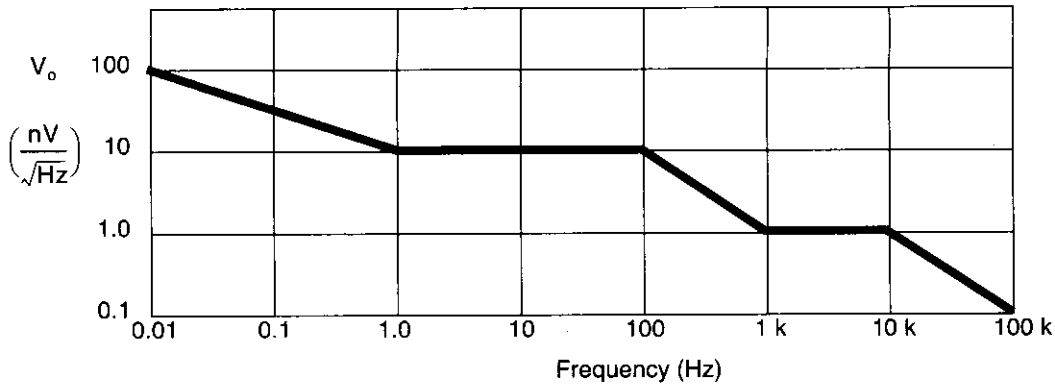


Fig. P4.14

4.16 Consider the two bipolar current mirrors shown in Fig. P4.16, where I_{in} is a 1-mA bias current plus a 100-μA(rms) signal current. Assuming that the base resistance for each transistor is $r_b = 330\ \Omega$ and dominates the output noise, estimate the resulting SNR (in dB) for the two current mirrors over a 50-MHz bandwidth (also assume the output noise is white).

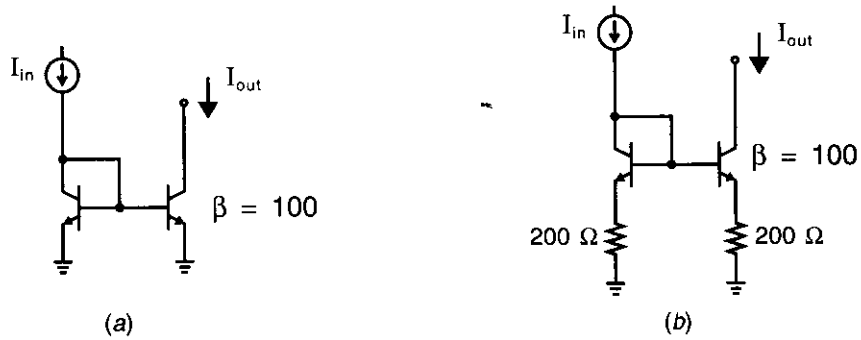


Fig. P4.16

4.17 Estimate the total output noise rms value for a low-pass filter, as shown in Fig. 4.15, when $C_f = 1$ nF, $R_f = 16$ k, $R_1 = 1.6$ k, and $R_2 = 0$. Also, find the SNR for an input signal equal to 100 mV rms. Assume that the noise voltage of the opamp is given by $V_n(f) = 20\ nV/\sqrt{Hz}$, that both its noise currents are $I_n(f) = 0.6\ pA/\sqrt{Hz}$, and that its unity-gain frequency equals 2 MHz.

4.18 Consider the CMOS differential input stage shown in Fig. 4.17, where Q_5 supplies a bias current of 100 μA, resulting in $g_{m1} = g_{m2} = 1$ mA/V and $g_{m3} = g_{m4} = 0.5$ mA/V. Find the equivalent input noise spectral density associated with thermal noise. If the bias current is doubled, how does the equivalent input noise density change?

CHAPTER

5 Basic Opamp Design and Compensation

This chapter describes the fundamental principles of basic opamp design. To illustrate many of these principles, the design of a traditional opamp—namely, the two-stage CMOS opamp—is used. This example illustrates compensation techniques needed to ensure stability in closed-loop amplifiers as well as to introduce a number of other important design techniques, such as ensuring zero systematic input-offset-voltage and process-insensitive lead compensation. Although the two-stage CMOS opamp is a classic circuit used in many integrated circuits, other more modern architectures have gained popularity recently. These more advanced architectures are described in Chapter 6. However, one should not overlook the potential of the two-stage opamp because fully-differential versions of this classic opamp are well-suited to low-voltage applications since they do not require cascode output stages. Fully-differential opamps are also described Chapter 6.

5.1 TWO-STAGE CMOS OPAMP

The two-stage circuit architecture has historically been the most popular approach for both bipolar and CMOS opamps, where a complementary process that has reasonable n-type and p-type devices is available. Although a CMOS version of this two-stage opamp is described here, a bipolar version is similar (but slightly more complicated). When properly designed, the two-stage opamp has a performance very close to more modern designs and is somewhat more suitable when resistive loads need to be driven.[1] Furthermore, it is an excellent example to illustrate many important design concepts that are also directly applicable to other more modern designs.

A block diagram of a typical two-stage CMOS opamp is shown in Fig. 5.1. "Two-stage" refers to the number of gain stages in the opamp. Figure 5.1 actually shows three stages—two gain stages and a unity-gain output stage. The output buffer is normally present only when resistive loads need to be driven. If the load is purely capacitive, then it is seldom included. The first gain stage is a differential-input single-ended output stage, often very similar to that shown in Fig. 3.19. The second gain stage is normally a common-source gain stage that has an active load, often very similar to that shown previously in Fig 3.4. Capacitor C_C is included to ensure stability

1. In a CMOS integrated circuit, opamp loads are often, but not always, purely capacitive.

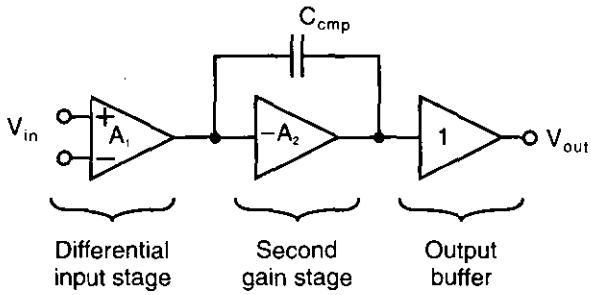


Fig. 5.1 A block diagram of a two-stage opamp.

when the opamp is used with feedback. Because C_C is between the input and the output of the high-gain second stage, it is often called a Miller capacitance since its effective capacitive load on the first stage is larger than its physical value.

An example of a practical CMOS version of the two-stage opamp is shown in Fig. 5.2. This example is used to illustrate many of the important design principles when realizing the two-stage amplifier.

It should be noted that the first stage has a p-channel differential input pair with an n-channel current-mirror active load. This is a complementary differential gain stage to that shown previously in Fig. 3.19. The trade-offs between having p-channel input transistors versus this stage and the alternative stage of Fig. 3.19 will be discussed later in this section. Also, the numbers next to the transistors represent reasonable transistor widths for a 1-μm process. Reasonable sizes for the lengths of the transistor might be somewhere between 1.5 and 2 times the minimum transistor length

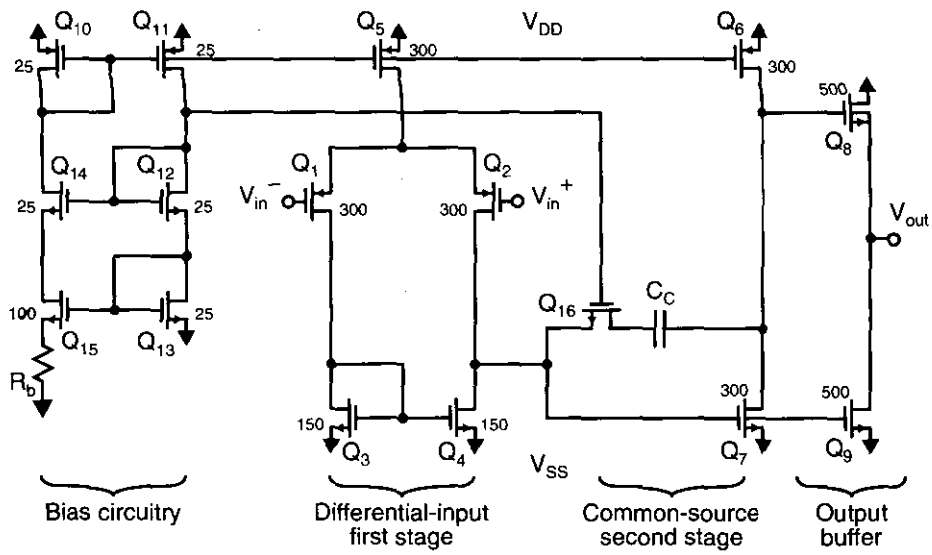


Fig. 5.2 A CMOS realization of a two-stage amplifier. All transistor lengths are 1.6 μm.

of a particular technology, whereas digital logic typically makes use of the minimum transistor length.

Opamp Gain

First we discuss the overall gain of the opamp. For low-frequency applications, this gain is one of the most critical parameters of an opamp.

The gain of the first stage has already been derived, resulting in (3.70), and is repeated here for convenience:

$$A_{v1} = g_{m1}(r_{ds2} \parallel r_{ds4}) \tag{5.1}$$

Recall from Chapter 1 that g_{m1} is given by

$$g_{m1} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 I_{D1}} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 \frac{I_{bias}}{2}} \tag{5.2}$$

Also, an approximation to the finite output impedance, r_{dsi}, of transistor Q_i is given by

$$r_{dsi} \approx \alpha \frac{L_i}{I_{Di}} \sqrt{V_{DGi} + V_{ti}} \tag{5.3}$$

where α is a technology-dependent parameter of around 5×10^6 $\sqrt{V}/m$. This equation is at best approximate and ignores all short-channel effects that become more important for shorter channel lengths in modern technologies.

The second gain stage is simply a common-source gain stage with a p-channel active load, Q_6. Its gain is given by $(-g_{m7})(r_{ds6} \parallel r_{ds7})$. Thus, we have

$$A_{v2} = -g_{m7}(r_{ds6} \parallel r_{ds7}) \tag{5.4}$$

The third stage is a common-drain buffer stage. This stage is often called a *source follower*, because the source voltage follows the gate voltage of Q_8, except for a level shift. As shown, the substrate of Q_8 is at the same voltage as the source of Q_8. This connection is only possible when Q_8 is realized in a well that is isolated from the bulk. Thus, for this example opamp, one would require a p-well or twin-well process to realize this output stage. Tying the substrate of Q_8 to its source eliminates gain degradations due to the body effect. This connection also results in a smaller dc voltage drop from the gate to the source of Q_8, which is a major limitation on the maximum positive output voltage. As Chapter 3 shows, the gain of this source-follower stage is given by

$$A_{v3} \cong \frac{g_{m8}}{G_L + g_{m8} + g_{ds8} + g_{ds9}} \tag{5.5}$$

where G_L is the load conductance being driven by the buffer stage. When it is not possible to tie the substrate of Q_8 to its source, as is the case when an n-well process is used (currently, a popular process), then the gain of the buffer stage is given by

$$A_{v3} \cong \frac{g_{m8}}{G_L + g_{m8} + g_{s8} + g_{ds8} + g_{ds9}} \tag{5.6}$$

where g_s is a body-effect conductance and is given by

$$g_s = \frac{g_m \gamma}{2\sqrt{V_{SB} + 2\phi_F}} \tag{5.7}$$

Voltage V_{SB} is the source-to-substrate (or bulk) voltage, γ is the body-effect constant (around $0.5\ V^{1/2}$), and $2\phi_F$ is twice the difference between the Fermi level in the bulk and the Fermi level of intrinsic silicon (around 0.7 V). Thus, g_s is around $g_m/5$.

EXAMPLE 5.1

Find the gain of the opamp shown in Fig. 5.2 using the following assumptions. Assume the bias current of the input differential pair is given by $I_{D5} = 100\ \mu A$, the power supplies are $V_{DD} = -V_{SS} = 2.5$ V, and $R_L = 10\ k\Omega$. Also assume the following process parameters: $\mu_n C_{ox} = 3\mu_p C_{ox} = 96\ \mu A/V^2$, $\gamma = 0.5\ V^{1/2}$, $\phi_F = 0.35$ V, $\alpha = 5 \times 10^6\ \sqrt{V}/m$, and $V_{tn} = -V_{tp} = 0.8$ V. To estimate output impedances, assume the drain-to-gate voltages of the first-stage transistors equal 0.5 V, whereas the transistors in the second and third stages have $V_{DGi} = 1$ V. Finally, do not assume the substrate of Q_8 is connected to its source—rather, assume it is connected to the negative power supply.

Solution

First the bias currents are calculated. Since $I_{D5} = 100\ \mu A$, we have

$$I_{D1} = I_{D2} = I_{D3} = I_{D4} = I_{D5}/2 = 50\ \mu A$$

$$I_{D6} = I_{D7} = (W_6/W_5)I_{D5} = 100\ \mu A$$

and

$$I_{D8} = I_{D9} = (W_9/W_7)I_{D7} = 167\ \mu A$$

We can now calculate the transconductances of Q_1, Q_2, Q_7, and Q_8. Using (5.2), we have $g_{m1} = g_{m2} = 0.775$ mA/V, $g_{m7} = 1.90$ mA/V, and $g_{m8} = 3.16$ mA/V.

Next, we need to estimate the output impedances of the transistors. To find these impedances, we use (5.3) together with the given approximations that the first-stage transistors have $V_{DGi} = 0.5$ V whereas the transistors in the other two stages have drain-to-gate voltages of 1 V. It should be mentioned here that these very rough drain-gate voltage approximations are reasonable since (5.3) is only of very moderate accuracy, perhaps 50 percent at best. These values for the transistor output impedances can be corrected later once a SPICE analysis is run. With the use of (5.3), we find that all of the transistors in the first stage have output impedances of

$$r_{ds1} = r_{ds2} = r_{ds3} = r_{ds4} = 182\ k\Omega$$

The output impedances of transistors in the second stage are

$$r_{ds6} = r_{ds7} = 107\ k\Omega$$

whereas the output impedances of transistors in the output stage are

$$r_{ds8} = r_{ds9} = 64\ k\Omega$$

The last parameter we need to calculate is the body-effect conductance of Q_8. Using (5.7) and assuming V_{out} is at ground results in $V_{SB8} = 2.5$ V, and thus we have $g_{S8} = 0.44$ mA/V.

Finally, using (5.1), we calculate that

$$A_{v1} = g_{m1}(r_{ds2} \| r_{ds4}) = 70.2\ \text{V/V}$$

Using (5.4),

$$A_{v2} = -g_{m7}(r_{ds6} \| r_{ds7}) = -102\ \text{V/V}$$

And using (5.6),

$$A_{v3} \cong \frac{g_{m8}}{G_L + g_{m8} + g_{s8} + g_{ds8} + g_{ds9}} = 0.85\ \text{V/V}$$

Thus, the total gain is equal to $A_{v1}A_{v2}A_{v3} = -6{,}090$ V/V. Once again, it should be mentioned here that this result is a rough approximation and should be verified using SPICE. The benefit of performing the hand calculations is to see how the gain is affected by different design parameters.

Frequency Response

We now wish to investigate the frequency response of the two-stage opamp at frequencies where the compensation capacitor, C_C, has caused the magnitude of the gain to begin to decrease, but still at frequencies well below the unity-gain frequency of the opamp. This corresponds to midband frequencies for many applications. This allows us to make a couple of simplifying assumptions. First, we will ignore all capacitors except the compensation capacitor, C_C, which normally dominates at all frequencies except around the unity-gain frequency of the opamp.[2] Second, we also assume that transistor Q_{16} is not present. This transistor operates as a resistor, which is included to achieve lead compensation and it has an effect only around the unity-gain frequency of the opamp, as we will see when we discuss compensation in Section 5.2. The simplified circuit used for analysis is shown in Fig. 5.3. It is worth mentioning that this simplified circuit is often used during system-level simulations when speed is more important than accuracy, but aspects such as slew-rate limiting (see next subsection) should be included in the simulation.

2. Recall that the unity-gain frequency of an opamp is that frequency where the magnitude of the open-loop opamp gain has decreased to one.

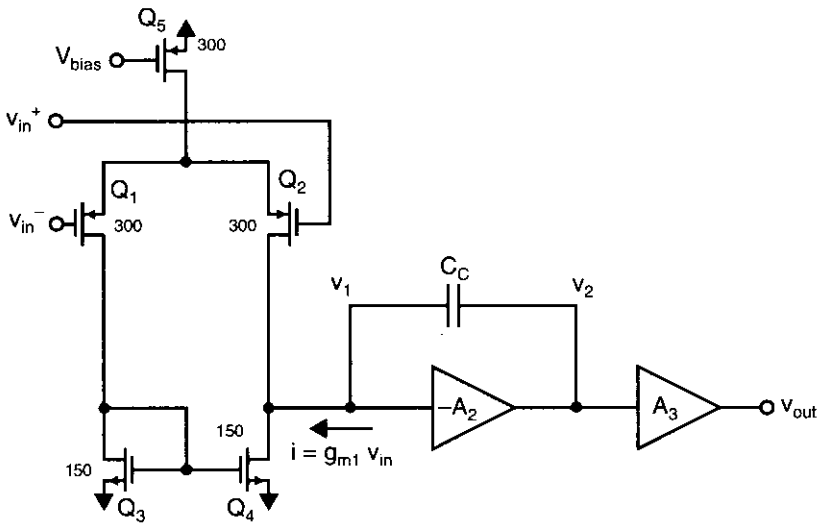


Fig. 5.3 A simplified model for the opamp used to find the midband frequency response.

The second stage introduces primarily a capacitive load on the first stage due to the compensation capacitor, C_C. Using Miller's Theorem [Sedra, 1991], one can show that the equivalent load capacitance, C_{eq}, at node v_1 is given by,

$$C_{eq} = C_C(1 + A_2) \approx C_C A_2 \tag{5.8}$$

The gain in the first stage can now be found using the small-signal model of Fig. 3.20, resulting in

$$A_1 = \frac{v_1}{v_{in}} = -g_{m1} Z_{out1} \tag{5.9}$$

where

$$Z_{out1} = r_{ds2} \parallel r_{ds4} \parallel \frac{1}{sC_{eq}} \tag{5.10}$$

For midband frequencies, the impedance of C_{eq} dominates, and we can write

$$Z_{out1} \cong \frac{1}{sC_{eq}} \cong \frac{1}{sC_C A_2} \tag{5.11}$$

For the overall gain, we have

$$A_v(s) \equiv \frac{v_{out}}{v_{in}} = A_3 A_2 A_1 \cong A_3 A_2 \frac{g_{m1}}{sC_C A_2} \tag{5.12}$$

using (5.9) and (5.11). If we further assume that $A_3 \cong 1$, then the overall gain, given in (5.12), simplifies to

$$A_v(s) = \frac{g_{m1}}{sC_C} \tag{5.13}$$

This simple equation can be used to find the approximate[3] unity-gain frequency. Specifically, to find the unity-gain frequency, ω_{ta}, we set $|A_v(j\omega_{ta})| = 1$, and solve for ω_{ta}. Performing such a procedure with (5.13), we obtain the following relationship:

$$\omega_{ta} = \frac{g_{m1}}{C_C} \tag{5.14}$$

Note here that the unity-gain frequency is directly proportional to g_{m1} and inversely proportional to C_C.

EXAMPLE 5.2

Using the same parameters as in Example 5.1, and assuming $C_C = 5$ pF, what is the unity-gain frequency in Hz?

Solution

Using $g_{m1} = 0.771$ mA/V and (5.14), we find that

$$\omega_{ta} = \frac{0.771 \times 10^{-3}}{5 \times 10^{-12}} = 154.2 \text{ Mrad/s}$$

Thus, we find that $f_{ta} = \omega_{ta}/(2\pi) = 24.5$ MHz.

Slew Rate

Another important high-frequency parameter of an opamp is its slew rate. The *slew rate* is the maximum rate at which the output changes when input signals are large. When the opamp of Fig. 5.2 is limited by its slew rate because a large input signal is present, all of the bias current of Q_5 goes into either Q_1 or Q_2, depending on whether v_{in} is negative or positive. When v_{in} is a large positive voltage, the bias current, I_{D5}, goes entirely through Q_1 and also goes into the current-mirror pair, Q_3, Q_4. Thus, the current coming out of the compensation capacitor, C_C, (i.e., I_{D4}) is simply equal to I_{D5} since Q_2 is off. When v_{in} is a large negative voltage, the current-mirror pair Q_3 and Q_4 is shut off because Q_1 is off, and now the bias current, I_{D5}, goes directly into C_C. In either case, the maximum current entering or leaving C_C is simply the total bias current, I_{D5}.

Defining the slew rate, SR, to be the maximum rate that v_2 can change, and recalling that $v_{out} \cong v_2$, we have

3. It is approximate because it is assumed that $A_3 \cong 1$ and that the higher-frequency poles are being ignored. Both of these effects will cause the unity-gain frequency to be slightly smaller than that calculated.

$$SR \equiv \left.\frac{dv_{out}}{dt}\right|_{max} = \frac{I_{C_C}|_{max}}{C_C} = \frac{I_{D5}}{C_C} \tag{5.15}$$

where we used the charge equation $q = CV$, which leads to $I = dq/dt = C(dV/dt)$. Since $I_{D5} = 2I_{D1}$, we can also write

$$SR = \frac{2I_{D1}}{C_C} \tag{5.16}$$

where I_{D1} is the original bias current of Q_1 with no signals present. Also, using (5.14), we have $C_C = g_{m1}/\omega_{ta}$, and substituting this into (5.16), we have

$$SR = \frac{2I_{D1}\omega_{ta}}{g_{m1}} \tag{5.17}$$

Recalling that

$$g_{m1} = \sqrt{2\mu_p C_{ox}\left(\frac{W}{L}\right)_1 I_{D1}} \tag{5.18}$$

we finally have another relationship for the slew-rate value.

$$SR = \frac{2I_{D1}}{\sqrt{2\mu_p C_{ox}(W/L)_1 I_{D1}}}\omega_{ta} = V_{eff1}\omega_{ta} \tag{5.19}$$

where

$$V_{eff1} = \sqrt{\frac{2I_{D1}}{\mu_p C_{ox}(W/L)_1}} \tag{5.20}$$

from Chapter 1. Normally one has little control over ω_{ta}, assuming a given maximum power dissipation is allowed. It is usually constrained to be less than two-thirds of the second-pole frequency, as we will see in Section 5.2, on compensation. As a result, *the only ways of improving the slew rate for the two-stage CMOS opamp is to increase* V_{eff1}, ω_{ta}, or both. As we will soon see, obtaining a high slew rate and unity-gain frequency are two of the major reasons for choosing p-channel input transistors rather than n-channel input transistors. It should be mentioned here that increasing V_{eff1} lowers the transconductance of the input stage. Although increasing V_{eff1} helps to minimize distortion, a lower transconductance in the first stage decreases the dc gain and increases the equivalent input thermal noise (see Chapter 4, Section 4.4).

EXAMPLE 5.3

Using the same parameters as in Example 5.1, and assuming $C_C = 5$ pF, what is the slew rate? What circuit changes could be made to double the slew rate but keep ω_{ta} and bias currents unchanged?

Solution

From (5.16), we have

$$SR = \frac{2 \times 50\ \mu A}{5\ pF} = 20\ V/\mu s \tag{5.21}$$

To double the slew rate while maintaining the same bias currents, C_C should be set to 2.5 pF. To maintain the same unity-gain frequency, g_{m1} should be halved, which can be accomplished by decreasing the widths of Q_1 and Q_2 by 4 (i.e., one could let $W_1 = W_2 = 75\ \mu m$).

Systematic Offset Voltage

When designing the two-stage CMOS opamp, if one is not careful, it is possible that the design will have an inherent (or systematic) input-offset voltage. Indeed, this was the case for many of the original designs used in production microcircuits. To see what is necessary to guarantee that no inherent input-offset voltage exists, consider the first two stages of the opamp, as shown in Fig. 5.4. To ensure that no systematic input-offset voltage exists, when the differential input voltage is zero (i.e, when $V_{in}^+ = V_{in}^-$), the output voltage of the first stage, V_{GS7}, should be that which is required to make I_{D7} equal to its bias current, I_{D6}. Specifically, the value of V_{GS7} should be given by

$$V_{GS7} = \sqrt{\frac{2I_{D6}}{\mu_n C_{ox}(W/L)_7}} + V_{tn} \tag{5.22}$$

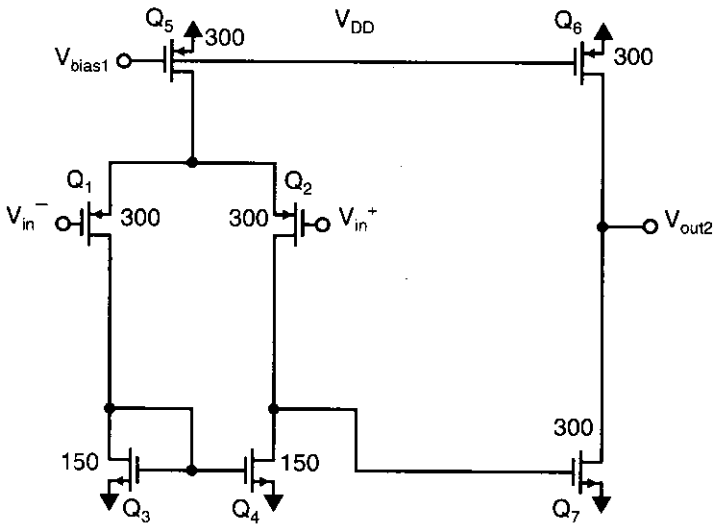


Fig. 5.4 The input and gain stages of the two-stage opamp.

When the differential input voltage is zero, the drain voltages of both Q_3 and Q_4 are equal by arguments of symmetry. Therefore, the output voltage of the first stage, V_{GS7}, is given by

$$V_{GS7} = V_{DS3} = V_{GS4} \tag{5.23}$$

This value is the voltage necessary to cause I_{D7} to be equal to I_{D6}, or, in other words, to satisfy (5.22). If this value is not achieved, then the output of the second stage (with Q_6, Q_7) would clip at either the negative or positive rail since this stage has such a high gain.[4] However, the gate-source voltage of Q_4 is given by

$$V_{GS4} = \sqrt{\frac{2I_{D4}}{\mu_n C_{ox}(W/L)_4}} + V_{tn} \tag{5.24}$$

so equating (5.22) and (5.24) to satisfy (5.23) results in

$$\sqrt{\frac{2I_{D4}}{\mu_n C_{ox}(W/L)_4}} = \sqrt{\frac{2I_{D6}}{\mu_n C_{ox}(W/L)_7}} \tag{5.25}$$

or, equivalently,

$$\frac{I_{D4}}{(W/L)_4} = \frac{I_{D6}}{(W/L)_7} \tag{5.26}$$

This equality, when the current density of Q_4 is equal to the current density of Q_7, guarantees that they both have the same effective gate-source voltages.

Since

$$\frac{I_{D6}}{I_{D4}} = \frac{I_{D6}}{I_{D5}/2} = \frac{(W/L)_6}{(W/L)_5/2} \tag{5.27}$$

we see that the necessary condition to ensure that no input-offset voltage is present is

$$\frac{(W/L)_7}{(W/L)_4} = 2\frac{(W/L)_6}{(W/L)_5} \tag{5.28}$$

which is satisfied for the sizes shown. Note that this analysis ignores the voltage drop of the level-shifter output stage (Q_8, Q_9) and any mismatches between the output impedance of p-channel and n-channel transistors. Fortunately, these effects cause only minor offset voltages, and by satisfying (5.28), offset voltages on the order of 5 mV or less can be obtained.

EXAMPLE 5.4

Consider the opamp of Fig. 5.2, where Q_3 and Q_4 are each changed to widths of 120 μm, and we want the output stage to have a bias current of 150 μA.

4. When the opamp is incorrectly designed to have an inherent input-offset voltage, feedback will cause the differential input voltage of the first stage to be equal to a nonzero voltage needed to ensure that the second stage does not clip.

Find the new sizes of Q_6 and Q_7, such that there is no systematic offset voltage.

Solution

Since I_{D6} determines the bias current of the output stage, and since it should have 50 percent more current than I_{D5}, its width should be 50 percent greater than W_5, resulting in

$$W_6 = 450\ \mu m \tag{5.29}$$

For Q_7, we use (5.28), which leads to

$$W_7 = \frac{2(450)(120)}{300} = 360\ \mu m \tag{5.30}$$

n-Channel or p-Channel Input Stage

The two-stage opamp discussed above has p-channel input transistors. It is also possible to realize a complementary opamp where the first stage has an n-channel differential pair and the second stage is a common-source amplifier having a p-channel input drive transistor. The choice of which configuration to use depends on a number of trade-offs that are discussed here.

First, the overall dc gain is largely unaffected by the choice since both designs have one stage with one or more n-channel driving transistors, and one stage with one or more p-channel driving transistors.

For a given power dissipation, and therefore bias current, having a p-channel input-pair stage maximizes the slew rate. This result is seen from (5.19) since p-channel input transistors for the first stage have a larger V_{eff} than would be the case for n-channel input transistors (assuming similar maximum widths have been chosen to maximize the gain). This slew-rate improvement can be one of the most important considerations, and thus most knowledgeable designers choose a p-channel input for the first stage.

Having a p-channel input first stage implies that the second stage has an n-channel input drive transistor. This arrangement maximizes the transconductance of the drive transistor of the second stage, which is critical when high-frequency operation is important. As we will see in the next section, the equivalent second pole, and therefore the unity-gain frequency as well, are both proportional to the transconductance of the second stage.

Another consideration is whether a p-channel or n-channel source-follower output stage is desired. Typically, an n-channel source follower is preferable because this will have less of a voltage drop. Also, since an n-channel transistor has a higher transconductance, the effect on the equivalent second pole due to its load capacitance is minimized, which is another important consideration. Finally, there is less degradation of the gain when small load resistances are being driven. The one disadvantage of having an n-channel source follower is that, for n-well processes, it is not possible to

connect the source to the substrate, thereby minimizing the voltage drop. Finally, note that, for opamps that drive purely capacitive loads, the buffer stage should not be included, and for this case, the output stage is clearly not a consideration.

Noise is another important consideration when choosing which input stage to use. Perhaps the major noise source of MOS opamps is due to 1/f noise caused by carriers randomly entering and leaving traps introduced by defects near the semiconductor surface. This 1/f noise source can be especially troublesome unless special circuit design techniques are used.[5] Typically, p-channel transistors have less 1/f noise than n-channel transistors since their majority carriers (holes) have less potential to be trapped in surface states. Thus, having a first-stage with p-channel inputs minimizes the output noise due to the 1/f noise. The same is not true when thermal noise is considered. When thermal noise is referred to the input of the opamp, it is minimized by using input transistors that have large transconductances (Chapter 4, Section 4.4), which unfortunately degrades the slew rate. However, when thermal noise is a major consideration, then a more modern architecture, such as a folded-cascode opamp (discussed in Chapter 6), is normally used.

In summary, when using a two-stage opamp, a p-channel input transistor for the first stage is almost always the best choice because it optimizes slew rate, unity-gain frequency, and minimizes 1/f noise, with the major disadvantage being an increase in wideband thermal noise.

5.2 FEEDBACK AND OPAMP COMPENSATION

This section discusses using opamps in closed-loop configurations and how to compensate an opamp to ensure that the closed-loop configuration is not only stable, but also has good settling characteristics. Although the two-stage opamp is used as an example, almost all the material discussed here applies to most other opamps as well.

Optimum compensation of opamps is typically considered to be one of the most difficult parts of the opamp design procedure. However, if the systematic approach taken here is used, then a straightforward procedure can be used that almost always results in a near-optimum compensation network.

Before discussing compensation, some properties of feedback and closed-loop amplifiers will first be reviewed.

First-Order Model of Closed-Loop Amplifier

A simple first-order model for the transfer function of a dominant-pole compensated opamp, $A(s)$, is given by

$$A(s) = \frac{A_0}{(1 + s/\omega_{p1})} \tag{5.31}$$

5. Some useful circuit techniques to reduce the effects of 1/f noise are correlated double sampling and chopper stabilization.

where A_0 is the dc gain of the opamp and ω_{p1} is the (real-axis) dominant pole. At this point, it should be mentioned that all the poles and zeros in this chapter occur on the real axis and are represented by the notations ω_p and ω_z, respectively.[6]

Recall that the definition of the unity-gain frequency of an opamp, ω_{ta}, is the frequency at which $|A(j\omega_{ta})| = 1$. We then have the following approximation, since $\omega_{ta} >> \omega_{p1}$:

$$|A(j\omega_{ta})| = 1 \cong \frac{A_0}{\omega_{ta}/\omega_{p1}} \tag{5.32}$$

Thus, we have the following important relationship for this first-order model:

$$\omega_{ta} \cong A_0\omega_{p1} \tag{5.33}$$

From here on, we will define ω_{ta} to be exactly equal to $A_0\omega_{p1}$, which is approximately equal to the unity-gain frequency of the opamp (assuming a first-order model for the opamp).[7] Substituting (5.33) into (5.31) for the case in which $\omega_{p1} << \omega << \omega_{ta}$, we have at midband frequencies

$$A(s) \cong \frac{\omega_{ta}}{s} \tag{5.34}$$

This approximate relationship is often used to analyze a closed-loop circuit for the effects of the opamp's finite bandwidth at midband frequencies.

An opamp with feedback can be modelled by the block diagram shown in Fig. 5.5 [Sedra, 1991]. Here, the feedforward amplifier, $A(s)$, models the open-loop response of the opamp. The feedback term, β, represents the feedback factor, which is assumed to be frequency independent; this is typically the case for amplifiers, but may not be the case for applications such as damped integrators. It can be shown using signal-flow graph analysis that the closed-loop gain, $A_{CL}(s)$, for this model is given by

$$A_{CL}(s) = \frac{A(s)}{1 + \beta A(s)} \tag{5.35}$$

At midband frequencies, the transfer function of the closed-loop amplifier may be found by substituting (5.34) into (5.35), resulting in

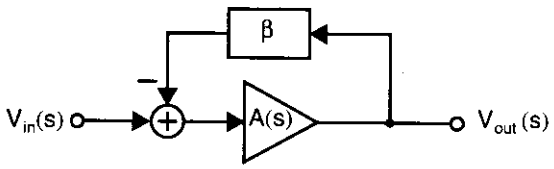


Fig. 5.5 A block diagram of a feedback circuit.

6. To be more precise, ω_p and ω_z are the inverses of the coefficients multiplying the s terms in the denominators and numerators, respectively. The actual poles and zeros are the negatives of these terms.
7. The unity-gain frequency of an optimally compensated opamp can be very different from ω_{ta} due to high-frequency poles and zeros that are ignored in the first-order model, especially when the closed-loop configuration has large gain. However, at midband frequencies (5.34) is still valid.

$$A_{CL}(s) \approx \frac{\omega_{ta}}{\beta\omega_{ta} + s} = \frac{1}{\beta}\frac{1}{(1 + s/\beta\omega_{ta})} \tag{5.36}$$

Thus, the closed-loop amplifier has a closed-loop gain at low frequencies approximately equal to $1/\beta$[8] and it has a –3-dB frequency given by

$$\omega_{-3\ dB} \cong \beta\omega_{ta} \tag{5.37}$$

It will be shown shortly that, for optimally compensated opamps, $\beta\omega_{ta} \cong \omega_t$, where ω_t is the unity-gain frequency of the open-loop transfer function. It will also be shown that, for optimally compensated amplifiers, ω_t is independent of the feedback factor, β, and dependent only on high-frequency poles and zeros. *Thus, the –3-dB frequency of the closed-loop gain of an optimally compensated amplifier is approximately equal to the unity-gain frequency of the loop gain, independent of the actual closed-loop gain.* However, the component sizes of the compensation network and ω_{ta} will be shown to depend on the desired closed-loop gain for an optimally compensated amplifier.

Linear Settling Time

The settling-time performance of integrated amplifiers is often an important design parameter. For example, in switched-capacitor circuits, the charge from one or more capacitors must be mostly transferred to a feedback capacitor within about half a clock period. This charge transfer is closely related to the opamp's step response. As a result, the *settling time* is defined to be the time it takes for an opamp to reach a specified percentage of its final value when a step input is applied.

This settling time consists of two distinct segments—linear and nonlinear settling time segments. The linear settling-time portion is due to the finite unity-gain frequency of the opamp, and thus it sets a minimum value for the overall settling time independent of the step size of the opamp's output. In contrast, the nonlinear settling time is due to slew-rate limiting, and thus this portion is strongly dependent on the output's step size. For example, for small step sizes in the output signal's level, the opamp may not reach a slew-rate limit at all, resulting in a nonlinear settling time of zero. Here, we discuss only the linear settling time by modelling the opamp as ideal but with a finite unity-gain frequency. Another simplification is to use the first-order opamp model, which has a 90-degree phase margin. Such a simplification results in a simple settling behavior that can be easily analyzed. Thus, the results here are used only to estimate the necessary unity-gain frequency for a circuit to settle within the linear settling-time segment. Simulations should be used in the latter parts of a design to determine more accurate settling-time estimates.

From (5.36), we recognize that the time constant of the closed-loop amplifier, τ, is given by

8. For inverting configurations, (5.35) should be modified to be $A_{CL}(s) = KA(s)/[1 + \beta A(s)]$, and the closed-loop gain at low frequencies is given by K/β. But these facts have no effect on the discussions to follow regarding optimum compensation because we are primarily interested in the loop gain, which is still given by $\beta A(s)$.

$$\tau = \frac{1}{\omega_{-3\,dB}} = \frac{1}{\beta\omega_{ta}} \tag{5.38}$$

The important result here is that the –3-dB frequency determines the settling-time response for a step input. Recall that the transient response of any first-order circuit is given by

$$x(t) = x(\infty) - [x(\infty) - x(t)]e^{-t/\tau} \tag{5.39}$$

where τ is the time constant of the circuit. For the closed-loop amplifier, the step response is found using (5.39) to be given by

$$v_{out}(t) = V_{step}(1 - e^{-t/\tau}) \tag{5.40}$$

Here, V_{step} is the size of the voltage step. With this exponential relationship, the time required for a first-order circuit to settle to within a specified value can be found. For example, if 1 percent accuracy is required, then one must allow $e^{-t/\tau}$ to reach 0.01, which is achieved at a time of 4.6τ. For settling to within a 0.1 percent accuracy, the settling time needed becomes approximately 7τ. Also, note that just after the step input, the slope of the output will be at its maximum, given by

$$\left.\frac{d}{dt}v_{out}(t)\right|_{t=0} = \frac{V_{step}}{\tau} \tag{5.41}$$

If the slew rate of the opamp is larger than this value, no slew-rate limiting would occur.

EXAMPLE 5.5

One phase of a switched-capacitor circuit is shown in Fig. 5.6, where the input signal can be modelled as a voltage step. If 0.1 percent accuracy is needed in the linear settling-time portion corresponding to 0.1 μs, find the required unity-gain frequency in terms of the capacitance values, C_1 and C_2. For $C_2 = 10C_1$, what is the necessary unity-gain frequency of the opamp? What unity-gain frequency is needed in the case in which $C_2 = 0.2C_1$?

Solution

We first note that a capacitive feedback network is used rather than a resistive one. A difficulty with this network in a nonswitched circuit is that no bias current

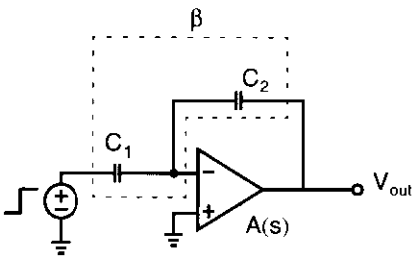


Fig. 5.6 One phase of a switched-capacitor circuit.

flows into the negative opamp input terminal. However, in a switched circuit using a CMOS opamp, this connection shown occurs only for a short time and does not cause any problems. Second, note that the feedback factor is simply a constant value since it is determined by a capacitive divider consisting of C_1 and C_2.

The feedback factor, β, is given by

$$\beta = \frac{1/(sC_1)}{1/(sC_1) + 1/(sC_2)} = \frac{C_2}{C_1 + C_2} \tag{5.42}$$

Now for 7τ settling within 0.1 μs, we see that τ must be less than 14.2 ns, and, since $\tau = 1/(\beta\omega_{ta})$, we see that

$$\omega_{ta} \geq \left(\frac{C_1 + C_2}{C_2}\right)\left(\frac{1}{14.2\text{ ns}}\right) \tag{5.43}$$

For the case in which $C_2 = 10C_1$, a unity-gain frequency of 12.3 MHz is needed, whereas in the case of $C_2 = 0.2C_1$, f_{ta} should be larger than 66.8 MHz.

Opamp Compensation

When one is compensating an opamp, the first-order model given by (5.31) is insufficient. It ignores high-frequency poles and zeros, which cause possible instability, even though they may be at frequencies greater than ω_t. It is important to *accurately* model the open-loop transfer function at higher frequencies where the loop-gain is approximately unity. Fortunately, when all poles and zeros are on the real axis, they can be modelled reasonably well by a single additional pole. Specifically, we can model $A(s)$ by

$$A(s) = \frac{A_0}{(1 + s/\omega_{p1})(1 + s/\omega_{eq})} \tag{5.44}$$

where ω_{p1} is the first dominant-pole frequency and ω_{eq} is the pole frequency that models higher-frequency poles. The relationship to approximate ω_{eq} given a set of real-axis poles, ω_{pi}, and zeros, ω_{zi}, is given by

$$\frac{1}{\omega_{eq}} \cong \sum_{i=2}^{m} \frac{1}{\omega_{pi}} - \sum_{i=1}^{n} \frac{1}{\omega_{zi}} \tag{5.45}$$

It should be noted here that the approximation in (5.45) is different than that given in [Sedra, 1991] since we are mostly interested in the phase shifts, rather than the attenuation, due to higher-frequency poles and zeros. In practice, ω_{eq} is found from simulation as the inverse of the frequency at which the transfer function has a –135° phase shift (–90° due to the dominant pole and another –45° due to the higher-frequency poles and zeros).

At frequencies much greater than the dominant pole frequency, $\omega >> \omega_{p1}$, we see that $1 + j\omega/\omega_{p1} \cong j\omega/\omega_{p1}$, and so (5.44) can be accurately approximated by

$$A(s) \cong \frac{\omega_{ta}}{s(1 + s/\omega_{eq})} \tag{5.46}$$

Note that this approximation result is especially valid at the unity-gain frequency of the loop (which we are presently interested in) since, from (5.33), $\omega_t/\omega_{p1} \cong (1/\beta)A_0$, which is almost certainly much greater than 1.

When the opamp is included in a circuit that has a feedback factor, β, then, at frequencies around ω_{ta}, the loop gain, $LG(s)$, is given by

$$LG(s) = \beta A(s) = \frac{\beta\omega_{ta}}{s(1 + s/\omega_{eq})} \tag{5.47}$$

The unity-gain frequency, ω_t, of the loop gain, $LG(s)$, can now be found by setting the magnitude of (5.47) equal to unity after substituting $s = j\omega_t$. Once this is done and the equation is rearranged, one can write,

$$\beta\frac{\omega_{ta}}{\omega_{eq}} = \frac{\omega_t}{\omega_{eq}}\sqrt{1 + \left(\frac{\omega_t}{\omega_{eq}}\right)^2} \tag{5.48}$$

This equation will be used shortly to relate a specified phase margin to the Q factor of the closed-loop configuration.

Note also that from (5.48) we have

$$\omega_{ta} = \frac{\omega_t}{\beta}\sqrt{1 + \left(\frac{\omega_t}{\omega_{eq}}\right)^2} \cong \frac{\omega_t}{\beta} \tag{5.49}$$

for the special case in which the unity-gain frequency is much less than the equivalent nondominant pole frequency (i.e., $\omega_t << \omega_{eq}$).

The phase margin, PM, is an often-used measure for how far an opamp with feedback is from becoming unstable. PM is defined as the difference between the loop-gain phase shift and –180°. From (5.47), the phase shift, $\angle LG(j\omega)$, is found as

$$\angle LG(j\omega) = -90° - \tan^{-1}(\omega/\omega_{eq}) \tag{5.50}$$

This equation implies that at the unity-gain frequency, $s = j\omega_t$, we have

$$PM = \angle LG(j\omega_t) - (-180°) = 90° - \tan^{-1}(\omega_t/\omega_{eq}) \tag{5.51}$$

and, therefore,

$$\omega_t/\omega_{eq} = \tan(90° - PM) \tag{5.52}$$

which implies that

$$\omega_t = \tan(90° - PM)\omega_{eq} \tag{5.53}$$

Equation (5.52) can be used to derive ω_t/ω_{eq} for a specified phase margin. *Note that the unity-gain frequency of the loop is independent of the feedback factor, β, and therefore, in the case of an optimally compensated amplifier, it is independent of the closed-loop gain as well.*

EXAMPLE 5.6

A closed-loop amplifier is compensated to have a 75° phase margin for $\beta = 1$. What is ω_t if $f_{eq} = \omega_{eq}/(2\pi) = 50\,\text{MHz}$? What is ω_{ta}?

Solution

Using (5.53), we have $\omega_t = 0.268\omega_{eq}$, which implies that the loop-gain unity-gain frequency, f_t, is given by $f_t = \omega_t/2\pi = 13.4$ MHz. Using (5.49), we also have, for $\beta = 1$, $\omega_{ta} = \omega_t\sqrt{1 + 0.268^2} = 1.035\omega_t$, which implies that $f_{ta} = \omega_{ta}/2\pi = 13.9$ MHz.

When considering optimum compensation, (5.34) is not accurate enough and one must use the more accurate relationship for $A(s)$ given in (5.46). Assuming β is frequency independent, substituting (5.44) into (5.35), and rearranging gives

$$A_{CL}(s) = \frac{A_{CL0}}{1 + \dfrac{s(1/\omega_{p1} + 1/\omega_{eq})}{1 + \beta A_0} + \dfrac{s^2}{(1 + \beta A_0)(\omega_{p1}\omega_{eq})}} \tag{5.54}$$

where

$$A_{CL0} = \frac{A_0}{1 + \beta A_0} \cong \frac{1}{\beta} \tag{5.55}$$

Thus, the closed-loop response near the unity-gain frequency, ω_t, is closely approximated by a second-order transfer function. It should be noted here that this approximation is inaccurate in the case where the high-frequency poles and zeros are quite close to ω_t. However, in this case, it would be extremely difficult to adequately compensate the opamp, and thus, this case is ignored.

The result of (5.54) can be equated to the general equation for a second-order all-pole transfer function, written as

$$H_2(s) = \frac{K\omega_0^2}{s^2 + \left(\dfrac{\omega_0}{Q}\right)s + \omega_0^2} = \frac{K}{1 + \dfrac{s}{\omega_0 Q} + \dfrac{s^2}{\omega_0^2}} \tag{5.56}$$

Recall that parameter ω_0 is called the resonant frequency and parameter Q is called the Q factor[9] [Sedra, 1991]. It is well known that, if $Q = \sqrt{1/2} \cong 0.707$, then the magnitude of the transfer function will have the widest passband without any peaking. Furthermore, for $Q = \sqrt{1/2}$, the –3-dB frequency is equal to ω_0. When the step response is investigated, restrictions on the Q factor can also be found to guar-

9. The Q factor is 1/2 times the inverse of the damping factor. The damping factor is an alternative method of indicating the pole locations in second-order transfer functions.

antee no peaking. Specifically, for there to be no peaking in the step response, it is necessary that both poles be real, which is equivalent to the requirement that $Q \leq 0.5$. In the case where $Q > 0.5$, the percentage overshoot of the output voltage is given by

$$\% \text{ overshoot} = 100e^{\frac{-\pi}{\sqrt{4Q^2-1}}} \tag{5.57}$$

Equating (5.54) with (5.56) and solving for ω_0 and Q results in

$$\omega_0 = \sqrt{(1+\beta A_0)(\omega_{p1}\omega_{eq})} \cong \sqrt{\beta\omega_{ta}\omega_{eq}} \tag{5.58}$$

and

$$Q = \frac{\sqrt{(1+\beta A_0)/\omega_{p1}\omega_{eq}}}{1/\omega_{p1} + 1/\omega_{eq}} \cong \sqrt{\frac{\beta A_0 \omega_{p1}}{\omega_{eq}}} = \sqrt{\frac{\beta\omega_{ta}}{\omega_{eq}}} \tag{5.59}$$

where the approximation of Q is valid since $\beta A_0 >> 1$ and $\omega_{p1} << \omega_{eq}$.

It is now possible to relate a specified phase margin to the Q factor. Equation (5.52) can be used to find ω_t/ω_{eq}. This result can be substituted into (5.48) to find $\beta(\omega_{ta}/\omega_{eq})$, which can then be substituted into (5.59) to find the equivalent Q factor. Finally, (5.57) can be used to find the corresponding percentage overshoot for a step input. This procedure gives us the information in Table 5.1.

Table 5.1 leads to some interesting observations. First, a frequency response with $Q \cong \sqrt{1/2}$ roughly corresponds to a phase margin of 65°. Therefore, one should design for a phase margin of at least 65°, given both process and temperature changes. Second, if one wants to ensure that there is no overshoot for a step input, then the phase margin should be at least 75°, again, given both process and temperature variations. Normally, values of 80° to 85° should be the nominal phase margin to account for these variations. These phase margins are much larger than what was traditionally thought to be necessary.

Finally, it is worth mentioning here that, when the feedback network is frequency independent and less than unity, (i.e., when $\beta \leq 1$), the worst-case phase margin occurs for $\beta = 1$. Thus, for a general-purpose opamp where $0 < \beta \leq 1$, if the opamp is compensated for $\beta = 1$, it is guaranteed to be stable for all other β, although it will not be optimally compensated and will be slower than necessary.

Table 5.1 The relationship between PM, ω_t/ω_{eq}, Q factor, and percentage overshoot

PM (Phase margin)	ω_t/ω_{eq}	Q factor	Percentage overshoot for a step input
55°	0.700	0.925	13.3%
60°	0.580	0.817	8.7%
65°	0.470	0.717	4.7%
70°	0.360	0.622	1.4%
75°	0.270	0.527	0.008%

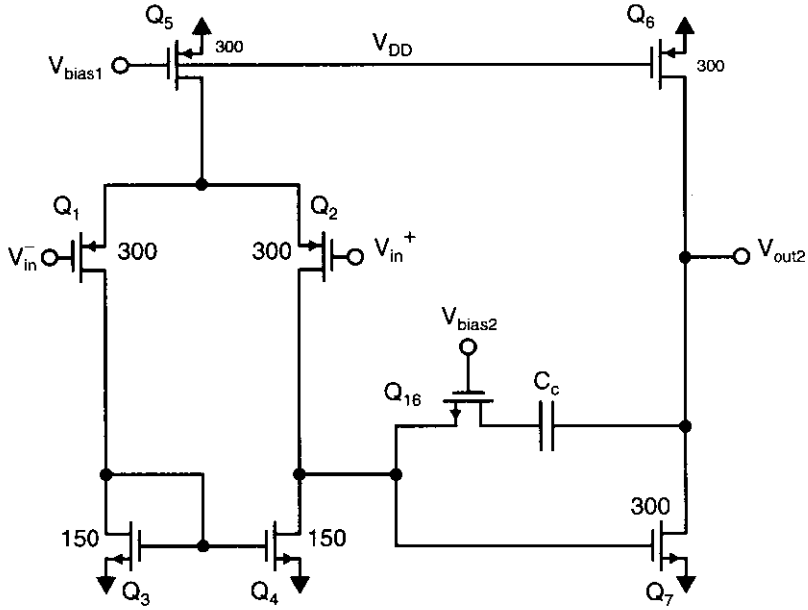


Fig. 5.7 The first two stages of the two-stage opamp, including the compensation network.

Compensating the Two-Stage Opamp

It is now possible to consider compensation of the two-stage opamp. To make the discussion easier to visualize, the first two stages of the opamp, including the compensation network, are shown in Fig. 5.7.

The capacitor, C_C, realizes what is commonly called dominant-pole compensation. It controls the dominant first pole, (i.e., ω_{p1}), and thereby the frequency, ω_{ta}, since (repeating (5.33) here for convenience),

$$\omega_{ta} = A_0\omega_{p1} \tag{5.60}$$

Transistor Q_{16} has $V_{DS16} = 0$ since no dc bias current flows through it, and therefore Q_{16} is hard in the triode region. Thus, this transistor operates as a resistor, R_C, of value given by

$$R_C = r_{ds16} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{16} V_{eff16}} \tag{5.61}$$

This transistor is included in order to realize a left-half-plane zero at frequencies around or slightly above ω_t. Without Q_{16}, we have a right-half-plane zero, which makes compensation much more difficult. The addition of such an extra left-half-plane zero is what is commonly called *lead-compensation*. It should be noted here that r_{ds} indicates the drain-source resistance of Q_{16} when it is in the triode region as opposed to the finite-output impedance of Q_{16} when it is in the active mode. The same notation, r_{ds}, is used to indicate the drain-source resistance in both cases—whether the transis-

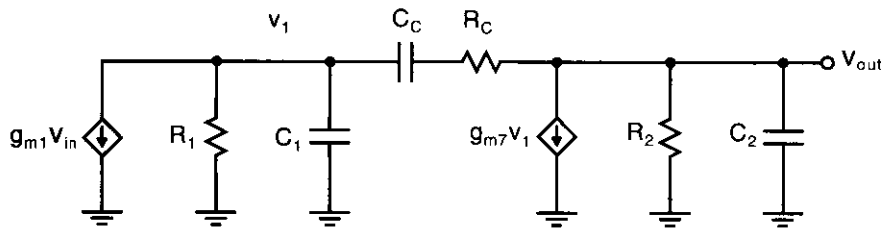


Fig. 5.8 A small-signal model of the two-stage opamp used for compensation analysis.

tor is in the active or the triode region. One simply has to check which region a transistor is operating in to ensure that the correct equation is used to determine r_{ds}.

A simplified small-signal model for this opamp is shown in Fig. 5.8, where the output buffer has again been ignored. Also, it is assumed that the first stage is much faster than the second stage and can therefore be modelled by a simple voltage-controlled current source. In the small-signal model, we have

$$R_1 = r_{ds4} \parallel r_{ds2} \tag{5.62}$$

$$C_1 = C_{db2} + C_{db4} + C_{gs7} \tag{5.63}$$

$$R_2 = r_{ds6} \parallel r_{ds7} \tag{5.64}$$

$$C_2 = C_{db7} + C_{db6} + C_{L2} \tag{5.65}$$

Note that if no output buffer is present, then C_{L2} is the output load capacitance. If an output buffer is present, then C_{L2} is the load capacitance introduced by the output buffer and C_1 also includes the capacitance needed to drive Q_9 in Fig. 5.2.

To show the need for R_C, we first assume $R_C = 0$ and perform nodal analysis at the nodes designated by v_1 and v_{out}. The following transfer function is obtained:

$$\frac{v_{out}}{v_{in}} = \frac{g_{m1}g_{m7}R_1R_2\left(1 - \dfrac{sC_C}{g_{m7}}\right)}{1 + sa + s^2b} \tag{5.66}$$

where

$$a = (C_2 + C_C)R_2 + (C_1 + C_C)R_1 + g_{m7}R_1R_2C_C \tag{5.67}$$

and

$$b = R_1R_2(C_1C_2 + C_1C_C + C_2C_C) \tag{5.68}$$

It is possible to find approximate equations for the two poles based on the assumption that the poles are real and widely separated.[10] This assumption allows us to express the denominator, $D(s)$, as[11]

10. If this assumption is not valid, then it is extremely difficult to properly compensate the opamp for unity-gain stability.

11. Note that the notation $[1 + (s/\omega_x)]$ implies that the root is in the left half plane if $\omega_x > 0$, whereas the root is in the right half plane if $\omega_x < 0$.

$$D(s) = \left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{p2}}\right) \cong 1 + \frac{s}{\omega_{p1}} + \frac{s^2}{\omega_{p1}\omega_{p2}} \tag{5.69}$$

Setting the coefficients of (5.66) equal to the coefficients of (5.69) and solving for ω_{p1} and ω_{p2} results in the following relationships. The dominant pole, ω_{p1}, is given by

$$\begin{aligned}\omega_{p1} &\cong \frac{1}{R_1[C_1 + C_C(1 + g_{m7}R_2)] + R_2(C_2 + C_C)} \\ &\cong \frac{1}{R_1C_C(1 + g_{m7}R_2)} \\ &\cong \frac{1}{g_{m7}R_1R_2C_C}\end{aligned} \tag{5.70}$$

whereas the nondominant pole, ω_{p2}, is given by

$$\begin{aligned}\omega_{p2} &\cong \frac{g_{m7}C_C}{C_1C_2 + C_2C_C + C_1C_C} \\ &\cong \frac{g_{m7}}{C_1 + C_2}\end{aligned} \tag{5.71}$$

Note that, from (5.66), another zero, ω_z, is located in the right half plane and is given by

$$\omega_z = \frac{-g_{m7}}{C_C} \tag{5.72}$$

This zero should come as no surprise since there are two signal paths having opposite signs in the small-signal circuit of Fig. 5.8—one through the voltage-controlled current source, $g_{m7}v_1$, and the other through the compensation capacitor, C_C.

From the preceding relationships for ω_{p1} and ω_{p2}, we can see that, as g_{m7} increases, the separation between the first and second poles increases. This separation tends to make the circuit more stable; hence, the use of a Miller capacitance for compensation is often called *pole-splitting compensation.* Also note that increasing C_C moves the dominant pole, ω_{p1}, to a lower frequency without affecting the second pole, ω_{p2}. This effect also makes the opamp more stable.

However, a problem arises due to the right-half-plane zero, ω_z. Because the zero is in the right half plane, it introduces negative phase shift, or phase lag, in the transfer function of the opamp. This makes stability more difficult. Making C_C larger does not help matters because this decreases the frequencies of both the first pole and the zero without making them more widely separated. Indeed, because of this right-half-plane zero, it is often impossible to choose C_C such that the step response has no overshoot (assuming $R_C = 0$). Fortunately, all is not lost, since introducing R_C allows adequate compensation, as we discuss in the next subsection.

Lead Compensation

If the small-signal model of Fig. 5.8 is reanalyzed with a nonzero R_C, then a third-order denominator results. The first two poles are still approximately at the frequencies

given by (5.70) and (5.71). The third pole is at a high frequency and has almost no effects. However, the zero is now determined by the relationship,

$$\omega_z = \frac{-1}{C_C(1/g_{m7} - R_C)} \tag{5.73}$$

This result allows the designer a number of possibilities. One could take

$$R_C = 1/g_{m7} \tag{5.74}$$

to eliminate the right-half-plane zero altogether. Alternatively, one could choose R_C to be even larger and thus move the right-half-plane zero into the left half plane to cancel the nondominant pole, ω_{p2}. Setting (5.73) equal to (5.71) and solving for R_C results in the following equation for R_C:

$$R_C = \frac{1}{g_{m7}}\left(1 + \frac{C_1 + C_2}{C_C}\right) \tag{5.75}$$

Unfortunately, C_2 is often not known a priori, especially when no output stage is present.

The third possibility (recommended by the authors) is to choose R_C even larger yet to move the now left-half-plane zero to a frequency slightly greater than the unity-gain frequency that would result if the lead resistor were not present—say, 20 percent larger [Roberge, 1975]. For this case, one should satisfy the following equation:

$$\omega_z = 1.2\omega_t \tag{5.76}$$

Assuming $R_C >> (1/g_{m7})$, then $\omega_z \cong 1/(R_C C_C)$, and recalling from (5.14) that $\omega_t \cong g_{m1}/C_C$, then one should choose R_C according to

$$R_C \cong \frac{1}{1.2 g_{m1}} \tag{5.77}$$

This approach leads to the following design procedure for lead compensation of a two-stage CMOS opamp:

1. Start by choosing, somewhat arbitrarily, $C'_C \cong 5$ pF.
2. Using SPICE, find the frequency at which a $-125°$ phase shift exists. Let the gain at this frequency be denoted A'. Also, let the frequency be denoted ω_t. This is the frequency that we would like to become the unity-gain frequency of the loop gain.
3. Choose a new C_C so that ω_t becomes the unity-gain frequency of the loop gain, thus resulting in a $55°$ phase margin. (Obtaining this phase margin is the reason we chose $-125°$ in step 2.) This can be achieved by taking C_C according to the equation

$$C_C = C'_C A' \tag{5.78}$$

It might be necessary to iterate on C_C a couple of times using SPICE.

4. Choose R_C according to

$$R_C = \frac{1}{1.2\omega_t C_C} \tag{5.79}$$

This choice will increase the unity-gain frequency by about 20 percent, leaving the zero near to the final resulting unity-gain frequency, which will end up about 15 percent below the equivalent second-pole frequency. The resulting phase margin is approximately 85°.[12] This allows a margin of 5° to account for processing variations without the poles of the closed-loop response becoming real. *This choice is also almost optimum lead compensation for almost any opamp when a resistor is placed in series with the compensation capacitor.* It might be necessary to iterate on R_C a couple of times to optimize the phase margin. However, one should check that the gain continues to steadily decrease at frequencies above the new unity-gain frequency; otherwise, the transient response can be poor. This situation sometimes occurs when unexpected zeros at frequencies only slightly greater than ω_t are present.

5. If, after step 4, or the phase margin is not adequate, then increase C_C while leaving R_C constant. This will move both ω_t and the lead zero to lower frequencies while keeping their ratio approximately constant, thus minimizing the effects of higher-frequency poles and zeros which, hopefully, do not also move to lower frequencies. In most cases, the higher-frequency poles and zeros (except for the lead zero) will not move to significantly lower frequencies when C_C is increased.

6. The final step is to replace R_C by a transistor. The size of the transistor can be chosen using equation (5.61), which is repeated here for convenience:

$$R_C = r_{ds16} = \frac{1}{\mu_n C_{ox}\left(\frac{W}{L}\right)_{16} V_{eff16}} \tag{5.80}$$

Finally, SPICE can be used again to fine-tune the device dimensions to optimize the phase margin to that obtained in steps 4 and 5.

Not only does the procedure just described apply to the two-stage opamp, but it (or a very similar procedure) has been found to be almost optimum for compensating most types of opamps.

EXAMPLE 5.7

An opamp has an open-loop transfer function given by

$$A(s) = \frac{A_0(1 + s/\omega_z)}{(1 + s/\omega_{p1})(1 + s/\omega_2)} \tag{5.81}$$

12. The first pole contributes –90° phase shift, the lead zero contributes approximately 45° phase shift, and the equivalent second pole contiguities approximately –40° phase shift.

Here, A_0 is the dc gain of the opamp and ω_z, ω_1, and ω_2 are the frequencies of a zero, the dominant pole, and the equivalent second pole, respectively. Assume that $\omega_2 = 2\pi \times 50$ MHz and that $A_0 = 10^4$.

(*a*) Assuming $\omega_z \to \infty$, find ω_{p1} and the unity-gain frequency, ω_t', so that the opamp has a unity-gain phase margin of 55°.

(*b*) Assuming $\omega_z = 1.2\omega_t'$ (where ω_t' is as found in part (*a*)), what is the new unity-gain frequency, ω_t? Also, find the new phase margin.

Solution

First, note that at frequencies much greater than the dominant-pole frequency (i.e., $\omega >> \omega_{p1}$), such as the unity-gain frequency, (5.81) can be approximated by

$$A(s) \cong \frac{A_0(1 + s/\omega_z)}{(s/\omega_{p1})(1 + s/\omega_2)} \tag{5.82}$$

(*a*) For $\omega_z \to \infty$ we use (5.82) to find the phase angle at ω_t':

$$\angle A(j\omega_t') = -90° - \tan^{-1}(\omega_t'/\omega_2) \tag{5.83}$$

For a phase margin of $PM = 55°$, we need

$$\angle A(j\omega_t') = -180° + PM = -125° \tag{5.84}$$

Setting (5.83) equal to (5.84) and solving for ω_t'/ω_2 results in

$$\begin{aligned} \tan^{-1}(\omega_t'/\omega_2) &= 35° \\ \Rightarrow \omega_t' &= 2.2 \times 10^8 = 2\pi \times 35 \text{ MHz} \end{aligned} \tag{5.85}$$

Next, setting $|A(j\omega_t')| = 1$ and again using (5.82) results in

$$\begin{aligned} \frac{A_0}{(\omega_t'/\omega_{p1})\sqrt{1 + (\omega_t'/\omega_2)^2}} &= 1 \\ \Rightarrow \omega_{p1} = \frac{\omega_t'\sqrt{1 + (\omega_t'/\omega_2)^2}}{A_0} &= 2\pi \times 4.28 \text{ kHz} \end{aligned} \tag{5.86}$$

(*b*) First, we set

$$\omega_z = 1.2\omega_t' = 2\pi \times 42 \text{ MHz} \tag{5.87}$$

To find the new unity-gain frequency, setting $|A(j\omega_t)| = 1$ in (5.82) now gives

$$\begin{aligned} \frac{A_0\sqrt{1 + (\omega_t/\omega_z)^2}}{(\omega_t/\omega_{p1})\sqrt{1 + (\omega_t/\omega_2)^2}} &= 1 \\ \Rightarrow \omega_t = \frac{A_0\omega_{p1}\sqrt{1 + (\omega_t/\omega_z)^2}}{\sqrt{1 + (\omega_t/\omega_2)^2}} \end{aligned} \tag{5.88}$$

This equation can be solved for the new unity-gain frequency. Both sides can be squared, resulting in a quadratic equation in ω_t, which can be solved exactly. Alternatively, the value for ω_t' found in part (*a*) can be used as an initial guess,

and (5.88) can be solved iteratively. After four iterations, one finds that $\omega_t = 2\pi \times 46.8$ MHz. This unity-gain frequency is a 34 percent increase over that found in part (*a*). We can solve for the new phase shift by using

$$\angle A(j\omega_t) = -90° + \tan^{-1}(\omega_t/\omega_z) - \tan^{-1}(\omega_t/\omega_2) = -85.0° \tag{5.89}$$

Note that this phase value gives a phase margin of 95°, which is a 40° improvement! Normally, the improvement would not be this great due to additional high-frequency poles and zeros (which have been ignored here). These high-frequency poles and zeros will typically degrade the phase margin by an additional 5° or 10°. Regardless of this degradation, the improvement from using lead compensation is substantial.

Making Compensation Independent of Process and Temperature

This section shows how lead compensation can be made process and temperature insensitive. Repeating equations (5.14) and (5.71) here, we have

$$\omega_t = \frac{g_{m1}}{C_C} \tag{5.90}$$

and

$$\omega_{p2} \cong \frac{g_{m7}}{C_1 + C_2} \tag{5.91}$$

We see here that the second pole is proportional to the transconductance of the drive transistor of the second stage, g_{m7}. Also, the unity-gain frequency is proportional to the transconductance of the input transistor of the first stage, g_{m1}. Furthermore, the ratios of all of the transconductances remain relatively constant over process and temperature variations since the transconductances are all determined by the same biasing network.[13] Also, most of the capacitances also track each other since they are primarily determined by gate oxides. Repeating (5.73), when a resistor is used to realize lead compensation, the lead zero is at a frequency given by

$$\omega_z = \frac{-1}{C_C(1/g_{m7} - R_C)} \tag{5.92}$$

Thus, if R_C can also be made to track the inverse of transconductances, and in particular $1/g_{m7}$, then the lead zero will also be proportional to the transconductance of Q_7. As a result, the lead zero will remain at the same relative frequency with respect to ω_t and ω_{p2}, as well as all other high-frequency poles and zeros. In other words, the lead compensation will be mostly independent of process and temperature variations.

13. The ratio μ_n/μ_p is relatively constant for a given process (say, a 0.6-μm process) but often varies more significantly from process to process (say, from 0.6-μm to 0.4-μm processes).

It turns out that R_C can be made proportional to $1/g_{m7}$ as long as R_C is realized by a transistor in the triode region that has an effective gate-source voltage proportional to that of Q_7. To see this result, recall that R_C is actually realized by Q_{16}, and therefore we have

$$R_C = r_{ds16} = \frac{1}{\mu_n C_{ox}(W/L)_{16} V_{eff16}} \tag{5.93}$$

Also, g_{m7} is given by

$$g_{m7} = \mu_n C_{ox}(W/L)_7 V_{eff7} \tag{5.94}$$

Thus, the product $R_C g_{m7}$, which we want to be a constant, is given by

$$R_C g_{m7} = \frac{(W/L)_7 V_{eff7}}{(W/L)_{16} V_{eff16}} \tag{5.95}$$

Therefore, all that remains is to ensure that V_{eff16}/V_{eff7} is independent of process and temperature variations since clearly the remaining terms depend only on a geometric relationship. The ratio V_{eff16}/V_{eff7} can be made constant by deriving V_{GS16} from the same biasing circuit used to derive V_{GS7}. Specifically, consider the circuit shown in Fig. 5.9, which consists of a bias stage, the second stage, and the compensation network of the two-stage opamp. First we must make $V_a = V_b$, which is possible by making $V_{eff13} = V_{eff7}$. These two effective gate-source voltages can be made equal by taking

$$\sqrt{\frac{2I_{D7}}{\mu_n C_{ox}(W/L)_7}} = \sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} \tag{5.96}$$

Squaring and simplifying, we see that the following condition must be satisfied:

$$\frac{I_{D7}}{I_{D13}} = \frac{(W/L)_7}{(W/L)_{13}} \tag{5.97}$$

However, the ratio I_{D7}/I_{D13} is set from the current mirror pair Q_6, Q_{11}, resulting in

$$\frac{I_{D7}}{I_{D13}} = \frac{(W/L)_6}{(W/L)_{11}} \tag{5.98}$$

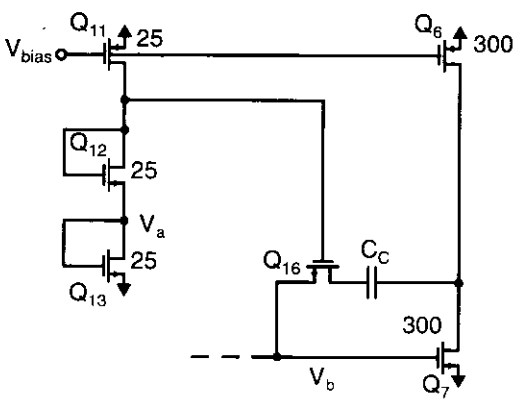


Fig. 5.9 The bias circuit, second stage, and compensation circuit of the two-stage opamp.

Thus, to make $V_{eff13} = V_{eff7}$, from (5.97) and (5.98), we need to satisfy the following relationship:

$$\frac{(W/L)_6}{(W/L)_7} = \frac{(W/L)_{11}}{(W/L)_{13}} \tag{5.99}$$

Assuming this condition is satisfied, we have $V_{eff13} = V_{eff7}$, and therefore we also have $V_a = V_b$. Since the gates of Q_{12} and Q_{16} are connected and their source voltages are the same, we also have $V_{eff12} = V_{eff16}$. Next, using these gate-source relationships and noting that $I_{D12} = I_{D13}$, we can write

$$\frac{V_{eff7}}{V_{eff16}} = \frac{V_{eff13}}{V_{eff12}} = \frac{\sqrt{\dfrac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}}}{\sqrt{\dfrac{2I_{D12}}{\mu_n C_{ox}(W/L)_{12}}}} = \sqrt{\frac{(W/L)_{12}}{(W/L)_{13}}} \tag{5.100}$$

Finally, substituting (5.100) into (5.95), we have the product $R_C g_{m7}$, given by

$$R_C g_{m7} = \frac{(W/L)_7}{(W/L)_{16}}\sqrt{\frac{(W/L)_{12}}{(W/L)_{13}}} \tag{5.101}$$

which is only dependent on geometry and not on processing or temperature variations.

As a result, we have guaranteed that the drain-source resistance of a transistor in the triode region is inversely matched to the transconductance of a different transistor. This relationship can be very useful for many other applications as well. Indeed, in the next subsection, we will see that it's quite simple to make all of the transconductances of transistors in a microcircuit match the conductance of a single off-chip resistor. *This approach results in the possibility of on-chip "resistors," realized by using triode-region transistors that are accurately ratioed with respect to a single off-chip resistor.* This relationship can be very useful in modern circuit design.

Biasing an Opamp to Have Stable Transconductances

We have seen that transistor transconductances are perhaps the most important parameters in opamps that must be stabilized. This stabilization can be achieved by using a circuit approach first proposed in [Steininger, 1990], in which transistor transconductances are matched to the conductance of a resistor. As a result, to a first-order effect, the transistor transconductances are independent of power-supply voltage as well as process and temperature variations.

The bias circuit is shown in Fig. 5.10. First, it is assumed that $(W/L)_{10} = (W/L)_{11}$. This equality results in both sides of the circuit having the same current due to the current-mirror pair Q_{10}, Q_{11}. As a result, we also must have $I_{D15} = I_{D13}$. Now, around the loop consisting of Q_{13}, Q_{15}, and R_B, we have

$$V_{GS13} = V_{GS15} + I_{D15}R_B \tag{5.102}$$

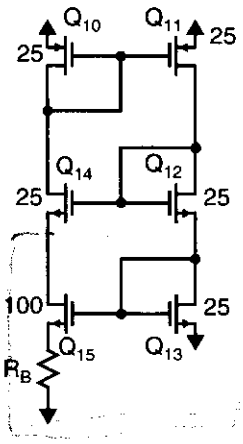


Fig. 5.10 A bias circuit that gives very predictable and stable transistor transconductances.

and recalling that $V_{effi} = V_{GSi} - V_t$, we can subtract the threshold voltage, V_t, from both sides, resulting in

$$V_{eff13} = V_{eff15} + I_{D15}R_B \tag{5.103}$$

This equation can also be written as

$$\sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} = \sqrt{\frac{2I_{D15}}{\mu_n C_{ox}(W/L)_{15}}} + I_{D15}R_B \tag{5.104}$$

and since $I_{D13} = I_{D15}$, we can also write

$$\sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{13}}} = \sqrt{\frac{2I_{D13}}{\mu_n C_{ox}(W/L)_{15}}} + I_{D13}R_B \tag{5.105}$$

Rearranging, we obtain

$$\frac{2}{\sqrt{2\mu_n C_{ox}(W/L)_{13}I_{D13}}}\left[1 - \sqrt{\frac{W/L_{13}}{W/L_{15}}}\right] = R_B \tag{5.106}$$

and recalling that $g_{m13} = \sqrt{2\mu_n C_{ox}(W/L)_{13}I_{D13}}$ results in the important relationship

$$g_{m13} = \frac{2\left[1 - \sqrt{\frac{(W/L)_{13}}{(W/L)_{15}}}\right]}{R_B} \tag{5.107}$$

Thus, the transconductance of Q_{13} is determined by geometric ratios only, independent of power-supply voltages, process parameters, temperature, or any other parameters with large variability. For the special case of $(W/L)_{15} = 4(W/L)_{13}$, we have simply

$$g_{m13} = \frac{1}{R_B} \tag{5.108}$$

Note that, not only is g_{m13} stabilized, but all other transconductances are also stabilized since all transistor currents are derived from the same biasing network, and, therefore, the ratios of the currents are mainly dependent on geometry. We thus have, for all n-channel transistors,

$$g_{mi} = \sqrt{\frac{(W/L)_i I_{Di}}{(W/L)_{13} I_{D13}}} \times g_{m13} \tag{5.109}$$

and for all p-channel transistors

$$g_{mi} = \sqrt{\frac{\mu_p}{\mu_n}\frac{(W/L)_i I_{Di}}{(W/L)_{13} I_{D13}}} \times g_{m13} \tag{5.110}$$

It should be noted here that the preceding analysis has ignored many second-order effects such as the transistor output impedance and the body effect. The body effect will modify the equation slightly, but the relationship will still depend primarily on geometry alone. The major limitation is due to the transistor output impedance. This effect can be made of little consequence by replacing the simple current mirrors with cascode mirrors. Although the use of cascode mirrors appears at first glance to require large power supplies, in Chapter 6 we will see that by using wide-swing current mirrors, this is not necessarily the case. A final limitation is that at high temperatures, the currents and effective gate-source voltages increase substantially to offset the mobility degradation and to keep the transconductances stable. In addition, the effective gate-source voltages also increase, thereby limiting signal swings more than is the case at normal temperatures. Since the carrier mobility is proportional to $T^{-3/2}$, this corresponds to a 27-percent reduction from room temperature of (300 °K) to 100 °C (373 °K). Thus, the effective gate-source voltages increase by 27 percent to keep the transistor transconductance unchanged, since

$$g_{mi} = \mu_i C_{ox}\left(\frac{W}{L}\right)_i V_{eff\text{-}i} \tag{5.111}$$

As long as the effective gate-source voltages have not initially been designed to be too large, this limitation is tolerable in most applications.[14] A typical value for effective gate-source voltages might be 0.2 V to 0.25 V at room temperature.

The bias circuit just presented is an example of a circuit having positive feedback. It is stable as long as the loop gain is less than unity, which is the case when $(W/L)_{15}/(W/L)_{13} > 1$. However, this circuit unfortunately can have a second stable state where all the currents are zero. To guarantee this condition doesn't happen, it is necessary to add a *start-up circuit* that only affects the operation if all the currents are zero at start up. Possible start-up circuits will be described in Chapter 6.

14. Also, if on-chip resistors, such as well or diffusion resistors, are used, this effect will be somewhat offset by their positive temperature-coefficient dependency.

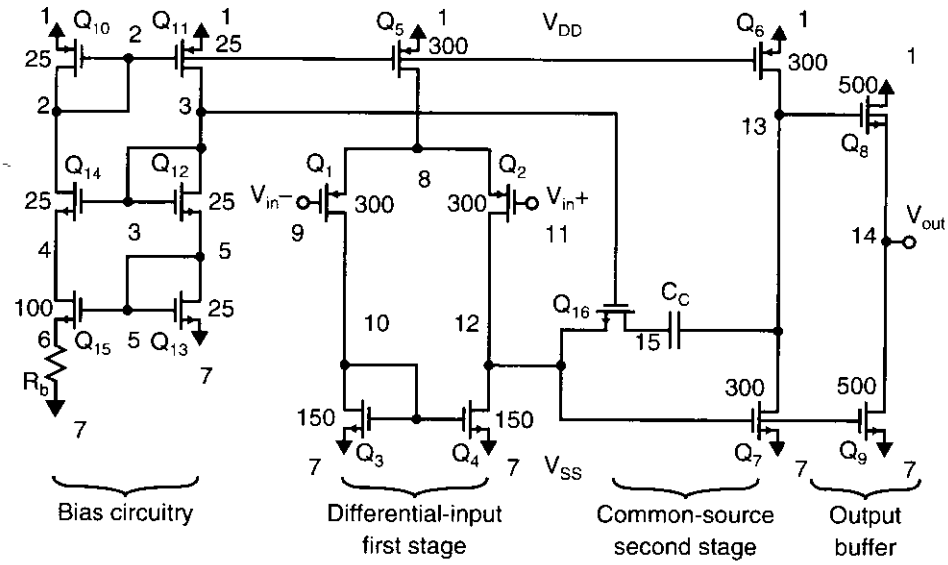


Fig. 5.11 All transistor lengths equal 1.6 μm.

5.3 SPICE SIMULATION EXAMPLES

In this section, SPICE simulation results are presented for Examples 5.1 and 5.2.

Simulations of Examples 5.1 and 5.2.

The circuit for Examples 5.1 and 5.2 is shown in Fig. 5.11. R_b is chosen such that Q_5 has a dc current of 100 μA. The size of Q_{16} is selected for proper lead compensation.

NETLIST:

vdd	1	0			dc 2.5
vss	7	0			dc –2.5
m10	2	2	1	1	PMOS w=25um l=1.6um
m11	3	2	1	1	PMOS w=25um l=1.6um
m14	2	3	4	7	NMOS w=25um l=1.6um
m12	3	3	5	7	NMOS w=25um l=1.6um
m15	4	5	6	7	NMOS w=100um l=1.6um
m13	5	5	7	7	NMOS w=25um l=1.6um
rb	6	7			8k
m5	8	2	1	1	PMOS w=300um l=1.6um
m1	10	9	8	1	PMOS w=300um l=1.6um
m2	12	11	8	1	PMOS w=300um l=1.6um
m3	10	10	7	7	NMOS w=150um l=1.6um
m4	12	10	7	7	NMOS w=150um l=1.6um
vin-	9	0			dc 0

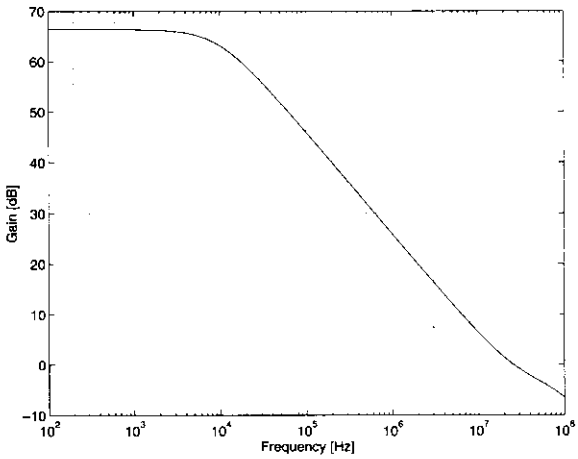


Fig. 5.12 A frequency plot of the two-stage opamp.

```
vin+    11   0                      dc 0 ac 1
m6      13   2    1    1            PMOS w=300um l=1.6um
m8      1    13   14   7            NMOS w=500um l=1.6um
m7      13   12   7    7            NMOS w=300um l=1.6um
m9      14   12   7    7            NMOS w=500um l=1.6um
cc      15   13                     5pF
m16     15   3    12   7            NMOS w=100um l=1.6um

.op
.ac dec 20 0.1k 100Meg
.print vdb(14)
.option post
.options brief
.lib "../cmos_models" library
.end
```

The frequency plot result is shown in Fig. 5.12, where we see that the dc gain is 67 dB (i.e., 2240 V/V). The gain of the opamp is different from the hand-calculation result from Example 5.1 because the values for the transistor output impedances in the example are rough estimations. The unity-gain frequency is shown to be 25 MHz, which corresponds closely to the hand-calculated result since this calculation does not depend on poorly modelled variables.

5.4 REFERENCES

P. R. Gray and R. G. Meyer. *Analysis and Design of Analog Integrated Circuits*, 3rd ed. John Wiley & Sons, New York, 1993.

J. K. Roberge. *Operational Amplifiers*. John Wiley & Sons, New York, 1975.

A. S. Sedra and K. C. Smith. *Microelectronic Circuits*, 3rd ed. Holt, Rinehart & Winston, New York, 1991.

J. M. Steininger, "Understanding Wide-Band MOS Transistors," *IEEE Circuits and Devices*, Vol. 6, No. 3, pp. 26–31, May 1990.

5.5 PROBLEMS

Unless otherwise stated, assume the following:

- n-channel MOS transistors:

  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8\ V$
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}(\Omega) = 8{,}000\ L(\mu m)/I_D(mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:

  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9\ V$
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}(\Omega) = 12{,}000\ L(\mu m)/I_D(mA)$ in active region,
  $C_j = 4.5 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.5 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

5.1 Assume that $I_{D5} = 100\ \mu A$, all transistor lengths are $1.2\ \mu m$, and $C_c = 10\ pF$. For the two-stage opamp of Fig. 5.2,

a. Estimate the –3-dB frequency of the first stage.

b. Estimate the unity-gain frequency of the opamp.

c. Find the slew rate.

5.2 Repeat Example 5.3, but let $C_C = 4$ pF. What circuit changes could be made to double the slew rate while keeping ω_t and C_C unchanged?

5.3 How should the sizes of Q_6 and Q_7 change in Problem 5.2 to maintain zero systematic offset voltage?

5.4 Assume Q_3 and Q_4 of the two-stage opamp of Fig. 5.4 are both 50 μm wide, all transistor lengths are $1.6\ \mu m$, and $I_{D5} = 100\ \mu A$. Estimate the inherent input-offset voltage.

5.5 Ignoring the body effect, what is the output voltage range of the opamp of Fig. 5.2? What is the range of the common-mode input voltage, assuming ±5 V power supplies are used? Assume $I_{D5} = 100\ \mu A$.

5.6 Consider the opamp shown in Fig. 5.2, where Q_{16} is placed on the output side of C_C (rather than where it is shown). Explain why this circuit would then oscillate for large positive output values.

5.7 Repeat Problem 5.5, but take the body effect into account and find approximate answers.

5.8 Consider a system response, $H(s)$, with one dominant pole, other real-axis poles, and zeros such that

$$H(s) = \frac{A_0 \times \Pi\left(1 + \frac{s}{\omega_{zi}}\right)}{\left(1 + \frac{s}{\omega_{p1}}\right)\Pi\left(1 + \frac{s}{\omega_{pi}}\right)}$$

Also consider another approximating function, $H_{app}(s)$, given by

$$H_{app}(s) = \frac{A_o}{\left(1 + \frac{s}{\omega_{p1}}\right)\left(1 + \frac{s}{\omega_{eq}}\right)}$$

Show that, when $H(s)$ is approximated by $H_{app}(s)$ such that the phase of the two systems are approximately equal at ω_t of $H(s)$, ω_{eq} should be equal to

$$\frac{1}{\omega_{eq}} = \sum \frac{1}{\omega_{pi}} - \sum \frac{1}{\omega_{zi}}$$

5.9 An opamp has its first pole at 3 kHz and has high-frequency poles at 130 MHz, 160 MHz, and 180 MHz. Using iteration, find the frequency where the phase shift is $-135°$ and, therefore, find the equivalent time constant that models the high-frequency poles. How does this compare to the estimate given by equation (5.45)?

5.10 A two-stage opamp has a compensation capacitor connected between the input and the output of the second stage. It has an equivalent second-pole frequency of 60 MHz. Assume the input transistors of the first stage have a transconductance of 0.775 mA/V and the gain of the output buffer is exactly unity. What is the required size of the compensation capacitor if the phase margin is to be 55° for the feedback configuration shown in Fig. P5.10?

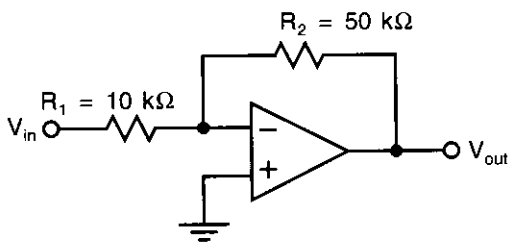


Fig. P5.10

5.11 For the result of Problem 5.10, what is a near-optimum size for a capacitor to be placed in parallel with R_2 in order to obtain lead compensation?

5.12 An opamp has an open-loop transfer function given by

$$H(s) = \frac{A_0}{\left(1 + \frac{s}{\omega_1}\right)\left(1 + \frac{s}{\omega_2}\right)}$$

Assume $A_0 = 10^4$ and $\omega_2 = 10^8$ rad/s. A zero, ω_z, is added such that the frequency of the zero is the same as the resulting open-loop unity-gain frequency, ω_t. Find ω_1 and ω_t so that the phase margin is 80°.

5.13 An opamp has an open-loop transfer function given by

$$A(s) \cong \frac{A_0(1 + s\tau_z)}{s\tau_1(1 + s\tau_2)}$$

Find the transfer function of the closed-loop amplifier, assuming a feedback factor β exists. Find approximate equations for the resonant frequency and the Q factor of the denominator of the transfer function of the closed-loop amplifier.

5.14 Prove that the frequencies of the first and second poles of the amplifier of Fig. 5.2 are still given by (5.70) and (5.71), even when lead compensation is used (i.e., when R_C is included in Fig. 5.8).

5.15 Design the bias circuit of Fig. 5.10 to have $V_{eff} = 0.25$ V for all transistors. For $Q_{12} - Q_{14}$ having sizes of 10 μm/1.2 μm and Q_{15} having a size of 40 μm/1.2 μm, what is the required size of R? What would the effective gate-source voltage be at 70 °C? Confirm your answer using SPICE.

CHAPTER

6 Advanced Current Mirrors and Opamps

The classical two-stage opamp was discussed in Chapter 5. While this two-stage opamp has been used in many commercial integrated circuits, recently a number of alternate opamp designs have been gaining in popularity and are the topic of this chapter. Since many of these opamps make use of more advanced current mirrors, a discussion of advanced current mirrors is presented first. Following the current-mirror section, the design of two modern opamps having single-ended outputs is presented. Next, design principles are discussed for realizing opamps having fully differential outputs. Finally, the basics of "current-feedback opamps" are presented, where we see that these opamps can have large gain-bandwidth products without the need for adjusting the compensation capacitance.

6.1 ADVANCED CURRENT MIRRORS

Wide-Swing Current Mirrors

As newer technologies with shorter channel lengths are used, it becomes more difficult to achieve reasonable opamp gains due to transistor output-impedance degradation caused by short-channel effects. As a result, designers are often forced to use cascode current mirrors. Unfortunately, the use of conventional cascode current mirrors limits the signal swings available, which may not be tolerated in certain applications. Fortunately, circuits exist that do not limit the signal swings as much as the current mirrors discussed in Chapter 3. One such circuit is shown in Fig. 6.1 and is often called the "wide-swing cascode current mirror" [Sooch, 1985; Babanezhad, 1987].

The basic idea of this current mirror is to bias the drain-source voltages of transistors Q_2 and Q_3 to be close to the minimum possible without them going into the triode region. Specifically, if the sizes shown in Fig. 6.1 are used, and assuming the classical square-law equations for long channel-length devices are valid, transistors Q_2 and Q_3 will be biased right at the edge of the triode region. Before seeing how these bias voltages are created, note that the transistor pair Q_3, Q_4 acts like a single diode-connected transistor in creating the gate-source voltage for Q_3. These two transistors operate very similarly to how Q_3 alone would operate if its gate were con-

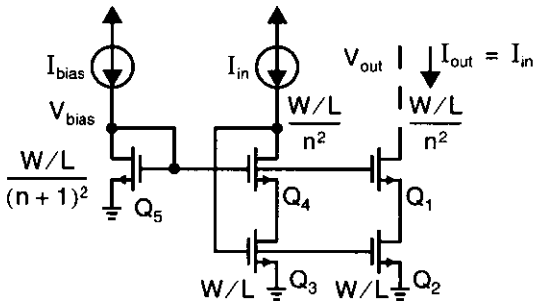


Fig. 6.1 The wide-swing cascode current mirror. I_{bias} typically is set to the nominal or maximum input current, I_{in}.

nected to its source. The reason for including Q_4 is to lower the drain-source voltage of Q_3 so that it is matched to the drain-source voltage of Q_2. This matching makes the output current, I_{out}, more accurately match the input current, I_{in}. If Q_4 were not included, then the output current would be a little smaller than the input current due to the finite output impedances of Q_2 and Q_3. Other than this, Q_4 has little effect on the circuit's operation.

To determine the bias voltages for this circuit, let V_{eff} be the effective gate-source voltage of Q_2 and Q_3, and assume all of the drain currents are equal. We therefore have

$$V_{eff} = V_{eff2} = V_{eff3} = \sqrt{\frac{2I_{D2}}{\mu_n C_{ox}(W/L)}} \tag{6.1}$$

Furthermore, since Q_5 has the same drain current but is $(n+1)^2$ times smaller, we have

$$V_{eff5} = (n+1)V_{eff} \tag{6.2}$$

Similar reasoning results in the effective gate-source voltages of Q_1 and Q_4 being given by

$$V_{eff1} = V_{eff4} = nV_{eff} \tag{6.3}$$

Thus,

$$V_{G5} = V_{G4} = V_{G1} = (n+1)V_{eff} + V_{tn} \tag{6.4}$$

Furthermore,

$$V_{DS2} = V_{DS3} = V_{G5} - V_{GS1} = V_{G5} - (nV_{eff} + V_{tn}) = V_{eff} \tag{6.5}$$

This drain-source voltage puts both Q_2 and Q_3 right at the edge of the triode region. Thus, the minimum allowable output voltage is now

$$V_{out} > V_{eff1} + V_{eff2} = (n+1)V_{eff} \tag{6.6}$$

A common choice for n might be simply unity, in which case the current mirror operates correctly as long as

$$V_{out} > 2V_{eff} \tag{6.7}$$

With a typical value of V_{eff} between 0.2 V and 0.25 V, the wide-swing current mirror can guarantee that all of the transistors are in the active (i.e., saturation) region even when the voltage drop across the mirror is as small as 0.4 V to 0.5 V.

However, there is one other requirement that must be met to ensure that all transistors are in the active region. Specifically, we need

$$V_{DS4} > V_{eff4} = nV_{eff} \tag{6.8}$$

to guarantee that Q_4 is in the active region. To find V_{DS4}, we note that the gate of Q_3 is connected to the drain of Q_4, resulting in

$$V_{DS4} = V_{G3} - V_{DS3} = (V_{eff} + V_{tn}) - V_{eff} = V_{tn} \tag{6.9}$$

As a result, one need only ensure that V_{tn} be greater than nV_{eff} for Q_4 to remain in the active region—not a difficult requirement.

It should be noted that this circuit was analyzed assuming the bias current, I_{bias}, equals the input current, I_{in}. Since, in general, I_{in} may be a varying current level, there is some choice as to what I_{bias} value should be chosen. One choice is to set I_{bias} to the largest expected value for I_{in}. Such a choice will ensure that none of the devices exit their active region, though the drain-source voltage of Q_2 and Q_3 will be larger than necessary except when the maximum I_{in} is applied. As a result, some voltage swing will be lost. Perhaps the more common choice in a wide-swing opamp is to let I_{bias} equal the nominal value of I_{in}. With this setting, some devices will enter triode and the output impedance will be reduced for larger I_{in} values (say, during slew-rate limiting), but such an effect during transient conditions can often be tolerated.

Before leaving this circuit, a few design comments are worth mentioning. In most applications, an experienced designer would take $(W/L)_5$ smaller than the size given in Fig. 6.1 to bias transistors Q_2 and Q_3 with slightly larger drain-source voltages than the minimum required (perhaps 0.1 V to 0.15 V larger). This increase accounts for the fact that practical transistors don't have a sharp transition between the triode and active regions. This increase would also help offset a second-order effect due to the body effect of Q_1 and Q_4, which causes their threshold voltages to increase, which tends to push Q_2 and Q_3 more into the triode region. In addition, to save on power dissipation, the bias branch consisting of Q_5 and I_{bias} might be scaled to have lower currents while keeping the same current densities and, therefore, the same effective gate-source voltages. A final common modification is to choose the lengths of Q_2 and Q_3 just a little larger than the minimum allowable gate length (as the drain-source voltage across them is quite small), but Q_1 and Q_4 would be chosen to have longer gate lengths since the output transistor (i.e., Q_1) often has larger voltages across it. A typical size might be twice the minimum allowable channel length. This choice of gate lengths helps eliminate detrimental short-channel effects, such as the drain-to-substrate leakage currents of Q_1 that might otherwise result. Minimizing the lengths of Q_2 and Q_3 maximizes the frequency response, as their gate-source capacitances are the most significant capacitances contributing to high-frequency poles.

Finally, it should be mentioned that this current mirror is presently becoming the most popular CMOS current mirror and is therefore very important in many analog

designs besides opamps. The next subsection will show how it can be incorporated into the constant g_m bias circuit described in Chapter 5.

Wide-Swing Constant-Transconductance Bias Circuit

It is possible to incorporate wide-swing current mirrors into the constant-transconductance bias circuit described in Chapter 5. This modification greatly minimizes most of the detrimental second-order imperfections caused by the finite-output impedance of the transistors, without greatly restricting signal swings. The complete circuit is shown in Fig. 6.2. This circuit is a modification of the circuit described in Fig 5.10, and has both wide-swing current mirrors and a start-up circuit.

The n-channel wide-swing cascode current mirror consists of transistors Q_1–Q_4, along with the diode-connected biasing transistor Q_5. The pair Q_3, Q_4 acts similarly to a diode-connected transistor at the input side of the mirror. The output current comes from Q_1. The gate voltages of cascode transistors Q_1 and Q_4 are derived by the diode-connected transistor Q_5. The current for this biasing transistor is actually derived from the bias loop via Q_{10} and Q_{11}.

Similarly, the p-channel wide-swing cascode current mirror is realized by Q_6–Q_9. Transistors Q_8 and Q_9 operate as a diode-connected transistor at the input side of the mirror. The current-mirror output current is the drain current of Q_6. The cascode transistors Q_6 and Q_9 have gate voltages derived from diode-connected Q_{14}, which has a bias current derived from the bias loop via Q_{12} and Q_{13}.

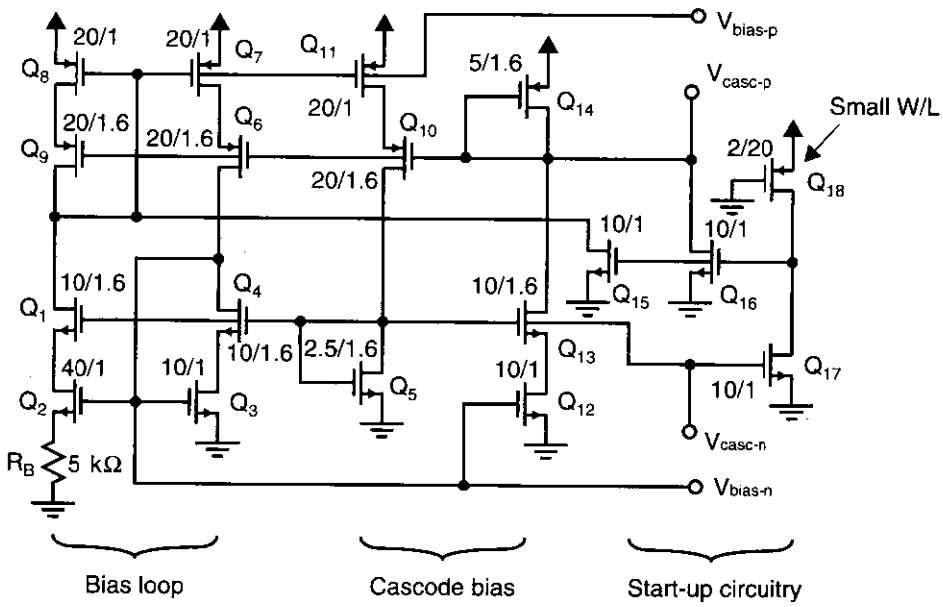


Fig. 6.2 A constant-transconductance bias circuit having wide-swing cascode current mirrors.

The bias loop does have the problem that at start-up it is possible for the current to be zero in all transistors, and the circuit will remain in this stable state forever. To ensure this condition does not happen, it is necessary to include start-up circuitry that affects only the bias-loop in the case that all currents in the loop are zero. An example of a start-up circuit consisting of transistors Q_{15}, Q_{16}, Q_{17}, and Q_{18} is also shown in Fig. 6.2. In the event that all currents in the bias loop are zero, Q_{17} will be off. Since Q_{18} operates as a high-impedance load that is always on, the gates of Q_{15} and Q_{16} will be pulled high. These transistors then will inject currents into the bias loop, which will start up the circuit. Once the loop starts up, Q_{17} will come on, sinking all of the current from Q_{18}, pulling the gates of Q_{15} and Q_{16} low, and thereby turning them off so they no longer affect the bias loop. This circuit is only one example of a start-up loop, and there are many other variations. For example, sometimes the p-channel transistor, Q_{18}, is replaced by an actual resistor (perhaps realized using a well resistor).

It is of interest to note that the bias circuit shown consists of four different loops—the main loop with positive feedback, the start-up loop that eventually gets disabled, and the two loops used for establishing the bias voltages for the cascode transistors. These latter two loops also constitute positive feedback but with very little gain.

Shown next to each transistor are reasonable W/L dimensions (in μm) of a possible realization that had its operation experimentally verified (it was realized in a 0.8-μm technology). Because this circuit allowed for accurately predictable transconductances, it also allows the performance of the realized opamp to be accurately predicted using moderately simple equations prior to realization.

Enhanced Output-Impedance Current Mirrors

Another variation on the cascode current mirror is often referred to as the *enhanced output-impedance current mirror*. A simplified form of this circuit is shown in Fig. 6.3, and it is used to increase the output impedance. The basic idea is to use a feedback amplifier to keep the drain-source voltage across Q_2 as stable as possible, irrespective of the output voltage. The addition of this amplifier ideally increases the

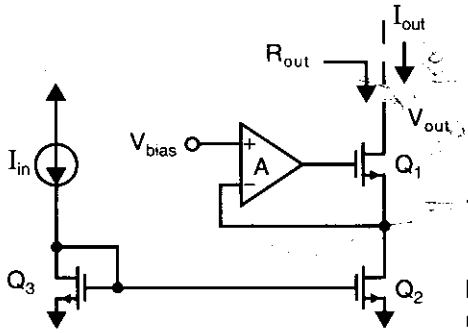


Fig. 6.3 The enhanced output-impedance current mirror.

output impedance by a factor equal to one plus the loop gain over that which would occur for a classical cascode current mirror. Specifically, using the results from Chapter 3,

$$R_{out} \cong g_{m1} r_{ds1} r_{ds2} (1 + A) \tag{6.10}$$

In practice, the output impedance might be limited by a parasitic conductance between the drain of Q_1 and its substrate due to short-channel effects. This parasitic conductance is a result of collisions between highly energized electrons resulting in electron-hole pairs to be generated with the holes escaping to the substrate. The generation of these electron-hole pairs is commonly called *impact ionization*.

It should be mentioned that the technique just described for output-impedance enhancement is not useful when bipolar transistors are used due to the finite base current of the transistors (see Problem 6.8).

EXAMPLE 6.1

Assume an amplifier, $A(s)$, has a transfer function given by

$$A(s) = \frac{A_0}{1 + s\tau'_1} \cong \frac{A_0}{s\tau'_1} \tag{6.11}$$

around its unity-gain frequency. Find approximate equations for any additional poles and/or zeros introduced by the gain enhancement circuit that are significant around the unity-gain frequency of the common-source amplifier shown in Fig. 6.4. Assume the gain of the common-source stage is given by

$$A_V(s) = \frac{V_{out}(s)}{V_{in}(s)} = -g_{m2}\left(R_{out}(s) \parallel \frac{1}{sC_L}\right) \tag{6.12}$$

where R_{out} is now a function of frequency and given by

$$R_{out}(s) = g_{m1} r_{ds1} r_{ds2} (1 + A(s)) \tag{6.13}$$

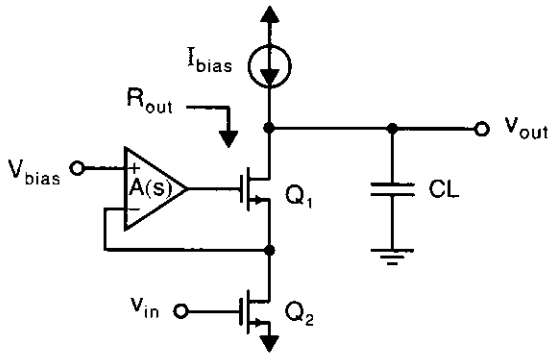


Fig. 6.4 A common-source amplifier with output-impedance enhancement.

Solution

Note that at high frequencies, it is not possible to assume that $|A(j\omega)| >> 1$. Working with conductances, we have

$$G_{out}(s) = \frac{1}{R_{out}(s)} = \frac{g_{ds1}g_{ds2}}{g_{m1}(1 + A(s))} \tag{6.14}$$

which can be inserted into (6.12) giving

$$\begin{aligned} A_V(s) &= \frac{-g_{m2}}{sC_L + G_{out}(s)} = \frac{-g_{m2}}{sC_L + \dfrac{g_{ds1}g_{ds2}}{g_{m1}(1 + A(s))}} \\ &= \frac{-g_{m2}(1 + A(s))}{sC_L(1 + A(s)) + \dfrac{g_{ds1}g_{ds2}}{g_{m1}}} \end{aligned} \tag{6.15}$$

Substituting (6.11) into (6.15) and rearranging gives

$$A_V(s) = \frac{-g_{m2}(A_0 + s\tau_1')}{s\left(A_0C_L + \dfrac{g_{ds1}g_{ds2}\tau_1'}{g_{m1}} + sC_L\tau_1'\right)} \tag{6.16}$$

From (6.16), we see that there is a left-hand-plane zero at a frequency given by

$$\omega_z = \frac{A_0}{\tau_1'} \tag{6.17}$$

which is the approximate unity-gain frequency of the amplifier, $A(s)$. Also, there are two poles, but the pole at dc is not accurate since we have made use of approximate equations that are valid only near the unity-gain frequency of the amplifier. However, the other pole location is a good approximation and it occurs in the left-half-plane given by (after some rearranging)

$$p_2 = \frac{A_0}{\tau_1'} + \frac{g_{ds1}g_{ds2}}{C_Lg_{m1}} = \frac{A_0}{\tau_1'} + \frac{1}{C_LR'} \tag{6.18}$$

Here, we have defined

$$R' = g_{m1}r_{ds1}r_{ds2} \tag{6.19}$$

which is approximately the output impedance of the cascode mirror without enhancement and is quite a large value. Furthermore, note that the term A_0/τ_1' is a value near the unity-gain frequency of the amplifier, and therefore, typically the first term in (6.18) dominates. Thus, the added pole frequency is approximately given by

$$p_2 \cong \frac{A_0}{\tau_1'} \cong \omega_z' \tag{6.20}$$

Therefore, like the new zero, the new pole also occurs around the unity-gain frequency of the amplifier, and their effects mostly cancel each other out. In addition, if the unity-gain frequency of the enhancement loop is greater than the unity-gain frequency of the opamp, the frequency response of the enhancement amplifier shouldn't have a large detrimental effect on the overall frequency response.

The enhanced output-impedance current mirror appears to have been originally proposed in [Hosticka, 1979] for single-ended input amplifiers and more recently described almost simultaneously in [Bult, 1990] and [Säckinger, 1990] for differential-input amplifiers. Bult proposed that complete opamps be used for each current mirror. In a modern fully differential opamp, this might increase the number of opamps by four, although these extra opamps can be scaled to have less power dissipation. The implementation proposed by Säckinger is shown in Fig. 6.5. The feedback amplifier in this case is realized by the common-source amplifier consisting of Q_3 and its current source I_{B1}. Assuming the output impedance of current source I_{B1} is approximately equal to r_{ds3}, the loop gain will be $(g_{m3}r_{ds3})/2$, and the final ideal output impedance will be given by

$$r_{out} \cong \frac{g_{m1}g_{m3}r_{ds1}r_{ds2}r_{ds3}}{2} \tag{6.21}$$

The circuit consisting of Q_4, Q_5, Q_6, I_{in}, and I_{B2} operates almost identically to a diode-connected transistor, but is used instead to guarantee that all transistor bias voltages are accurately matched to those of the output circuitry consisting of Q_1, Q_2, Q_3, and I_{B1}. As a result, I_{out} will very accurately match I_{in}.

The Säckinger realization is much simpler than that proposed by Bult, but has a major limitation in that the signal swing is significantly reduced. This reduction is a result of Q_2 and Q_5 being biased to have drain-source voltages much larger than the minimum required. Specifically, their drain-source voltages are given by

$$V_{DS2} = V_{DS5} = V_{eff3} + V_{tn} \tag{6.22}$$

rather than the minimum required, which would be equal to V_{eff2}. This limitation is especially harmful for modern technologies operating with power supply voltages of

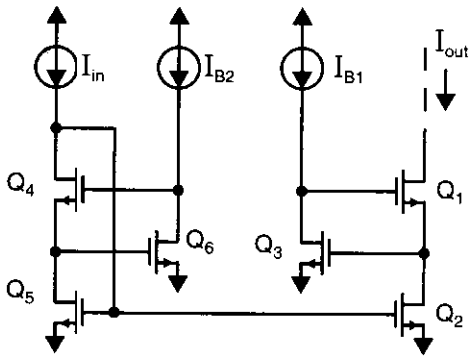


Fig. 6.5 The Säckinger implementation of the enhanced output-impedance current mirror.

3.3 V or lower. In the next subsection, an alternative realization is described that combines the wide-swing current mirror with the enhanced output-impedance circuit.

Wide-Swing Current Mirror with Enhanced Output Impedance

This current mirror was originally described in [Gatti, 1990] for use in current-mode continuous-time filters and was then used in the design of wide-signal-swing opamps [Coban, 1994; Martin, 1994]. It is very similar to the enhanced output-impedance current mirrors of [Säckinger, 1990], except that diode-connected transistors used as level shifters have been added in front of the common-source enhancement amplifiers, as shown in Fig. 6.6. At the output side, the level shifter is the diode-connected transistor, Q_4, biased with current I_{bias}. The circuitry at the input basically acts as a diode-connected transistor while ensuring that all bias voltages are matched to the output circuitry so that I_{out} accurately matches I_{in}. Shown next to each transistor is a reasonable width in μm. Note that for the case in which

$$I_{bias} = I_{in}/7 \tag{6.23}$$

all transistors are biased with nearly the same current density, except for Q_3 and Q_7. As a result, all transistors have the same effective gate-source voltages, V_{eff}, except for Q_3 and Q_7, which have gate-source voltages of $2V_{eff}$ because they are biased at four times the current density. Thus, we find that the gate voltage of Q_3 equals

$$V_{G3} = 2V_{eff} + V_{tn} \tag{6.24}$$

and the drain-source voltage of Q_2 is given by

$$V_{DS2} = V_{S4} = V_{G3} - V_{GS4} = (2V_{eff} + V_{tn}) - (V_{eff} + V_{tn}) = V_{eff} \tag{6.25}$$

Therefore, Q_2 is biased on the edge of the triode region and the minimum output voltage is given by

$$V_{out} > V_{DS2} + V_{eff1} = 2V_{eff} \tag{6.26}$$

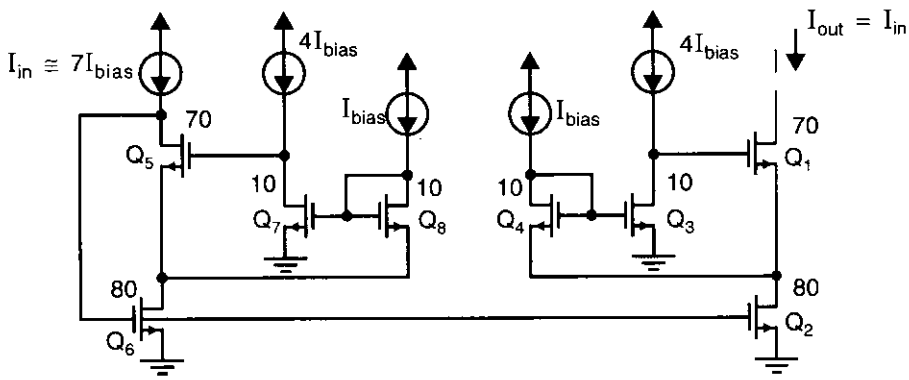


Fig. 6.6 A wide-swing current mirror with enhanced output impedance. Current source $7I_{bias}$ is typically set to the nominal or maximum input current, I_{in}.

With the shown values for the W/L ratios, the power dissipation would be almost doubled over that of a classical cascode current mirror, assuming I_{bias} is set to one-seventh the nominal or maximum input current value, I_{in}. However, it is possible to bias the enhancement circuitry at lower densities and thereby save on power dissipation, albeit at the expense of speed, as the additional poles introduced by the enhancement circuitry would then be at lower frequencies.

A slightly modified version of the wide-swing enhanced output-impedance mirror is shown in Fig. 6.7. This variation obtains the bias voltage of the cascode transistor of the input diode-connected branch from the bias-generation circuitry of Fig. 6.2. This circuit suffers slightly with regard to large-signal dc matching, but has less power dissipation and area than the circuit of Fig. 6.6. Notice also that Q_2 of Fig. 6.6 has now been separated into two transistors, Q_2 and Q_5, in Fig. 6.7. This change allows one to design an opamp without the output-impedance enhancement, but with wide-swing mirrors. Then, when output-impedance enhancement is desired, one simply includes the appropriate enhancement amplifiers and the only change required to the original amplifier is to connect the gates of the output cascode mirrors of the appropriate current mirrors to the enhancement amplifiers rather than the bias-generation circuit. Finally, it has been found through simulation that the modified circuit of Fig. 6.7 is less prone to instability than the circuit of Fig. 6.6.

It is predicted that this current mirror will become more important as the newer technologies force designers to use smaller power-supply voltages, which limit voltage swings. Also, the short-channel effects of newer technologies make the use of current mirrors with enhanced output impedance more desirable.

There are a couple of facts that designers should be aware of when they use output-impedance enhancement. The first fact is that sometimes it may be necessary to add local compensation capacitors to the enhancement loops to prevent ringing during transients. The second fact is that the inclusion of output-impedance enhancement can substantially slow down the settling times for large-signal transients. This is because during large-signal transients, the outputs of the enhancement loops can slew to voltages very different from those required after settling, and it can take a while for the outputs to

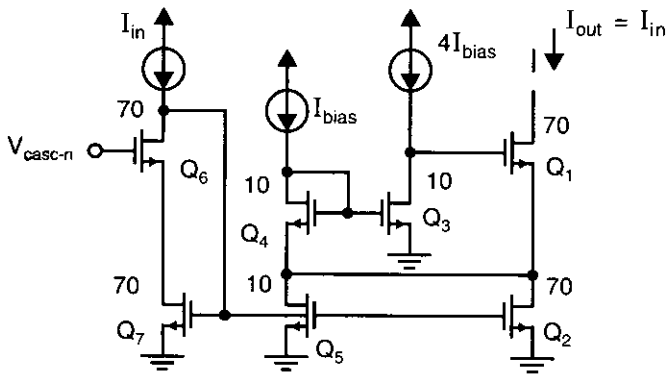


Fig. 6.7 A modified wide-swing current mirror with enhanced output impedance.

Fig. 6.8 (*a*) A symbol representing a current mirror. (*b*) Example of a simple current mirror.

return to the necessary voltages. A typical settling-time difference might be around a 50-percent increase. Whether this is worth the substantial increase in gain (perhaps as much as 30 dB or so) depends on the individual application.

Current-Mirror Symbol

It should now be clear that there are a number of different current-mirror circuits, each having various advantages and disadvantages, that can be used in a particular circuit application. Thus, from an architectural perspective, it is often desirable to describe a circuit without showing which particular current mirror is used. In these cases, we make use of the current-mirror symbol shown in Fig. 6.8(*a*). The arrow is on the input side of the mirror, which is the low-impedance side (with an impedance typically equal to $1/g_m$). The arrow also designates the direction of current flow on the input side. The ratio, $1{:}K$, represents the current gain of the mirror. For example, the current mirror shown in Fig. 6.8(*a*) might be realized by the simple current mirror shown in Fig. 6.8(*b*). Finally, it should be mentioned that for illustrative purposes, some circuits in this book will be shown with a particular current mirror (such as the wide-swing mirror) to realize a circuit architecture. However, it should be kept in mind that similar circuits using almost any of the current mirrors just described (or described elsewhere) are possible.

6.2 FOLDED-CASCODE OPAMP

Many modern integrated CMOS opamps are designed to drive only capacitive loads. With such capacitive-only loads, it is not necessary to use a voltage buffer to obtain a low output impedance for the opamp. As a result, it is possible to realize opamps having higher speeds and larger signal swings than those that must also drive resistive loads. These improvements are obtained by having only a single high-impedance node at the output of an opamp that drives only capacitive loads. The admittance seen at all other nodes in these opamps is on the order of a transistor's transconductance, and thus they have relatively low impedance. By having all internal nodes of relatively low impedance, the speed of the opamp is maximized. It should also be mentioned that these low node impedances result in reduced voltage signals at all nodes other than the output node; however, the current signals in the various transis-

tors can be quite large.[1] With these opamps, we shall see that the compensation is usually achieved by the load capacitance. Thus, as the load capacitance gets larger, the opamp usually becomes more stable but also slower. One of the most important parameters of these modern opamps is their transconductance value (i.e., the ratio of the output current to the input voltage). Therefore, some designers refer to these modern opamps as transconductance opamps, or *Operational Transconductance Amplifiers* (OTAs).

An example of an opamp with a high-output impedance is the folded-cascode opamp, as shown in Fig. 6.9. The design shown is a differential-input single-ended output design. Note that all current mirrors in the circuit are wide-swing cascode current mirrors, discussed previously. The use of these mirrors results in high-output impedance for the mirrors (compared to simple current mirrors), thereby maximizing the dc gain of the opamp. The basic idea of the folded-cascode opamp is to apply cascode transistors to the input differential pair but using transistors opposite in type from those used in the input stage. For example, the differential-pair transistors consisting of Q_1 and Q_2 are n-channel transistors in Fig. 6.9, whereas the cascode transistors consisting of Q_5 and Q_6 are p-channel transistors. This arrangement of opposite-type transistors allows the output of this single gain-stage amplifier to be taken at the same bias-voltage levels as the input signals. It should be mentioned that even though a folded-cascode amplifier is basically a single gain stage, its gain can be quite reasonable, on the order of 700 to 3,000. Such a high gain occurs because the gain is determined by the product

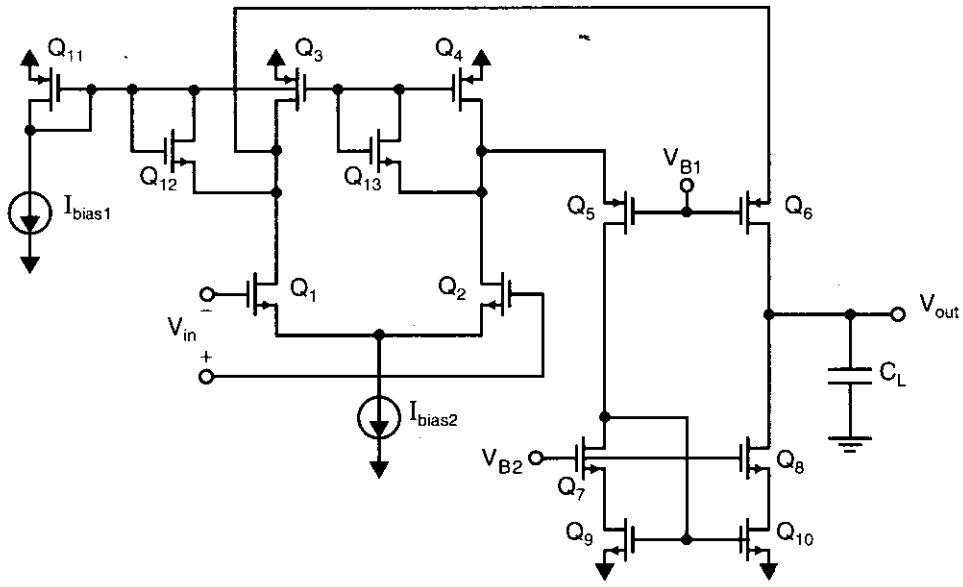


Fig. 6.9 A folded-cascode opamp.

1. Because of their reduced voltage signals but large current signals, these types of opamps are sometimes referred to as *current-mode opamps*.

of the input transconductance and the output impedance, and the output impedance is quite high due to the use of cascode techniques.

The shown differential-to-single-ended conversion is realized by the wide-swing current mirror composed of Q_7, Q_8, Q_9, and Q_{10}. In a differential-output design, these might be replaced by two wide-swing cascode current sinks, and common-mode feedback circuitry would be added, as is discussed in a later section in this chapter.

It should be mentioned that a simplified bias network is shown so as not to complicate matters unnecessarily when the intent is to show the basic architecture. In practical realizations, Q_{11} and I_{bias1} might be replaced by the constant-transconductance bias network of Fig. 6.2. In this case, V_{B1} and V_{B2} would be connected to $V_{casc\text{-}p}$ and $V_{casc\text{-}n}$, respectively, of Fig. 6.2 to maximize the output voltage signal swing.

An important addition of the folded-cascode opamp shown is the inclusion of two extra transistors, Q_{12} and Q_{13}. These two transistors serve two purposes. One is to increase the slew-rate performance of the opamp, as will be discussed in Example 6.2. However, more importantly, during times of slew-rate limiting, these transistors prevent the drain voltages of Q_1 and Q_2 from having large transients where they change from their small-signal voltages to voltages very close to the negative power-supply voltage. Thus the inclusion of Q_{12} and Q_{13} allows the opamp to recover more quickly following a slew-rate condition.

The compensation is realized by the load capacitor, C_L, and realizes dominant-pole compensation. In applications where the load capacitance is very small, it is necessary to add additional compensation capacitance in parallel with the load to guarantee stability. If lead compensation is desired, a resistor can be placed in series with C_L. While lead compensation may not be possible in some applications, such as when the compensation capacitance is mostly supplied by the load capacitance, it is more often possible than many designers realize (i.e., it is often possible to include a resistor in series with the load capacitance).

The bias currents for the input differential-pair transistors are equal to $I_{bias2}/2$. The bias current of one of the p-channel cascode transistors, Q_5 or Q_6, and hence the transistors in the output-summing current mirror as well, is equal to the drain current of Q_3 or Q_4 minus $I_{bias2}/2$. This drain current is established by I_{bias1} and the ratio of $(W/L)_3$, or $(W/L)_4$, to $(W/L)_{11}$. Since the bias current of one of the cascode transistors is derived by a current subtraction, for it to be accurately established, it is necessary that both I_{bias1} and I_{bias2} be derived from a single-bias network. In addition, any current mirrors used in deriving these currents should be composed of transistors realized as parallel combinations of unit-size transistors. This approach eliminates inaccuracies due to second-order effects caused by transistors having non-equal widths.

Small-Signal Analysis

In a small-signal analysis of the folded-cascode amplifier, it is assumed that the differential output current from the drains of the differential pair, Q_1, Q_2, is applied to the load capacitance, C_L. Specifically, the small-signal current from Q_1 passes directly

from the source to the drain of Q_6 and thus C_L, while the current from Q_2 goes indirectly through Q_5 and the current mirror consisting of Q_7 to Q_{10}[2]. Although these two paths have slightly different transfer functions due to the poles caused by the current mirror, when an n-channel transistor current mirror is used, a pole-zero doublet occurs at frequencies quite a bit greater than the opamp unity-gain frequency and can usually be ignored. Hence, ignoring high-frequency poles and zero, the approximate small-signal transfer function for the folded-cascode opamp is given by

$$A_V = \frac{V_{out}(s)}{V_{in}(s)} = g_{m1} Z_L(s) \tag{6.27}$$

Here, g_{m1} is the transconductance of each of the transistors in the input differential pair and $Z_L(s)$ is the impedance to ground seen at the output node. This impedance consists of the parallel combination of the output load capacitance, the impedance of any additional network added for stability, and the output impedance of the opamp (i.e., looking into the drains of Q_6 and Q_8). Thus, when the compensation is realized by the output capacitance only, we have

$$A_V = \frac{g_{m1} r_{out}}{1 + s r_{out} C_L} \tag{6.28}$$

where r_{out} is the output impedance of the opamp. This impedance is quite high, on the order of $g_m r_{ds}^2/2$ or greater if output-impedance enhancement is used.

For mid-band and high frequencies, the load capacitance dominates, and we can ignore the unity term in the denominator and thus have

$$A_V \cong \frac{g_{m1}}{sC_L} \tag{6.29}$$

from which the unity-gain frequency of the opamp is found to be

$$\omega_t = \frac{g_{m1}}{C_L} \tag{6.30}$$

Therefore, for large load capacitances, maximizing the transconductance of the input transistors maximizes the bandwidth, assuming the load capacitance is large enough so that the unity-gain frequency is much less than the limit imposed by the second poles. The transconductance of the input transistors is maximized by using wide n-channel devices and ensuring that the input transistor pair's bias current is substantially larger than the bias current of the cascode transistors and current mirror. Note that this approach also maximizes the dc gain (i.e., $g_{m1} r_{out}$) since not only does it maximize g_{m1}, but it also maximizes r_{out} by resulting in all transistors connected to the output node being biased at lower current levels (for a given total power dissipation). A practical upper limit on the ratio of the bias currents of the input transistors to the currents of the cascode transistors might be around four or so. If too high a ratio is used, the bias currents of the cascode transistors cannot be well established since they

2. This is based on the assumption that g_{m5} and g_{m6} are much larger than g_{ds3} and g_{ds4}.

are derived by current subtractions. Another advantage of having very large transconductances for the input devices is that the thermal noise due to this input pair is reduced. Thus, since much of the bias current in folded-cascode opamps flows through the input differential pair, these opamps often have a better thermal noise performance than other opamp designs having the same power dissipation.

When the load is a series combination of a compensation capacitance, C_L, and a lead resistance, R_C, the small-signal transfer function is given by

$$A_V = \frac{g_{m1}}{\dfrac{1}{r_{out}} + \dfrac{1}{R_C + 1/sC_L}} \cong \frac{g_{m1}(1 + sR_CC_L)}{sC_L} \tag{6.31}$$

where the approximate term is valid at mid and high frequencies. Here we see that, as in Chapter 5, R_C can be chosen to place a zero at 1.2 times the unity-gain frequency.

The second poles of this opamp are primarily due to the time constants introduced by the impedance and parasitic capacitances at the sources of the p-channel cascode transistors, Q_5 and Q_6. The impedances at these nodes is one over the transconductances of the cascode transistors.[3] Since p-channel transistors are used here, possibly biased at lower currents, these impedances are typically substantially greater than the source impedances of most n-channel transistors in the signal path. As a consequence, when high-frequency operation is important, these impedances can be reduced by making the currents in the p-channel cascode transistors around the same level as the bias currents of the input transistors. The parasitic capacitance at the sources of the cascode transistors is primarily due to the gate-source capacitances of the cascode transistors as well as the drain-to-bulk and drain-to-gate capacitances of the input transistors and the current-source transistors Q_3 and Q_4. Therefore, minimizing junction areas and peripheries at these two nodes is important.

Slew Rate

The diode-connected transistors, Q_{12} and Q_{13}, are turned off during normal operation and have almost no effect on the opamp. However, they substantially improve the operation during times of slew-rate limiting [Law, 1983]. To appreciate their benefit, consider first what happens during times of slew-rate limiting when they are not present. Assume there is a large differential input voltage that causes Q_1 to be turned on hard and Q_2 to be turned off. Since Q_2 is off, all of the bias current of Q_4 will be directed through the cascode transistor Q_5, through the n-channel current mirror, and out of the load capacitance. Thus, the output voltage will decrease linearly with a slew-rate given by

3. This is true only at high frequencies, where the output impedance of the amplifier is small due to the load and/or compensation capacitance. This is the case of interest here. At low frequencies, the impedance looking into the source of Q_6 is given by $R_{s6} = 1/g_{m6} + R_L/(g_{m6}r_{ds6})$, where R_L is the resistance seen looking out the drain of Q_6.

$$SR = \frac{I_{D4}}{C_L} \tag{6.32}$$

Also, since all of I_{bias2} is being diverted through Q_1, and since this current is usually designed to be greater than I_{D3}, both Q_1 and the current source I_{bias2} will go into the triode region, causing I_{bias2} to decrease until it is equal to I_{D3}. As a result, the drain voltage of Q_1 approaches that of the negative power-supply voltage. When the opamp is coming out of slew-rate limiting, the drain voltage of Q_1 must slew back to a voltage close to the positive power supply before the opamp operates in its linear region again. This additional slewing time greatly increases the distortion and also increases the transient times during slew-rate limiting (which occurs often for opamps used in switched-capacitor applications).

Next, consider the case where the diode-connected transistors, Q_{12} and Q_{13}, are included. *Their main purpose is to clamp the drain voltages of* Q_1 *or* Q_2 *so they don't change as much during slew-rate limiting.* A second, more subtle effect dynamically increases the bias currents of both Q_3 and Q_4 during times of slew-rate limiting. This increased bias current results in a larger maximum current available for charging or discharging the load capacitance. To see this increase in bias current, consider a case similar to the one just described, where a large differential input causes Q_1 to be fully on while Q_2 is off. In this case, the diode-connected transistor Q_{12} conducts with the current through Q_{12} coming from the diode-connected transistor Q_{11}. Thus, the current in Q_{11} increases, causing the currents in bias transistors Q_3 and Q_4 to also increase, until the sum of the currents of Q_{12} and Q_3 are equal to the bias current I_{bias2}. Note that the current in Q_4 also increases since it is equal to the current in Q_3. This increase in bias current of Q_4 results in an increase of the maximum current available for discharging C_L. In summary, not only are the voltage excursions less, but the maximum available current for charging or discharging the load capacitance is also greater during times of slew-rate limiting.

EXAMPLE 6.2

Find reasonable transistor sizes for the folded-cascode opamp shown in Fig. 6.9 to satisfy the following design parameters. Also find the unity-gain frequency and slew rate, both without and with the clamp transistors. What would be a reasonable size for a compensation resistor in series with C_L, if lead compensation were desired?

- Assume ±2.5-V power supplies are used and limit the power dissipation of the opamp to 2 mW.
- Set the ratio of the current in the input transistors to that of the cascode transistors to be 4:1. Also, set the bias current of Q_{11} to be 1/30th that of Q_3 (or Q_4) such that its current can be ignored in the power dissipation calculation.
- The maximum transistor width should be 300 μm and channel lengths of 1.6 μm should be used in all transistors.

- All transistors should have effective gate-source voltages of around 0.25 V except for the input transistors, whose widths should be set to the maximum value of 300 μm. Also, round all transistor widths to the closest multiple of 10 μm, keeping in mind that if a larger transistor is to be matched to a smaller one, the larger transistor should be built as a parallel combination of smaller transistors.
- Finally, assume the load capacitance is given by $C_L = 10$ pF, and that $\mu_n C_{ox} = 3\mu_p C_{ox} = 96\ \mu A/V^2$.

Solution

The total current in the opamp, I_{total}, excluding the current in the bias network, is equal to $I_{D3} + I_{D4}$, which is equal to $2(I_{D1} + I_{D6})$. Defining $I_B \equiv I_{D5} = I_{D6}$ and noting that we are asked to make $I_{D1} = 4I_{D6}$, we have

$$I_{total} = 2(I_{D1} + I_{D6}) = 2(4I_B + I_B) = 10I_B \tag{6.33}$$

Since the power dissipation is specified to be 2 mW, we have

$$I_B = I_{D5} = I_{D6} = \frac{I_{total}}{10} = \frac{P_{diss}/5\ V}{10} = 40\ \mu A \tag{6.34}$$

This bias current for I_B implies that $I_{D3} = I_{D4} = 5I_{D5} = 200\ \mu A$, and $I_{D1} = I_{D2} = 4I_{D5} = 160\ \mu A$. Now, we let all transistor channel lengths be 1.6 μm. This choice allows us to immediately determine the sizes of most transistors using

$$\left(\frac{W}{L}\right)_i = \frac{2I_{Di}}{\mu_i C_{ox} V_{effi}^2} \tag{6.35}$$

and then round to the closest multiple of 10 in transistor widths. When the widths found were larger than 300 μm, they were limited to 300 μm. Table 6.1 lists reasonable values for the resulting dimensions of all transistors. Note that the larger widths are exactly divisible by the smaller widths of a transistor of the same type and thus allows larger transistors to be realized as a parallel combination of smaller transistors. The width of Q_{11} was determined from the requirement that $I_{D11} = I_{D3}/30 = 6.6\ \mu A$. The widths of Q_{12} and Q_{13} were somewhat arbitrarily chosen to equal the width of Q_{11}. Note that, as requested, the widths of the input transistors were set to a value of 300 μm to maximize their transconductance. The transconductance of the input transistors

Table 6.1 The transistor dimensions (in μm) of the opamp of Fig. 6.9.

Q_1	300/1.6	Q_6	60/1.6	Q_{11}	10/1.6
Q_2	300/1.6	Q_7	20/1.6	Q_{12}	10/1.6
Q_3	300/1.6	Q_8	20/1.6	Q_{13}	10/1.6
Q_4	300/1.6	Q_9	20/1.6		
Q_5	60/1.6	Q_{10}	20/1.6		

is given by

$$g_{m1} = \sqrt{2I_{D1}\mu_n C_{ox}(W/L)_1} = 2.4 \text{ mA/V} \tag{6.36}$$

The unity-gain frequency of the opamp is given by

$$\omega_t = \frac{g_{m1}}{C_L} = 2.4\times10^8 \text{ rad/s} \Rightarrow f_t = 38 \text{ MHz} \tag{6.37}$$

A reasonable size for the lead resistor, R_C, in series with C_L, is given by

$$R_C = \frac{1}{1.2C_L\omega_t} = \frac{1}{1.2g_{m1}} = 347\ \Omega \tag{6.38}$$

The slew rate without the clamp transistors is given by

$$SR = \frac{I_{D4}}{C_L} = 20 \text{ V/}\mu\text{s} \tag{6.39}$$

When the clamp transistors are included, during slew-rate limiting, we have

$$I_{D12} + I_{D3} = I_{bias2} \tag{6.40}$$

But

$$I_{D3} = 30I_{D11} \tag{6.41}$$

and

$$I_{D11} = 6.6\ \mu\text{A} + I_{D12} \tag{6.42}$$

Substituting (6.41) and (6.42) into (6.40) and solving for I_{D11} gives

$$I_{D11} = \frac{I_{bias2} + 6.6\ \mu\text{A}}{31} \tag{6.43}$$

which implies that the value of I_{D11} during slew-rate limiting equals 10.53 μA and $I_{D3} = I_{D4} = 30I_{D11} = 0.32$ mA, which is substantially larger than the slew current available without the clamp transistors. This larger bias current will give a slew-rate value of

$$SR = \frac{I_{D4}}{C_L} = 32 \text{ V/}\mu\text{s} \tag{6.44}$$

More importantly, the time it takes to recover from slew-rate limiting will be substantially decreased.

6.3 CURRENT-MIRROR OPAMP

Another popular opamp often used when driving on-chip capacitive-only loads is shown in simplified form in Fig. 6.10. It is immediately seen that all nodes are low impedance except for the output node. By using good current mirrors having high

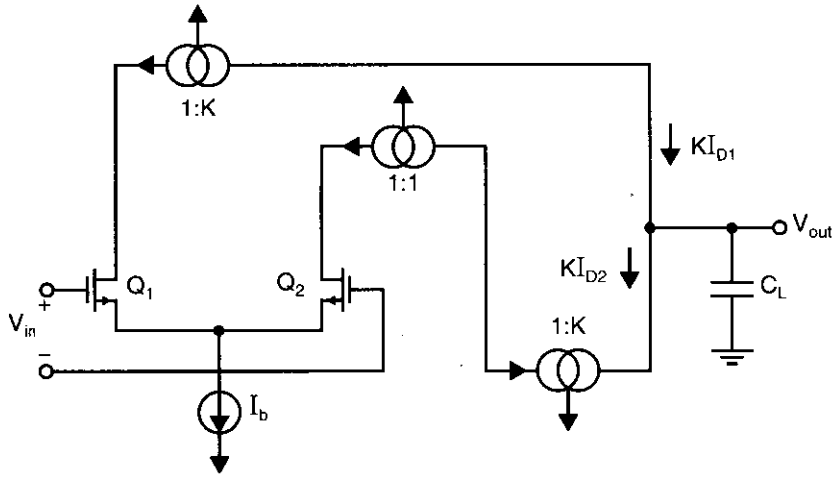


Fig. 6.10 A simplified current-mirror opamp.

output impedance, a reasonable overall gain can be achieved. A more detailed example of a current-mirror opamp is shown in Fig. 6.11.

The overall transfer function of this opamp will closely approximate a single-pole operation. In a similar analysis to that given for the folded-cascode opamp, we have

$$A_V = \frac{V_{out}(s)}{V_{in}(s)} = Kg_{m1}Z_L(s) = \frac{Kg_{m1}r_{out}}{1 + sr_{out}C_L} \cong \frac{Kg_{m1}}{sC_L} \tag{6.45}$$

The K factor is the current gain from the input transistors to the output sides of the current mirrors connected to the output node. Using (6.45), we can solve for the unity-gain frequency, resulting in

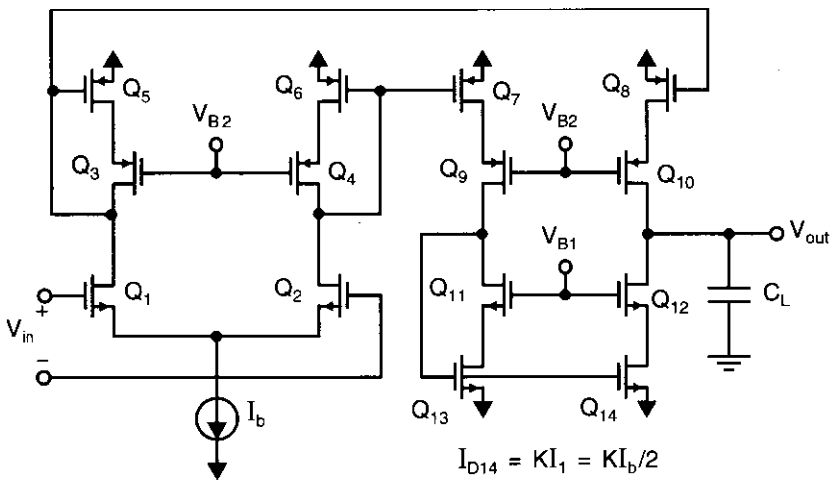


Fig. 6.11 A current-mirror opamp with wide-swing cascode current mirrors.

$$\omega_t = \frac{Kg_{m1}}{C_L} = \frac{K\sqrt{2I_{D1}\mu_n C_{ox}(W/L)_1}}{C_L} \tag{6.46}$$

If the power dissipation is specified, the total current,

$$I_{total} = (3+K)I_{D1} \tag{6.47}$$

is known for a given power-supply voltage. Substituting (6.47) into (6.46), we obtain

$$\omega_t = \frac{K\sqrt{2\left(\frac{I_{total}}{3+K}\right)\mu_n C_{ox}(W/L)_1}}{C_L} = \frac{K}{\sqrt{3+K}}\frac{\sqrt{2I_{total}\mu_n C_{ox}(W/L)_1}}{C_L} \tag{6.48}$$

Obviously, for larger values of K, the opamp's transconductance is larger (i.e., Kg_{m1}), and therefore, the unity-gain frequency is also larger. This simple result assumes the unity-gain frequency is limited by the load capacitance rather than any high-frequency poles caused by the time constants of the internal nodes. A practical upper limit on K might be around five. The use of large K values also maximizes the gain for I_{total} fixed since r_{out} is roughly independent of K for large K.

In the circuit shown in Fig. 6.11, the important nodes for determining the non-dominant poles are the drain of Q_1, primarily, and the drains of Q_2 and Q_9, secondly. Increasing K increases the capacitances of these nodes while also increasing the equivalent resistances.[4] As a result, the equivalent second pole moves to lower frequencies. If K is increased too much, an increase in C_L will be required to keep ω_t below the frequency of the equivalent second pole to maintain stability. Thus, increasing K decreases the bandwidth when the equivalent second poles dominate. In the case where the load capacitance is small, the equivalent second pole will limit the unity-gain frequency of the opamp, and if it is very important that speed is maximized, K might be taken as small as one. From experience it has been found that a reasonable compromise for a general-purpose opamp might be to let $K = 2$.

The slew rate of the current-mirror opamp is found by assuming there is a very large input voltage. In this case, all of the bias current of the first stage will be diverted through either Q_1 or Q_2. This current will be amplified by the current gain from the input stage to the output stage, (i.e., K). Thus, during slew-rate limiting, the total current available to charge or discharge C_L is KI_b. The slew rate is therefore given by

$$SR = \frac{KI_b}{C_L} \tag{6.49}$$

For a given total power dissipation, this slew rate is maximized by choosing a large K value. For example, with $K = 4$ and during slew-rate limiting, 4/5 of the total bias current of the opamp will be available for charging or discharging C_L. This result gives a current-mirror opamp superior slew rates when compared to a folded-cascode

4. Increasing K implies that the currents at the input sides of the current mirrors are smaller, for a given total power dissipation. Also, the widths of transistors at the input sides of current mirrors will be smaller. Both of these effects cause the transconductances of transistors at the input side of current mirrors to be smaller, and therefore the impedances (in this case, 1/gm) will be larger.

opamp, even when the clamp transistors have been included in the folded-cascode opamp. Also, there are no problems with large voltage transients during slew-rate limiting for the current-mirror opamp.

In summary, due primarily to the larger bandwidth and slew rate, the current-mirror opamp is usually preferred over a folded-cascode opamp. However, it will suffer from larger thermal noise when compared to a folded-cascode amplifier because its input transistors are biased at a lower proportion of the total bias current and therefore have a smaller transconductance. As seen in the chapter on noise, a smaller transconductance results in larger thermal noise when the noise is referred back to the input of the opamp.

EXAMPLE 6.3

Assume the current-mirror opamp shown in Fig. 6.11 has all transistor lengths equal to 1.6 μm and transistor widths as given in Table 6.2. Notice that $K = 2$. Assume $C_L = 10$ pF, $\mu_n C_{ox} = 3\mu_p C_{ox} = 96\ \mu A/V^2$, the total power dissipation is specified to be 2 mW (ignoring the power dissipation of the bias network, which would normally be small), and ±2.5 V power supplies are used.

Find the slew rate and the unity-gain frequency, assuming the equivalent second pole does not dominate. Estimate the equivalent second pole, assuming $C_{ox} = 1.92\ fF/\mu m^2$. Would it be necessary to increase C_L if a 75° phase margin were required without using lead compensation? What if lead compensation were used?

Solution

Since the total bias current is given by $(3 + K)I_b/2$, we have

$$I_b = \frac{2I_{total}}{(3+K)} = \frac{2P_{diss}/5\ V}{(3+K)} = 160\ \mu A \tag{6.50}$$

Thus, the bias currents of all transistors, except those in the output stage, are 80 μA. The bias currents of transistors in the output stage are twice that of the input stage, which is 160 μA. The transconductance of the input transistors is

$$g_{m1} = \sqrt{2I_{D1}\mu_n C_{ox}(W/L)_1} = 1.7\ mA/V \tag{6.51}$$

Table 6.2 The transistor sizes (in μm) for the opamp used in Example 6.3.

Q_1	300/1.6	Q_7	60/1.6	Q_{13}	30/1.6
Q_2	300/1.6	Q_8	120/1.6	Q_{14}	60/1.6
Q_3	60/1.6	Q_9	60/1.6		
Q_4	60/1.6	Q_{10}	120/1.6		
Q_5	60/1.6	Q_{11}	30/1.6		
Q_6	60/1.6	Q_{12}	60/1.6		

The transconductance of the opamp will be K times this or 3.4 mA/V. The unity-gain frequency is now given by

$$\omega_t = \frac{Kg_{m1}}{C_L} = 3.4\times10^8 \text{rad/s} \Rightarrow f_t = 54 \text{ MHz} \tag{6.52}$$

This result should be compared to the 38 MHz unity-gain frequency of the 2 mW folded-cascode opamp of the previous example, showing that current-mirror opamps are superior when driving large capacitive loads.

Continuing, the slew rate is given by

$$SR = \frac{KI_b}{C_L} = 32 \text{ V}/\mu\text{s} \tag{6.53}$$

which compares favorably to 20 V/μs for the folded-cascode opamp without the clamp transistors (although it equals the value of 32 V/μs for the folded-cascode opamp with the clamp transistors).

When estimating the equivalent second pole, the junction capacitances will be ignored to simplify matters. In reality, the drain-to-bulk capacitances of the input transistors could contribute a significant factor, and this amount can be determined using simulation. The dominant node almost certainly will occur at the drain of Q_1. The impedance at this node is given by

$$R_1 = 1/g_{m5} = 2.29 \text{ k}\Omega \tag{6.54}$$

In addition, the capacitance will be primarily due to the gate-source capacitances of Q_5 and Q_8. We therefore have

$$C_1 = C_{gs5} + C_{gs8} = (1+K)C_{gs5} \tag{6.55}$$

resulting in

$$C_1 = (1+K)C_{ox}W_5 L_5 = 0.553 \text{ pF} \tag{6.56}$$

With these values of impedance, the time constant for this node is given by

$$\tau_1 = R_1C_1 = 1.27 \text{ ns} \tag{6.57}$$

In a similar manner, we can calculate the impedances and, hence, the time constant for the drain of Q_2 to be $R_2 = 2.29$ kΩ, $C_2 = 0.369$ pF, and $\tau_2 = R_2C_2 = 0.85$ ns. The other important time constant comes from the parasitic capacitors at the drain of Q_9. Here, we have

$$R_3 = 1/g_{m13} = 1.86 \text{ k}\Omega \tag{6.58}$$

and

$$C_3 = C_{gs13} + C_{gs14} = (1+K)C_{ox}W_{13} L_{13} = 0.276 \text{ pF} \tag{6.59}$$

resulting in $\tau_3 = 0.51$ ns. The time constant of the equivalent second pole can now be estimated to be given by

$$\tau_{2eq} = \tau_1 + \tau_2 + \tau_3 = 2.6 \text{ ns} \tag{6.60}$$

This result gives an equivalent second pole at the approximate frequency of

$$p_{2eq} = \frac{1}{\tau_{2eq}} = 3.8\times10^8 \text{ rad/s} = 2\pi \times 61 \text{ MHz} \tag{6.61}$$

If lead compensation is not used and a 75° phase margin is required, the unity-gain frequency must be constrained to be less than 0.27 times the equivalent second pole from Table 5.1 of Chapter 5. Specifically, the unity-gain frequency must be less than 16.5 MHz. To achieve this unity-gain frequency, we need to increase C_L by the factor 54 MHz/16.5 MHz, resulting in C_L being 32.7 pF.

If lead compensation is to be used, the unity-gain frequency can be chosen so that only a 55° phase margin is achieved *before* the lead resistor is added. This approach would allow the unity-gain frequency to be as high as 0.7 times the equivalent second-pole frequency, which is 42.7 MHz. This unity-gain frequency can be realized by increasing C_L to 10 pF times 54 MHz/42.7 MHz, resulting in the new value of C_L being 12.6 pF. After the lead resistor is added, the unity-gain frequency increases by about 20 percent to 51 MHz, which is *over a three times improvement as compared to when lead compensation is not used.* The benefits of lead compensation are obvious.

As a final comment, notice that the delay through the signal path from the drain of Q_2 to the output (i.e., $\tau_2 + \tau_3 = 1.36$ ns) is approximately the same as the delay through the signal path from the drain of Q_1 to the output, which is $\tau_1 = 1.27$ ns. This near equivalence helps to minimize harmful effects caused by the pole-zero doublet introduced at high frequencies (i.e., around the frequency of the second pole) by the existence of two signal paths.

6.4 LINEAR SETTLING TIME REVISITED

We saw in Chapter 5 that the time constant for linear settling time was equal to the inverse of $\omega_{-3\,dB}$ for the closed-loop circuit gain. We also saw that $\omega_{-3\,dB}$ is given by the relationship

$$\omega_{-3\,dB} = \beta\omega_t \tag{6.62}$$

However, while for the classical two-stage CMOS opamp the unity-gain frequency remains relatively constant for varying load capacitances, the unity-gain frequencies of the folded-cascode and current-mirror amplifiers are strongly related to their load capacitance. As a result, their settling-time performance is affected by both the feedback factor as well as the effective load capacitance.

For the folded-cascode opamp, its unity-gain frequency is given by

$$\omega_t = \frac{g_{m1}}{C_L} \tag{6.63}$$

whereas for a current-mirror opamp, it is equal to

$$\omega_t = \frac{Kg_{m1}}{C_L} \tag{6.64}$$

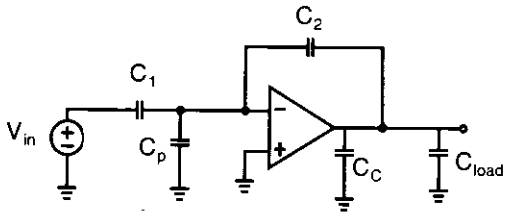


Fig. 6.12 A general circuit for determining the −3-dB frequency of a closed-loop amplifier.

Thus we see that for both these high-output impedance opamps, their unity-gain frequency is inversely proportional to the load capacitance. To determine the −3-dB frequency of a closed-loop opamp, consider the general case shown in Fig. 6.12. At the opamp output, C_{load} represents the capacitance of the next stage that the opamp must drive, while C_C is a compensation capacitance that might be added to maintain a sufficient phase margin. At the input side, C_p represents parasitic capacitance due to large transistors at the opamp input as well as any switch capacitance.

The feedback network is due to capacitances C_1, C_2, and C_p resulting in

$$\beta = \frac{1/[s(C_1 + C_p)]}{1/[s(C_1 + C_p)] + 1/(sC_2)} = \frac{C_2}{C_1 + C_p + C_2} \tag{6.65}$$

The effective load capacitance seen by the opamp output is found by analyzing this circuit from a series-shunt feedback perspective. Specifically, treating the negative opamp input as an open circuit, the effective load capacitance, C_L, is seen to be given by the parallel combination of C_C and C_{load}, as well as that seen looking into C_2. The capacitance seen looking into C_2 is equal to the series combination of C_2 together with $C_1 + C_p$. Combining all these capacitances together, we have

$$C_L = C_C + C_{load} + \frac{C_2(C_1 + C_p)}{C_1 + C_p + C_2} \tag{6.66}$$

EXAMPLE 6.4

Consider the current-mirror opamp with no lead compensation in Example 6.3 being used in the circuit shown in Fig. 6.12 with $C_1 = C_2 = C_C = C_{load} = 5$ pF. What is the linear settling time required for 0.1 percent accuracy?

Solution

First, the gate-source capacitances of the input devices of the current-mirror opamp can be calculated to be

$$C_{gs1} = 300 \times 1.6 \times 1.92 \text{ fF}/\mu\text{m}^2 = 0.92 \text{ pF} \tag{6.67}$$

The capacitance seen looking into the inverting input of the opamp is one-half this value since the gate-source capacitances of the two input devices are in series. Thus, the parasitic capacitance, C_P, is 0.46 pF. Therefore, the effective load capacitance is given by

$$C_L = 5 + 5 + \frac{5(5 + 0.46)}{5 + 5 + 0.46} = 12.61 \text{ pF} \tag{6.68}$$

which results in

$$\omega_t = \frac{Kg_{m1}}{C_L} = \frac{2 \times 1.7 \text{ mA/V}}{12.61 \text{ pF}} = 2.70 \times 10^8 \text{ rad/s} \tag{6.69}$$

or equivalently,

$$f_t = 42.9 \text{ MHz} \tag{6.70}$$

Now, the feedback factor, β, is seen to be

$$\beta = \frac{5}{5 + 0.46 + 5} = 0.48 \tag{6.71}$$

resulting in

$$\tau = \frac{1}{\beta\omega_t} = 7.8 \text{ ns} \tag{6.72}$$

Finally, for 0.1 percent accuracy, we need a linear settling time of 7τ or 54 ns.

6.5 FULLY DIFFERENTIAL OPAMPS

Most modern high-performance analog integrated circuits make use of fully differential signal paths. With opamps, this technique results in differential outputs as well as inputs, and hence, they are referred to as fully differential opamps. One of the major driving forces behind the use of fully differential signals is to help reject noise from the substrate as well as from pass-transistor switches turning off in switched-capacitor applications. The reason for this noise rejection is that if the circuit is built in a symmetric manner (sometimes referred to as a balanced circuit), then ideally the noise will affect both signal paths identically, and will then be rejected, since only the difference between signals is of importance. In other words, the noise will have no effect on the differential signal, which is the signal of interest, since both sides of the differential signal see the same noise. In reality, this rejection only partially occurs since, unfortunately, the mechanisms introducing the noise are usually nonlinear with respect to voltage levels. For example, substrate noise will usually feed in through junction capacitances, which are nonlinear with voltage. Also, the clock feedthrough noise introduced by switches turning off usually has some voltage-dependent nonlinearities that can cause more noise to feed into one path than the other, thereby injecting a differential noise. However, almost certainly, the noise rejection of a fully differential design will be much better than that for a single-ended output design.

One drawback of using fully differential opamps is that a common-mode feedback (CMFB) circuit must be added. This extra circuitry is needed to establish the

common-mode (i.e., average) output voltage. Ideally, it will keep this common-mode voltage immovable, preferably close to halfway between the power-supply voltages, even when large differential signals are present. Without it, the common-mode voltage is left to drift, since, although the opamp is placed in a feedback configuration, the common-mode loop gain is not typically large enough to control its value. Such is not the case with differential signals as the differential loop gain is typically quite large.

As we will see in the next section, the design of a good common-mode feedback circuit is nontrivial. First, it usually must have a speed performance comparable to the unity-gain frequency of the differential path. This speed requirement is necessary, otherwise noise on the power supplies may be significantly amplified such that the output signal becomes clipped (or distorted). Second, straightforward continuous-time designs often don't work when large differential signals are present, which may result in large common-mode signals being injected. Typically, these continuous-time CMFB circuits are the major limitation on maximum allowable signals, often limiting the signals substantially more than the differential signal-path circuitry does. Also, even when switched-capacitor CMFB circuits (described in the next section) are used, although the signal swings are not usually limited, the CMFB circuitry is often a source of noise injection and can also substantially increase the load capacitance that needs to be driven.

Another drawback of fully differential opamps is that in many designs, the single-ended slew-rate in one direction is substantially reduced when it is compared to that of equivalent single-ended output designs. This slew-rate reduction occurs because the maximum current for slewing is often limited by fixed-bias currents in the output stages. For example, in the fully differential amplifiers to be presented shortly, the negative-going current in the output stage is fixed. The unity-gain frequency, however, is usually increased because one of the current-mirrors is typically eliminated from the signal path.

Regardless of some of the limitations just described, a well-designed, fully differential amplifier works very well and can substantially improve the noise rejection. For this reason, differential designs are becoming more and more popular, perhaps constituting a majority in new microcircuit designs.

Fully Differential Folded-Cascode Opamp

A simplified schematic of a fully differential folded-cascode opamp is shown in Fig. 6.13. The n-channel current-mirror of Fig. 6.9 has been replaced by two cascode current sources, one composed of Q_7 and Q_8, the other composed of Q_9 and Q_{10}. In addition, a common-mode feedback (CMFB) circuit has been added. The gate voltage on the drive transistors of these current sources is determined by the output voltage, V_{cntrl}, of the common-mode feedback circuit. The inputs to the CMFB circuit are the two outputs of the fully differential amplifier. The CMFB circuit will detect the average of these two outputs and force it to be equal to a predetermined value, as we shall see in the next section.

Note that when the opamp output is slewing, the maximum current available for the negative slew rate is limited by the bias currents of Q_7 or Q_9. If the CMFB circuit

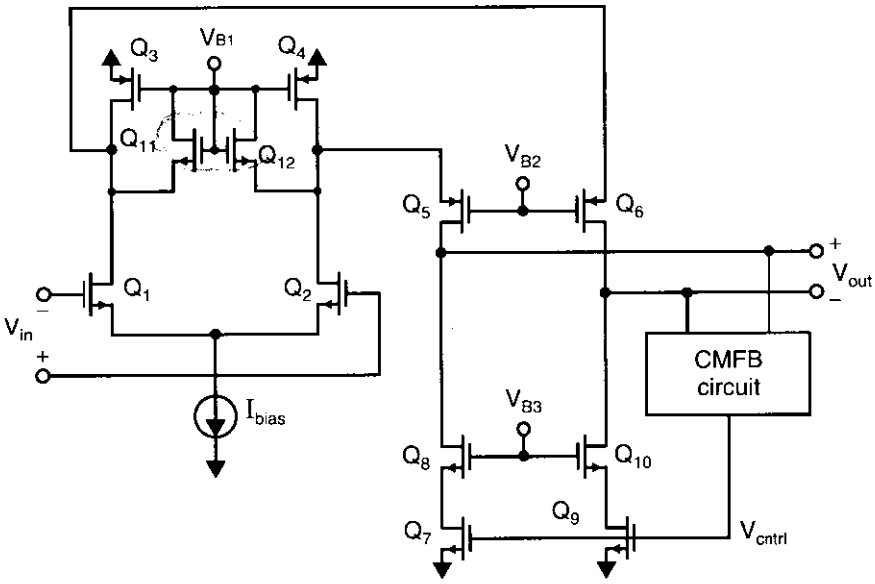


Fig. 6.13 A fully differential folded-cascode opamp.

is very fast, these will be increased dynamically to some degree during slewing, but seldom to the degree of a single-ended output fully differential opamp. For this reason, fully differential folded-cascode opamps are usually designed with the bias currents in the output stage equal to the bias currents in the input transistors. Also, to minimize transient voltage changes during slew-rate limiting, clamp transistors Q_{11} and Q_{12} have been added, as in the single-ended design presented earlier. If these transistors are not included (as has historically been the case), the time to recover from slew-rate limiting of the folded-cascode amplifier can be comparatively poor.

Each signal path now consists of only one node in addition to the output nodes, namely the drain nodes of the input devices. These nodes will most certainly be responsible for the equivalent second poles. In cases where the load capacitance is small, and it is important to maximize the bandwidth, a complementary design (i.e., n- and p-channel transistors and the power supplies interchanged) would be preferable to maximize the frequency of the equivalent second poles. This complementary circuit would result in the impedance at the drains of the input devices being the inverse of the transconductance of n-channel transistors, rather than p-channel transistors, resulting in smaller impedances and therefore faster time constants. The trade-off involved is that the opamp's transconductance, and therefore the opamp's dc gain, would become smaller due to the input transistors now being p-channel transistors. Also, the CMFB circuitry could possibly be slower, as the current sources being modulated would also have to be p-channel drive transistors. In any case, a complementary design is often a reasonable choice for high-speed fully differential designs.

Alternative Fully Differential Opamps

An alternative design is the fully differential current-mirror opamp, as shown in simplified form in Fig. 6.14. As with the folded-cascode design, this circuit can be realized as shown or in a complementary design, with p-channel input transistors, n-channel current mirrors, and p-channel bias current sources. Which design is preferable is primarily determined by whether the load capacitance or the equivalent second poles are limiting the bandwidth, and by whether maximizing the dc gain or the bandwidth is more important. n-channel inputs are preferable in the former case and p-channel inputs are preferable in the latter case, for reasons similar to those discussed for the folded-cascode opamp. Also, n-channel input transistors will give lower thermal noise (due to their larger transconductances), whereas p-channel inputs will result in less opamp input-referred 1/f noise.

If a general-purpose fully differential opamp is desired, then this design with large p-channel input transistors, a current gain of $K = 2$, and wide-swing enhanced output-impedance cascode mirrors and current sources is probably a good choice, compared to all other fully differential alternatives.

One of the limitations of the fully differential opamps seen so far is that the maximum current at the output for single-ended slewing in one direction is limited by fixed current sources. It is possible to modify the designs to get bidirectional drive capability at the output, as is shown in Fig. 6.15. This opamp is similar to the current mirror design of Fig. 6.14, but now the current mirrors at the top (i.e., the p-channel current mirrors) have been replaced by current mirrors having two outputs. The first

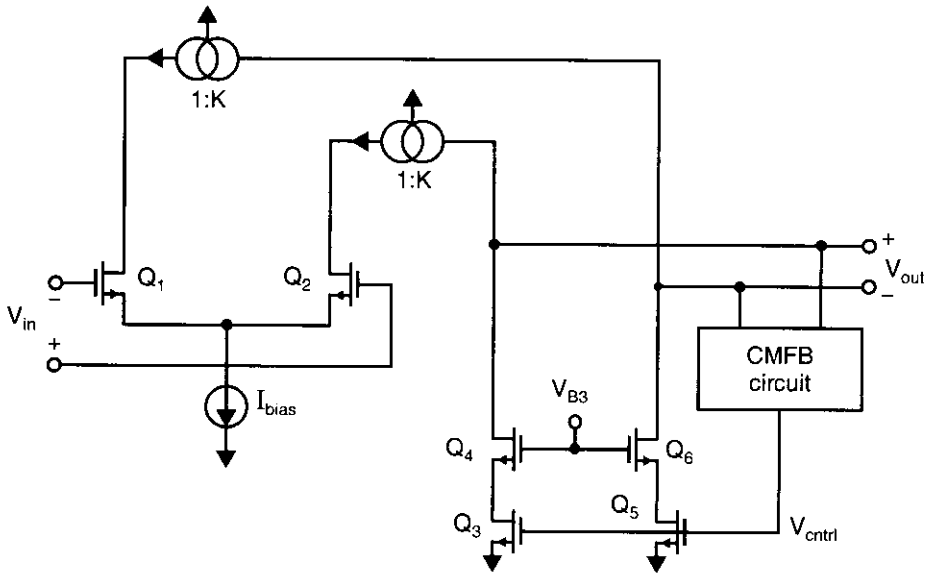


Fig. 6.14 A fully differential current-mirror opamp.

output has a current gain of K and goes to the output of the opamp, as before. The second output has a gain of one and goes to a new n-channel current mirror that has a current gain of K, where it is mirrored a second time and then goes to the opamp's opposite output, as shown. Assume now the differential output is slewing in the positive direction. For this case, the differential input will be very positive and the slewing current going into V_{out+} will be KI_{bias}, but due to the additional n-channel mirrors, the current being sinked from V_{out-} will also be KI_{bias}. This circuit has an improved slew rate at the expense of slower small-signal response due to the addition of the extra outputs on the p-channel mirrors and the additional n-channel mirrors, although in many applications this trade-off may be well merited. Of course, common-mode feedback must be added to the circuit of Fig. 6.15.

A third alternative, which makes use of a class-AB input stage, is shown in Fig. 6.16 [Castello, 1985]. Note here that for simplicity, the CMFB circuit is not shown, although it is required. This design has two differential-pair input stages connected in parallel. Each differential pair is composed of an n-channel and a p-channel transistor, and some level-shift transistors. Specifically, one input stage consists of a level-shift circuit composed of source follower Q_1 and diode-connected Q_2, and a differential pair consisting of p-channel transistor Q_3 and n-channel transistor Q_4. The other input stage consists of level shifter Q_5 and Q_6, and differential pair Q_7 and Q_8. In addition, there are four current mirrors to sum the drain currents of all differential pair transistors at the output nodes. At each output node, currents originally coming from the n-channel transistor of one differential pair and from the p-channel transistor of the other differential pair are summed.

The advantage of the input stage in this opamp is that during slew-rate limiting, one differential pair will turn off, but the total current in the other differential pair will dynamically increase substantially. For example, assuming a large positive input

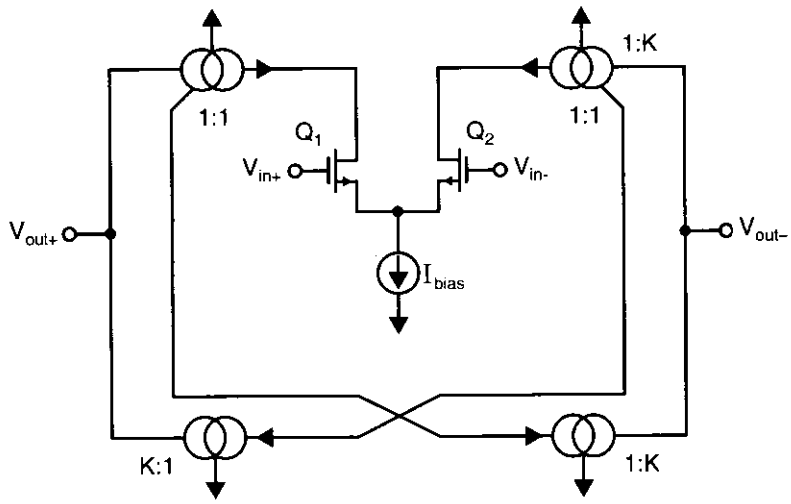


Fig. 6.15 A fully differential opamp with bidirectional output drive.

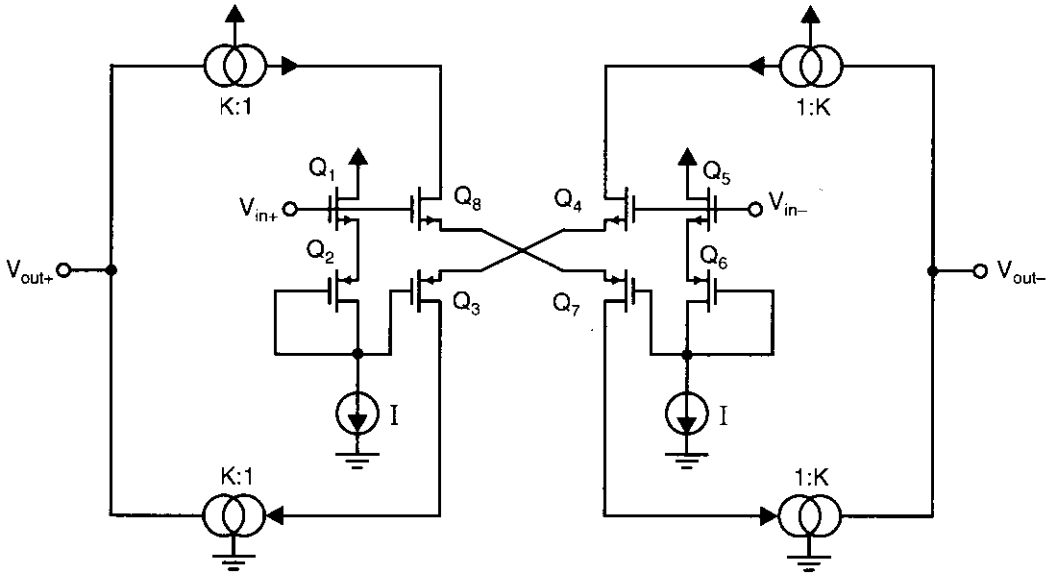


Fig. 6.16 A class AB fully differential opamp. CMFB circuit not shown.

voltage, the pair Q_3, Q_4 turns off, while the current through the pair Q_7, Q_8 increases. As a result, the increased current through Q_8 goes through a current mirror to the V_{out+} node, causing it to slew in a positive direction, whereas the current from Q_7 goes through a current mirror to the V_{out-} node, causing it to slew in a negative direction. The currents at this time of slew-rate limiting will be much larger than the small-signal bias currents, giving this opamp a very large slew-rate performance.

The disadvantage of this design is that the level-shift circuitry required at the input increases the noise and adds additional parasitics, which contribute to the equivalent second pole. In addition, the common-mode range of the input must remain at least $2V_t + 3V_{eff}$ above the lower power supply (and typically higher for the slew-rate performance to be maintained). This is a major problem when 5-V power supplies are being used, and it effectively eliminates this design from consideration for use with 3.3-V power supply voltages. However, for applications where the power-supply voltages are large, the load capacitances are large, and the slew rate is very important, this approach is quite reasonable.

Another alternative for a fully differential opamp is to use two single-ended output opamps with their inputs connected in parallel and each of their outputs being one output side of the fully differential circuit. An example of such an approach using current-mirror subcircuits is shown in Fig. 6.17. This design also has a fairly large slew rate when compared to the simpler fully differential current-mirror opamp of Fig. 6.14, but has additional current mirrors and complexity. Another advantage of this design is that the noise voltages due to the input transistors sum in a squared

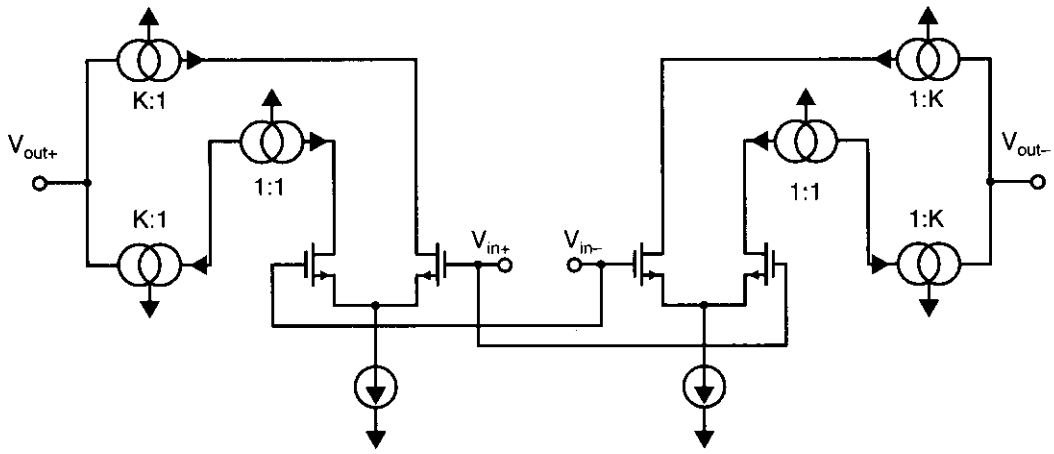


Fig. 6.17 A fully differential opamp composed of two single-ended output current-mirror opamps. CMFB circuit not shown.

fashion, while the two signal paths sum linearly.[5] As a result, this design has an improvement in signal-to-noise ratio of the input-referred gain by a factor of $\sqrt{2}$, or 3 dB. Also, the increase in total power dissipation is not significant when K is moderately large. As for the previous design, the compensation of the common-mode feedback loop is more difficult than for designs having fixed current sources biasing the output stages.

Another alternative design is shown in somewhat simplified form in Fig. 6.18. The advantage of this circuit is that due to the use of both n-channel and p-channel transistors in the two differential input pairs, the input common-mode voltage range is increased [Babanezhad, 1988; Hogervorst, 1992; Coban, 1994]. This feature can be particularly important when low power-supply voltages are being used. When the input common-mode voltage range is close to one of the power-supply voltages, one of the input differential pairs will turn off, but the other one will remain active. In an effort to keep the opamp gain relatively constant during this time, the bias currents of the still-active differential pair are dynamically increased. For example, if the input common-mode voltage range was close to the positive power-supply voltage, Q_3 and Q_4 would turn off, and Q_6 would conduct all of I_2. This current would go through current mirror M_1 and increase the bias current of the differential pair consisting of Q_1 and Q_2, which is still active. A similar situation occurs if the input common-mode voltage is near the negative power-supply rail. With careful design, it has been reported that the transconductance of the input stage can be held constant to within 15 percent of its nominal value with an input common-mode voltage

5. See Chapter 4 regarding noise.

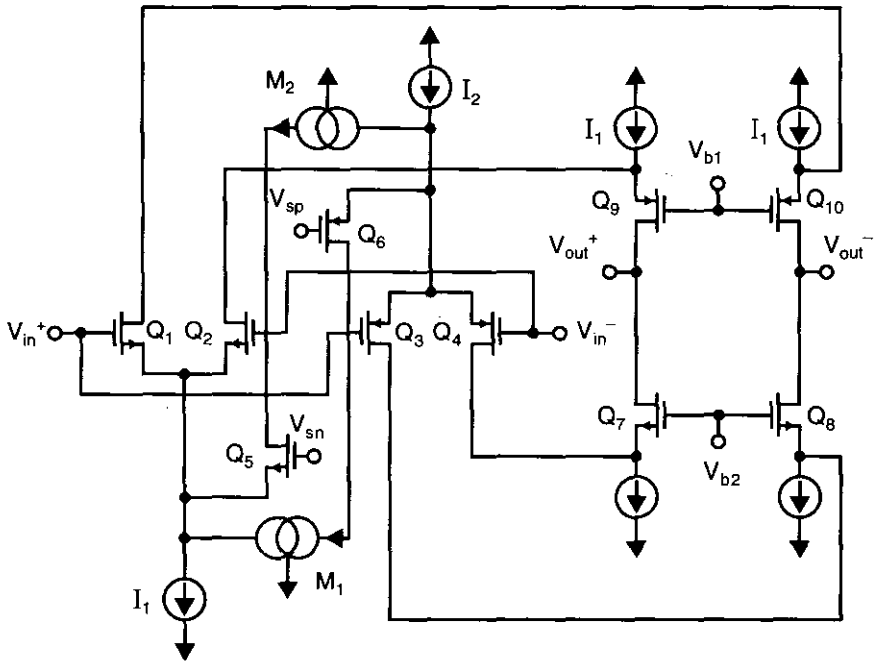


Fig. 6.18 An opamp having rail-to-rail input common-mode voltage range. CMFB circuit not shown.

range as large as the difference between the power-supply voltages [Coban, 1994].

6.6 COMMON-MODE FEEDBACK CIRCUITS

Typically, when using fully-differential opamps in a feedback application, the applied feedback determines the differential signal voltages, but does not affect the common-mode voltages. It is therefore necessary to add additional circuitry to determine the output common-mode voltage and to control it to be equal to some specified voltage, usually about halfway between the power-supply voltages. This circuitry, referred to as the *common-mode feedback* (CMFB) *circuitry,* is often the most difficult part of the opamp to design.

There are two typical approaches to designing CMFB circuits—a continuous-time approach and a switched-capacitor approach. The former approach is often the limiting factor on maximizing the signal swings, and, if nonlinear, may actually introduce common-mode signals. The latter approach is typically only used in switched-capacitor circuits, since in continuous-time applications it introduces clock-feedthrough glitches.

An example of a continuous-time CMFB circuit is shown in Fig. 6.19 [Martin, 1985; Whatly, 1986]. To illustrate its operation, assume a common-mode output voltage of zero and that V_{out+} is equal in magnitude, but opposite in sign, to V_{out-}. Furthermore, assume the two differential pairs have infinite common-mode input

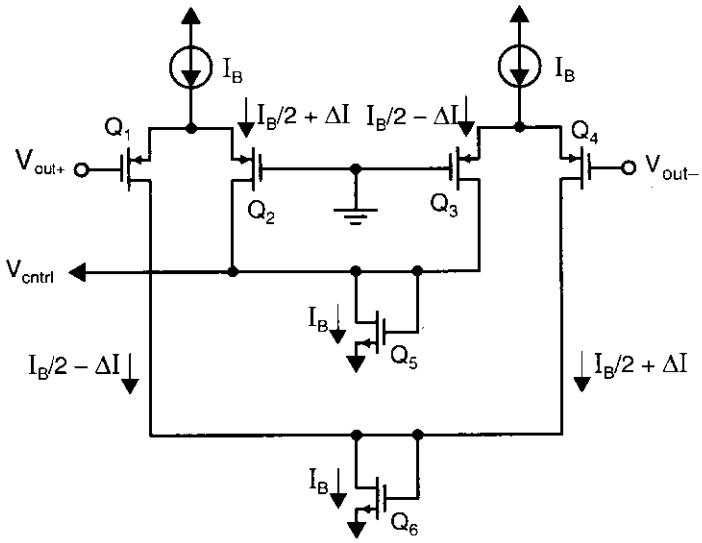


Fig. 6.19 An example of a continuous-time CMFB circuit.

rejection, which implies that the large-signal output currents of the differential pairs depend only on their input differential voltages. Since the two pairs have the same differential voltages being applied, the current in Q_1 will be equal to the current in Q_3, while the current in Q_2 will equal the current in Q_4. This result is valid independent of the nonlinear relationship between a differential pair's input voltage and its large-signal differential drain currents. Now, letting the current in Q_2 be denoted as $I_{D2} = I_B/2 + \Delta I$, where I_B is the bias current of the differential pair and ΔI is the large-signal current change in I_{D2}, the current in Q_3 is given by $I_{D3} = (I_B/2) - \Delta I$, and the current in Q_5 is given by

$$I_{D5} = I_{D2} + I_{D3} = (I_B/2 + \Delta I) + ((I_B/2) - \Delta I) = I_B \tag{6.73}$$

Thus, as long as the voltage V_{out+} is equal to the negative value of V_{out-}, the current through diode-connected Q_5 will not change even when large differential signal voltages are present. Since the voltage across diode-connected Q_5 is used to control the bias voltages of the output stage of the opamps, this means that when no common-mode voltage is present, the bias currents in the output stage will be the same regardless of whether a signal is present or not. Note, however, that the above result does not remain valid if the output voltage is so large that transistors in the differential pairs turn off.

Next consider what happens when a common-mode voltage other than zero[6] is present. For example, assume a positive common-mode signal is present. This positive voltage will cause the currents in both Q_2 and Q_3 to increase, which

6. For this example, it is assumed that positive and negative power supplies are present and that the desired common-mode voltage is zero. In a single-supply application, the gates of Q_2 and Q_3 would be connected to a bias voltage equal to the desired output common-mode voltage.

causes the current in diode-connected Q_5 to increase, which in turn causes its voltage to increase. This voltage is the bias voltage that sets the current levels in the n-channel current sources at the output of the opamp. Thus, both current sources will have larger currents pulling down to the negative rail, which will cause the common-mode voltage to decrease, bringing the common-mode voltage back to zero (or almost zero). Thus, as long as the common-mode loop gain is large enough, and the differential signals are not so large as to cause transistors in the differential pairs to turn off, the common-mode output voltage will be kept very close to ground.

The size of the differential signals that can be processed without one of the differential-pair signals turning off is maximized if the differential-pair transistors are designed to have large effective gate-source voltages. Alternatively, source degeneration can be used to allow them to have larger input signals without all of the current being directed to one side of the differential pair. However, even when this maximization is performed, the CMFB circuit still limits the differential signals to be less than what can be processed by the rest of the opamp.

Finally, it should be noted that when realizing the CMFB circuit of Fig. 6.19, the current sources I_B should be high-output impedance cascode current sources to ensure good common-mode rejection of the two differential pairs.

A modified version of the CMFB circuit of Fig. 6.19 is shown in Fig. 6.20 [Duque-Carillo, 1993]. This version adds an additional transistor (Q_7) and an additional current source to make use of the other sides of the differential pairs, which were previously unused in Fig. 6.19. This change effectively doubles the common-mode gain of the circuit for little additional complexity. As reported in [Duque-Carillo, 1993], it has very good linearity (better than 0.01%).

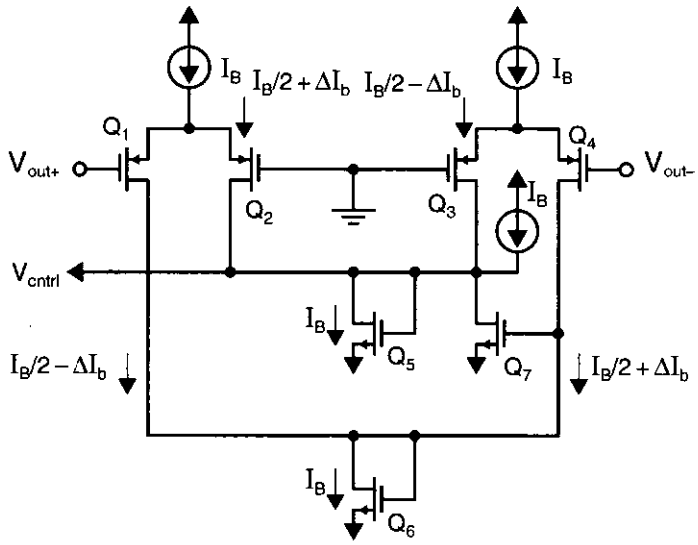


Fig. 6.20 A modified CMFB circuit having twice the common-mode gain as that of Fig. 6.19.

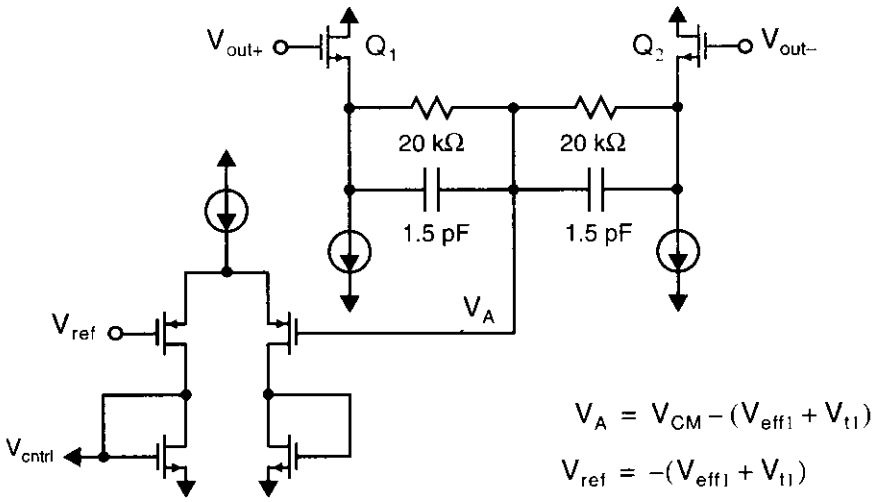


Fig. 6.21 An alternative continuous-time CMFB circuit.

An alternative approach for realizing CMFB circuits is shown in Fig. 6.21 [Banu, 1988]. This circuit generates the common-mode voltage of the output signals (minus a dc level shift) at node V_A. This voltage is then compared to a reference voltage, V_{ref}, using a separate amplifier. Although this approach works well, it has a major limitation in that the voltage drop across the source-follower transistors Q_1, Q_2 severely limits the differential signals that can be processed (unless transistors with native threshold voltages, such as 0.3 volts, are available). This limitation is particularly important when small power-supply voltages are used. In addition, the additional nodes of the common-mode feedback circuitry make the circuit slightly more difficult to compensate. When bipolar transistors are available, as is the case in a BiCMOS process, this approach is much more desirable.

An important consideration when designing CMFB circuits is that they be well compensated. Otherwise, the injection of common-mode signals can cause them to ring or even possibly become unstable. Thus, when the circuit is being designed, the phase margin and step-response of the common-mode loop should be found and verified by simulation. Often, the compensation of the common-mode loop can be realized using the same compensation capacitors used to stabilize the differential loop. This multipurpose compensation is achieved by connecting two compensation (or loading) capacitors between the opamp outputs and ground (or some other reference voltage). The other approach would be to use a single compensation capacitor connected directly between the two outputs, but this method only compensates the differential loop. It should be mentioned that by having as few nodes in the common-mode loop as is possible, compensation is simplified without having to severely limit the speed of the CMFB circuit. For this reason, the CMFB circuit is usually used to control current sources in the output stage of the opamp, as opposed to the current sources in the input stage of the opamp. The high speed of the CMFB circuit is neces-

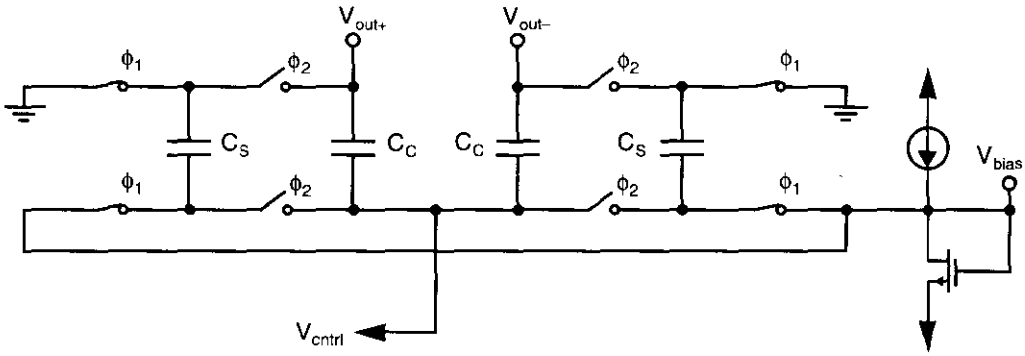


Fig. 6.22 A switched-capacitor CMFB circuit.

sary to minimize the effects of high-frequency common-mode noise, which could be amplified, causing the opamp outputs to saturate. Finally, it should be stated that designing continuous-time CMFB circuits that are both linear and operate with low power-supply voltages is an area of continuing research.

A third approach for realizing CMFB circuits is based on the use of switched-capacitor circuits. An example of this approach is shown in Fig. 6.22 [Senderowicz, 1982; Castello, 1985]. In this approach, capacitors labelled C_C generate the average of the output voltages, which is used to create control voltages for the opamp current sources. The dc voltage across C_C is determined by capacitors C_S, which are switched between bias voltages and between being in parallel with C_C. This circuit acts much like a simple switched-capacitor low-pass filter having a dc input signal. The bias voltages are designed to be equal to the difference between the desired common-mode voltage and the desired control voltage used for the opamp current sources.

The capacitors being switched, C_S, might be between one-quarter and one-tenth the sizes of the nonswitched capacitors, C_C. Using larger capacitance values overloads the opamp more than is necessary during the phase ϕ_2, and their size is not critical to circuit performance. Reducing the capacitors too much causes common-mode offset voltages due to charge injection of the switches. Normally, all of the switches would be realized by minimum-size n-channel transistors only, except for the switches connected to the outputs, which might be realized by transmission gates (i.e., parallel n-channel and p-channel transistors both having minimum size) to accommodate a wider signal swing.

In applications where the opamp is being used to realize switched-capacitor circuits, switched-capacitor CMFB circuits are generally preferred over their continuous-time counterparts since they allow a larger output signal swing.

6.7 CURRENT-FEEDBACK OPAMPS

The last type of opamp described in this chapter is referred to as a *current-feedback opamp*. This type of opamp has become popular recently, particularly in applications

where high gain and high speed are required [Comlinear, 1985; Bowers, 1990]. One of this opamp's advantages is that its closed-loop gain can be changed, when used in a feedback application, without significantly affecting its loop gain. Because of this feature, a single compensation capacitor of one size can be used irrespective of the gain selected. As will be seen, the opamp is similar to many of the high-speed CMOS opamps described above, in that all nodes are low impedance except for the node where the compensation capacitor is located. Currently, the most popular technology used for realizing current-feedback opamps is a high-speed complementary-bipolar technology having vertical pnp transistors that have speeds comparable to vertical npn transistors.

A simplified example of a current-feedback opamp is shown in Fig. 6.23, where R_o is the impedance at the output of the current mirrors. The input signal, v_{in}, is applied to a high-impedance input, while a feedback current connects to a low-impedance node labelled v_n. The impedance seen looking into this low-impedance node is on the order of $1/(2g_{m1})$.

Ignoring stability issues, the basic operation of this current-feedback opamp can be understood as follows. The voltage at v_n is equal to the input voltage, v_{in}, due to the class-AB unity-gain buffer stage consisting of Q_1, Q_2, and the two diodes. Now assuming R_o is very large and ignoring C_C (which is included for compensation), a small feedback current, i_f, results in a large output voltage, v_{out}. Thus, during stable

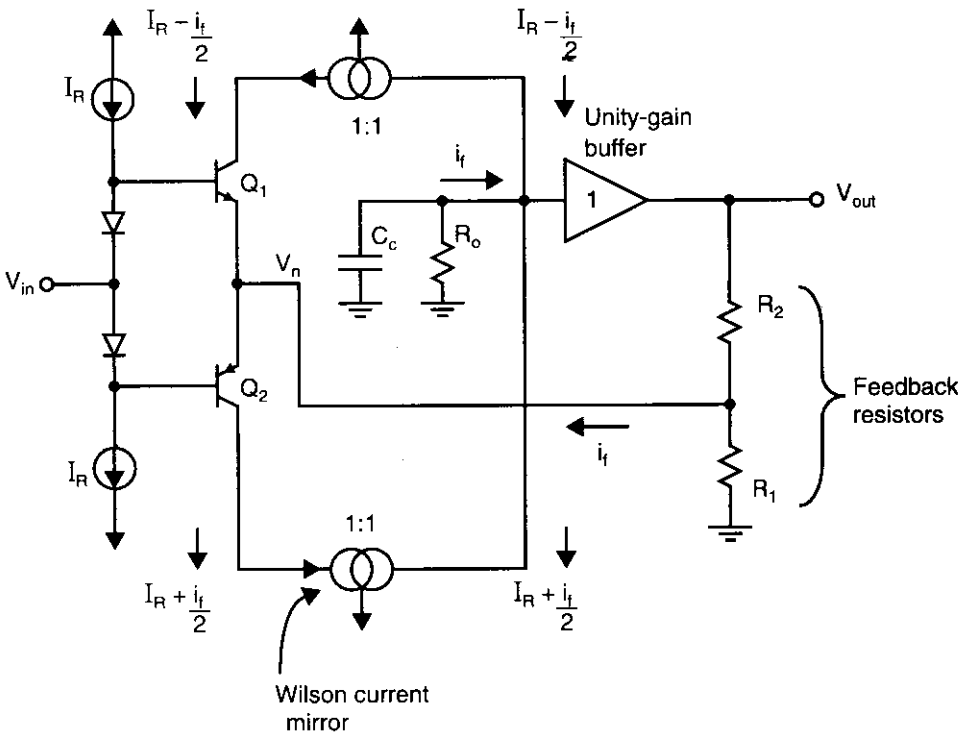


Fig. 6.23 A current-feedback opamp.

operation, $i_f \approx 0$ such that v_{out} goes to a finite voltage value. Finally, v_{out} is seen to be related to $v_n = v_{in}$ due to the two resistors, R_1 and R_2. Specifically, noting that $i_f = 0$, we have

$$\frac{v_{in}}{R_1} = \frac{v_{out} - v_{in}}{R_2} \tag{6.74}$$

which can be rearranged to give the following input/output relationship:

$$\frac{v_{out}}{v_{in}} = \frac{R_1 + R_2}{R_1} \tag{6.75}$$

To understand the feedback nature of the current-feedback opamp, we make some simplifying assumptions here. First, assume that the transistor current gain, β, is infinite. Also assume that the impedance looking into the emitters of Q_1 and Q_2, which is $1/(g_{m1} + g_{m2})$, is much smaller than $R_1 \parallel R_2$. Finally, assume the feedback current, i_f, splits equally between the two transistors Q_1 and Q_2. Although this equal current splitting assumption is unnecessary (since the two currents will add to i_f in any case), it does make the explanation simpler.

The signal being fed back in the opamp is the signal current i_f coming from the feedback network of R_1 and R_2 and going into the emitters of Q_1 and Q_2. Recalling that $v_n = v_{in}$, this current is given by

$$i_f = i_{R_2} - i_{R_1} = \frac{v_{out} - v_{in}}{R_2} - \frac{v_{in}}{R_1} = \frac{v_{out}}{R_2} - v_{in}\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{6.76}$$

Now, with Q_1 and Q_2 having infinite current gains, β, the collector currents of Q_1 and Q_2 are both equal to their bias currents, I_R, added to their fed-back signal currents, $i_f/2$. This current is mirrored by current mirrors whose outputs are summed at the high-impedance node. Note that this high-impedance node sees an impedance to ground equal to the compensation capacitor, C_C, in parallel with a high resistance, R_o. This resistance, R_o, is large, typically on the order of $(g_m r_o^2)/4$. The loop gain for this circuit can be found by breaking the loop at the top of R_2 and injecting a signal, v_x, at the top of R_2. By breaking the loop at the low-impedance output of the unity-gain buffer, there is only a small error in ignoring the loading effect of the feedback network. The loop gain is then equal to the ratio of V_{out} to v_x, while zeroing the input signal, v_{in}. The loop gain, $LG(s)$, is seen to be

$$LG(s) = \left(\frac{-I_f}{V_x}\right)\left(\frac{V_{out}}{I_f}\right) \tag{6.77}$$

The first term in (6.77) can be found using (6.76) for $v_{in} = 0$ and letting $V_x = V_{out}$. We have

$$\frac{I_f}{V_x} = \frac{1}{R_2} \tag{6.78}$$

assuming again that $V_x = V_{in}$. For the second term, note from Fig. 6.23 that the small-signal portions of the currents from the current mirrors sum together and both

flow out of the parallel combination of C_C and R_o, resulting in

$$V_{out} = \frac{-I_f}{sC_C + 1/R_o} \tag{6.79}$$

Substituting (6.78) and (6.79) into (6.77), we have

$$LG(s) = \frac{R_o/R_2}{1 + sC_CR_o} \cong \frac{1}{sC_CR_2} \tag{6.80}$$

where the approximation is valid at high frequencies. Thus, we see here that the high-frequency behavior of the loop gain is independent of R_1. As a result, one can change R_1 to realize various closed-loop gains without affecting the unity-gain frequency or the closed-loop stability. Using (6.80), we see the unity-gain frequency, ω_t, is given by

$$\omega_t \cong \frac{1}{C_CR_2} \tag{6.81}$$

To find the closed-loop transfer function, we can substitute (6.79) into (6.76) and simplify to obtain

$$A_{cl}(s) = \frac{V_{out}}{V_{in}} = \frac{1/R_1 + 1/R_2}{1/R_1 + 1/R_2 + sC_c} = \frac{R_o(R_1 + R_2)}{(R_o + R_2)R_1}\left(\frac{1}{1 + sC_C(R_o \parallel R_2)}\right) \tag{6.82}$$

For $R_o >> R_2$, which is typically a good assumption, we have

$$A_{cl}(s) \cong \frac{R_1 + R_2}{R_1}\left(\frac{1}{1 + sC_CR_2}\right) \tag{6.83}$$

Thus, the –3-dB frequency of the closed-loop transfer function is given by

$$\omega_{-3\ dB} = \frac{1}{C_CR_2} \tag{6.84}$$

which is the same as the unity-gain frequency of the open-loop response, and is independent of R_1 and the low-frequency closed-loop gain. Therefore C_C can be chosen once, for a given R_2, and will remain optimum to a first-order approximation, irrespective of R_1 and the closed-loop gain. This independence is one of the main benefits of using current-feedback amplifiers. Specifically, they can attain very high 3-dB frequency values for large closed-loop gains, or in other words, large gain-bandwidth products. For example, if a current-feedback amplifier has a closed-loop gain of unity and a –3-dB frequency of 100 MHz, it has a gain-bandwidth product of 100 MHz. Now, by simply reducing R_1, the same amplifier might have a closed-loop gain of 20, but its –3-dB frequency remains near 100 MHz, and thus its gain-bandwidth product is near 2 GHz! It should be noted here, though, that above 100 MHz the gain would most likely fall off quite a bit faster than –20 dB/decade, since the nondominant poles will be only somewhat greater than 100 MHz.

This compensation independence does not occur for the more common voltage-feedback amplifiers. With voltage-feedback amplifiers, as the closed-loop gain is

increased, the closed-loop –3-dB frequency is decreased by the same amount if the compensation capacitor remains the same. As a result, the gain-bandwidth product for a voltage-feedback amplifier remains constant for a given compensation capacitor. For example, if a voltage-feedback amplifier has a closed-loop gain of unity and a –3-dB frequency of 100 MHz, when the closed-loop gain is changed to 20, the –3-dB frequency lowers to 5 MHz. However, in this case, the gain above 5 MHz will fall off at –20 dB/decade until somewhat greater than 100 MHz (when the nondominant poles come into effect). In addition, it should be noted that for a given fixed closed-loop gain, voltage-feedback amplifiers can achieve similar gain-bandwidth products to their current-feedback counterparts if one is allowed to change their compensation capacitor. Specifically, for larger closed-loop gains, the feedback factor β, and thus loop gain, is decreased; as a result, a smaller compensation capacitor can be used to obtain the same phase margin.

In most practical realizations of current-feedback amplifiers, Darlington-pair transistors are used in the input stage to decrease the input bias currents required. Unfortunately, this approach makes the opamp somewhat noisier and increases the opamp input offset voltages somewhat. Another limitation is the requirement that R_1 be taken much larger than the impedances looking into the emitters of Q_1 and Q_2, namely $1/(g_{m1} + g_{m2})$. As a result, R_1, and therefore R_2 as well, need to be larger than would be desired for optimum high-frequency performance. This limitation is more severe in high-gain applications where $R_1 << R_2$. Additionally, these current-feedback opamps typically make use of a purely resistive feedback network and may become difficult to compensate if reactive components (such as a capacitor) are used in the feedback network. Regardless of these limitations, current-feedback opamps exhibit excellent high-frequency characteristics and are quite popular in many video and telecommunications applications.

6.8 SPICE SIMULATION EXAMPLES

In this section, simulation results for Example 6.2 and Example 6.3 are presented.

Simulation of Example 6.2

The circuit for Example 6.2 is shown in Fig. 6.24. First, the clamp transistors, Q_{12} and Q_{13}, are excluded. Bias voltages V_{B1} and V_{B2} are set such that all transistors are active, and an ac analysis is performed to determine the unity-gain frequency. Next, the opamp is placed in unity-gain configuration, and a step response is performed to determine the slew rate. Finally, the clamp transistors are included, and the resulting slew rate is determined.

NETLIST:

```
vdd     1     0     dc 2.5
vss     2     0     dc -2.5
```

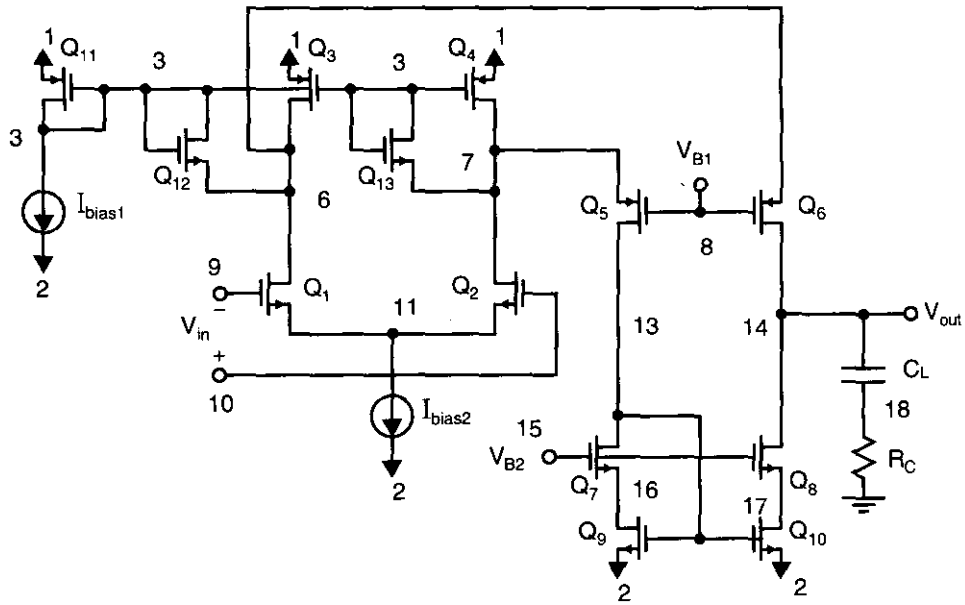


Fig. 6.24 A folded-cascode opamp.

m11	3	3	1	1	PMOS w = 10um l = 1.6um
ibias1	3	2			dc 6.6μ
m3	6	3	1	1	PMOS w = 300um l = 1.6um
m4	7	3	1	1	PMOS w = 300um l = 1.6um
m1	6	9	11	2	NMOS w = 300um l = 1.6um
m2	7	10	11	2	NMOS w = 300um l = 1.6um
ibias2	11	2			dc 320μ
m5	13	8	7	1	PMOS w = 60um l = 1.6um
m6	14	8	6	1	PMOS w = 60um l = 1.6um
m7	13	15	16	2	NMOS w = 20um l = 1.6um
m8	14	15	17	2	NMOS w = 20um l = 1.6um
m9	16	13	2	2	NMOS w = 20um l = 1.6um
m10	17	13	2	2	NMOS w = 20um l = 1.6um
Cl	14	18			10p
Rc	18	0			347
vb1	8	0			dc 0.75
vb2	15	0			dc –1
vin10	10	0			dc 0 ac 1
vin9	9	0			dc 0

The frequency plot of the folded-cascode opamp is shown in Fig. 6.25, where we find that the unity-gain frequency occurs at about 40 MHz. The step response of the folded-cascode opamp without clamp transistors is shown in Fig. 6.26(*a*), where we

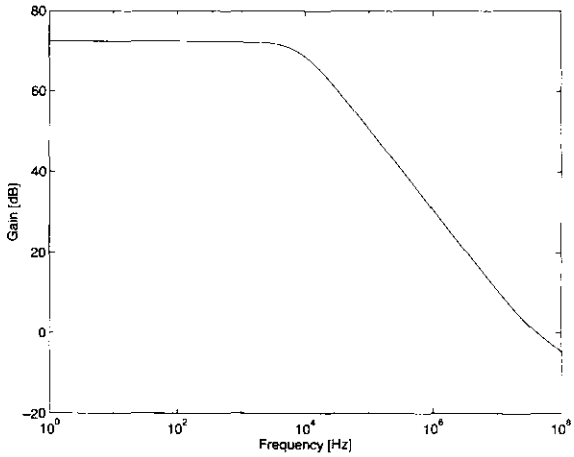
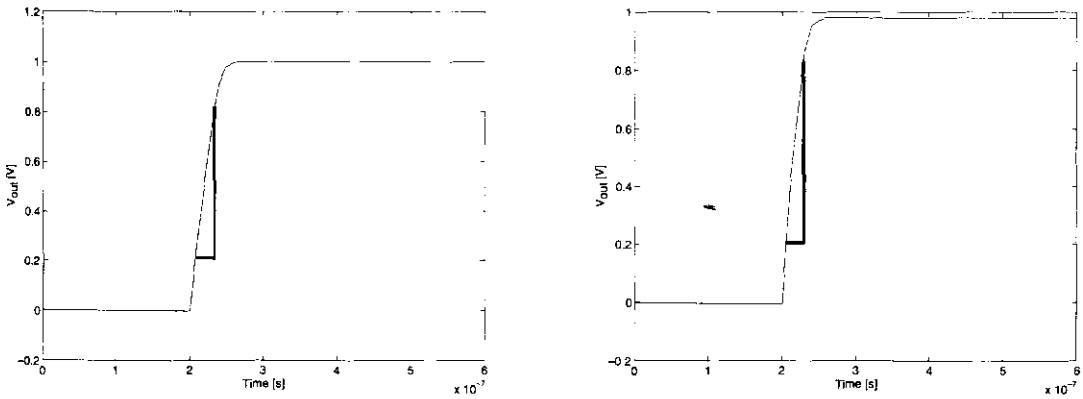


Fig. 6.25 Frequency plot of the folded-cascode opamp.



(*a*) (*b*)

Fig. 6.26 The step response of the folded-cascode opamp (*a*) without clamp transistors and (*b*) with clamp transistors.

find the slew rate is about 23 V/μs. The step response of the folded-cascode opamp with clamp transistors is shown in Fig. 6.26(*b*), where we find that the slew rate is now 27 V/μs. While this slew-rate improvement is not that substantial, it should be noted that the main advantage of including the clamp transistors is to help the opamp recover from slew-rate conditions more quickly.

Simulation of Example 6.3

The circuit for Example 6.3 is shown in Fig. 6.27. First, an ac analysis is performed to determine the unity-gain frequency. Then the opamp is placed in unity-gain

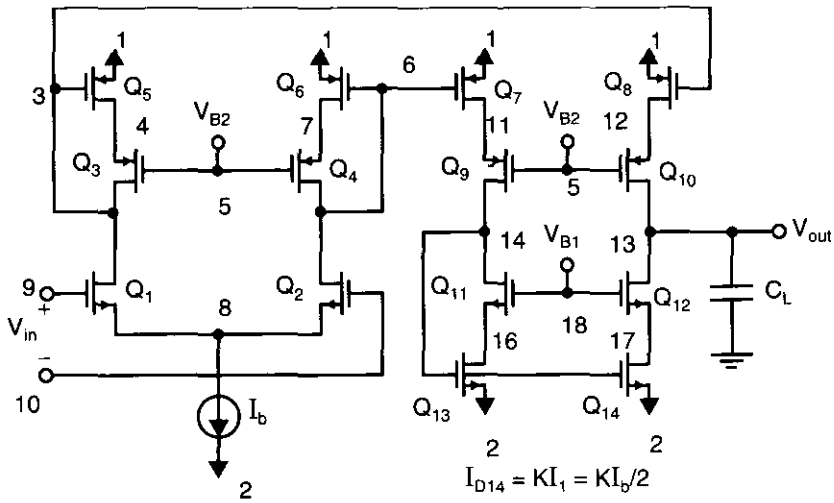


Fig. 6.27 A current-mirror opamp.

configuration, and a step response is performed to determine the slew rate.

NETLIST:

vdd	1	0			dc 2.5
vss	2	0			dc –2.5
m5	4	3	1	1	PMOS w = 60um l = 1.6um
m6	7	6	1	1	PMOS w = 60um l = 1.6um
m3	3	5	4	1	PMOS w = 60um l = 1.6um
m4	6	5	7	1	PMOS w = 60um l = 1.6um
m1	3	9	8	2	NMOS w = 300um l = 1.6um
m2	6	10	8	2	NMOS w = 300um l = 1.6um
ib	8	2			160um
m7	11	6	1	1	PMOS w = 60um l = 1.6um
m8	12	3	1	1	PMOS w = 120um l = 1.6um
m9	14	5	11	1	PMOS w = 60um l = 1.6um
m10	13	5	12	1	PMOS w = 120um l = 1.6um
m11	14	18	16	2	NMOS w = 30um l = 1.6um
m12	13	18	17	2	NMOS w = 60um l = 1.6um
m13	16	14	2	2	NMOS w = 30um l = 1.6um
m14	17	14	2	2	NMOS w = 60um l = 1.6um
cl	13	0			10p
vb2	5	0			dc 0.6
vb1	18	0			dc –1.24
vin1	9	0			dc 0 ac 1
vin2	10	0			dc 0

The frequency plot of the opamp is shown in Fig. 6.28(*a*), where it is seen that unity-gain frequency is about 50 MHz. The step response of the opamp is shown in Fig. 6.28(*b*), where it can be determined that the slew rate is about 30 V/μs.

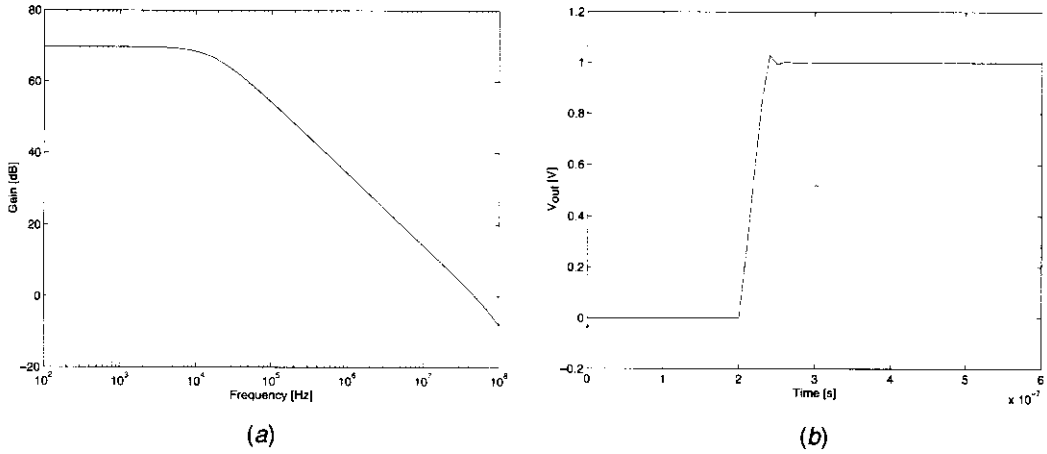


Fig. 6.28 Simulation results for a current-mirror opamp. (*a*) Frequency response. (*b*) Step response.

6.9 REFERENCES

J. N. Babanezhad, "A Rail-to-Rail CMOS Opamp," *IEEE J. of Solid-State Circuits,* Vol. 23, no. 6, pp. 1414–1417, December 1988.

J. N. Babanezhad and R. Gregorian, "A Programmable Gain/Loss Circuit," *IEEE J. of Solid-State Circuits,* Vol. 22, no. 6, pp. 1082–1090, December 1987.

M. Banu, J. M. Khoury, and Y. Tsividis, "Fully Differential Operational Amplifiers with Accurate Output Balancing," *IEEE J. of Solid-State Circuits,* Vol. 23, no. 6, pp. 1410–1414, December 1988.

D. F. Bowers, "Applying 'Current-Feedback' to Voltage Amplifiers," *Analog IC Design: The Current-Mode Approach.* C. Toumazou, F. J. Lidgey, and D. G. Haigh, eds., Peter Peregrinus, London, 1990.

K. Bult and G. J. G. M. Geelen, "A Fast-Settling CMOS Opamp for SC Circuits with 90-dB DC Gain," *IEEE J. of Solid-State Circuits,* Vol. 25, no. 6, pp. 1379–1384, December 1990.

R. Castello and P. R. Gray, "A High-Performance Micropower Switched-Capacitor Filter," *IEEE J. of Solid-State Circuits,* Vol. 20, no. 6, pp. 1122–1132, December 1985.

A. Coban and P. Allen, "A 1.75-V Rail-to-Rail CMOS Opamp," *Proceedings of the IEEE Int. Symp. on Circuits and Systems,* Vol. 5, pp. 5.497–5.500, London, June 1994.

Comlinear Corporation, "A New Approach to Opamp Design," *Comlinear Corporation Application Note* 300-1, March 1985.

J. F. Duque-Carillo, "Control of the Common-Mode Component in CMOS Continuous-Time Fully Differential Signal Processing," *Analog Integrated Circuits and Signal Processing, An International Journal,* Kluwer Academic Publishers, September 1993.

U. Gatti, F. Maloberti, and G. Torelli, "A Novel CMOS Linear Transconductance Cell for Continuous-Time Filters," proceedings of the *IEEE Int. Symp. on Circuits and Systems,* pp. 1173–1176, New Orleans, May 1990.

R. Hogervorst, et al., "CMOS Low-Voltage Operational Amplifiers with Constant-G_m Rail-to-Rail Input Stage," proceedings of the *IEEE Int. Symp. on Circuits and Systems,* pp. 2876–2879, San Diego, May 1992.

B. Hosticka, "Improvement of the Gain of MOS Amplifiers," *IEEE J. of Solid-State Circuits,* Vol. SC-14, no. 14, pp. 1111–1114, December 1979.

S. Law, Private conversation, Xerox Corp., 1983.

K. Martin, Class notes, UCLA, 1985.

K. Martin, Laboratory notes (independently derived, albeit after [Coban 1994]), 1994.

E. Säckinger and W. Guggenbühl, "A High-Swing, High-Impedance MOS Cascode Circuit," *IEEE J. of Solid-State Circuits*, Vol. 25, no. 1, pp. 289–298, February 1990.

D. Senderowicz, et al., "A Family of Differential NMOS Analog Circuits for a PCM Codec Filter Chip," *IEEE J. of Solid-State Circuits*, Vol. 17, no. 6, pp. 1014–1023, December 1982.

C. C. Shih and P. R. Gray, "Reference Refreshing Cyclic Analog-to-Digital and Digital-to-Analog Converters," *IEEE J. of Solid-State Circuits*, Vol. 21, no. 4, pp. 544–554, August 1986.

N. S. Sooch, "MOS Cascode Current Mirror," U.S. patent no. 4,550,284, October 1985.

R. A. Whatly, "Fully Differential Operational Amplifier with DC Common-Mode Feedback," U.S. patent no. 4573020, February 1986.

6.10 PROBLEMS

Unless otherwise stated, assume the following:

- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8\ V$
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}(\Omega) = 8{,}000L(\mu m)/I_D(mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9\ V$
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}(\Omega) = 12{,}000L(\mu m)/I_D(mA)$ in active region
  $C_j = 4.5 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.5 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

6.1 Calculate the output impedance of the two-transistor diode-connected circuit shown in Fig. P6.1 using small-signal analysis. Assume both transistors are in the active region, ignore the body effect, and assume $g_{m1} = g_{m2}$, $r_{ds1} = r_{ds2}$, and $g_m r_{ds} >> 1$.

6.2 Assume n is chosen to be 1 in Fig. 6.1 and all transistors are taken to be equal sizes, except $(W/L)_5$ is chosen so that $V_{DS3} = V_{DS2} = V_{eff3} + 0.15\ V$, where V_{eff} is chosen to be 0.2 V for all transistors except Q_5. Ignoring the body effect, find the required size for all transistors assuming they all have lengths equal to 1.6 μm. Assume that $I_{bias} = 50\ \mu A$.

6.3 What are the drain-source voltages of Q_3 and Q_4 from Problem 6.2 if the lengths of Q_3 and Q_4 are decreased to 1 μm to maximize speed?

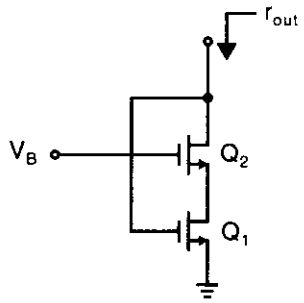


Fig. P6.1

6.4 What is the value required for R_B in Fig. 6.2 to give $V_{eff3} = 0.2$ V?

6.5 Assuming that μ_n is proportional to $T^{-3/2}$, what would V_{eff3} be at 100 °C if it was 0.2 V at 20 °C?

6.6 Assume that R_B from Fig. 6.2 has a temperature dependence of +0.3%/°C. What would V_{eff3} be at 100 °C if it was 0.2 V at 20 °C? Assume μ_n varies as in Problem 6.5.

6.7 Using small-signal analysis and ignoring the body effect, show that the circuit of Fig. 6.3 has an output impedance given by $r_{out} \cong g_{m1} r_{ds1} r_{ds2}(1 + A)$.

6.8 For the circuit shown in Fig. P6.8, using small-signal analysis, find the output impedance. What is an upper limit on the output impedance assuming $R_E >> r_\pi$? Compare this to the upper limit on the output impedance of a simple cascode mirror (i.e., $A = 0$) for the same case.

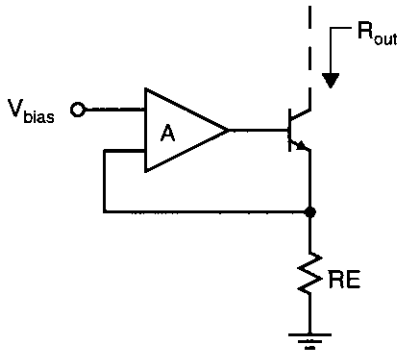


Fig. P6.8

6.9 For the circuit shown in Fig. 6.6 with the W/L ratios shown and assuming all lengths are 1.6 μm, estimate the output impedance. Assume that all current sources are ideal and that $I_{in} = 7I_{bias}$. Compare this to the output impedance of a wide-swing cascode current mirror where the gate of Q_1 was connected to $V_{casc\text{-}n}$ from the bias-generation circuitry of Fig. 6.2. Assume $I_{bias} = 50\ \mu A$.

6.10 For the folded-cascode amplifier shown in Fig. 6.9, with the transistor sizes given in Table 6.1, find the unity-gain frequency and the slew rate. Assume

±2 V power supplies are used, that C_L is 10 pF, and that the power dissipation is 1 mW.

6.11 For the folded-cascode amplifier shown in Fig. 6.9, with the transistor sizes given in Table 6.1, estimate the approximate frequency of the second pole caused by the parasitic capacitances at the drains of Q_1 and Q_2. Assume ±2 V power supplies are used, that C_L is 10 pF, and that the power dissipation is 1 mW. To simplify matters, junction capacitances can be ignored. What should C_L be to achieve 70° phase margin? What would the corresponding unity-gain frequency and slew rate be?

6.12 For the amplifier of Problem 6.11 with lead compensation, find the load capacitance, C_L, for a 70° phase margin. What size resistor should be used? What is the final unity-gain frequency and slew rate? Assume lead compensation with $\omega_z = 1.2\omega_t$ improves the initial unity-gain frequency by 20 percent and the phase margin by 30 degrees.

6.13 For the folded-cascode amplifier shown in Fig. 6.9, assume the bias currents of Q_1 and Q_2 are K times greater than the bias currents in Q_5 and Q_6. Assume the total bias current in the opamp, I_{total}, is determined by a specified power dissipation. Derive an equation for the unity-gain frequency in terms of K and I_{total}. Then show that the unity-gain frequency is maximized by taking K large.

6.14 Repeat Problem 6.13, but this time show that the dc gain is optimized by taking K large.

6.15 Derive an equation for the ratio of the unity-gain frequency of the folded-cascode amplifier of Fig. 6.9 to the unity-gain frequency of the current-mirror opamp of Fig. 6.11 in terms of K and I_{total} assuming both amplifiers have the same size input transistors, total power dissipation, and load capacitances. For the folded-cascode opamp, define K to be I_{D2}/I_{D7}. What is this ratio for K equal to 1, 2, and 4?

6.16 Consider the current-feedback opamp described in Example 6.13, in a feedback configuration shown in Fig. 6.12, with $C_1 = 1$ pF and $C_2 = C_C = C_{load} = 5$ pF. What is the linear settling time for 1 percent accuracy?

6.17 For the case described in Example 6.4, and assuming a current-mirror opamp described in Example 6.3, also assume the input voltage is a 1-V step change. Would the opamp slew rate limit? If so, how close must the output voltage come to its final value before its output voltage rate of change is less than the slew rate? How long would this take? How much longer would be required for linear settling to 1 percent of the total voltage change at the output?

6.18 What are the single-ended slew rates of the fully differential folded-cascode and current-mirror opamps, assuming $K = 2$ and the current densities are the same as those of Example 6.2 and Example 6.3, respectively? Calculate the slew rates in both the positive and the negative directions. Assume the load capacitances are 10 pF between each output and ground. For the folded-

cascode design, assume the clamp transistors Q_{11} and Q_{12} of Fig. 6.13 have not been included. Also, assume the current sources biasing the output stages do not change during transients.

6.19 Derive the slew rate of the fully differential opamp of Fig. 6.15 in terms of C_L, K, and I_{bias}.

6.20 Derive an equation for the unity-gain frequency of the fully differential opamp shown in Fig. 6.16 in terms of K and I_{total}.

6.21 Assume the CMFB circuit of Fig. 6.19 has ±2.5 V power supply voltages, and that the current sources require 0.5 V across them in order to have all transistors remain in the active regions. Ignoring the body effect, what V_{eff} bias voltage should be used for the p-channel transistors to maximize signal swing? What is the maximum single-ended signal swing before the gain of the common-mode feedback circuitry goes to zero and why?

6.22 Repeat Problem 6.21, but do not ignore the body effect. Assume the n-wells of the p-channel transistors are connected to V_{DD}.

6.23 Derive the low-frequency gain of the current-feedback amplifier of Fig. 6.23 without assuming that the impedance looking into the emitters of Q_1 and Q_2 is zero.

6.24 Consider the circuit shown in Fig. P6.24, where a high-output-impedance opamp is used. Capacitor C_p is the parasitic capacitance at the opamp input, while C_o is the output capacitance. Assuming linear settling, show that the optimal value for minimizing the time constant of this circuit is given by

$$C_{1,opt} = \sqrt{MC_oC_p}$$

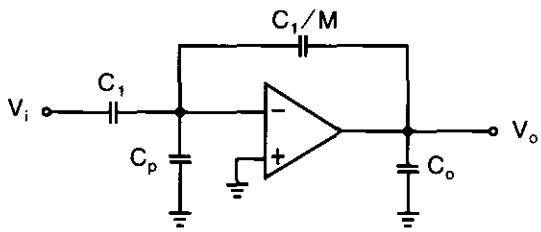


Fig. P6.24

6.25 Based on the results of Problem 6.24, derive τ_{min} for $M = 1$, $C_o = 1$ pF, and $C_p = 0.05$ pF. Sketch τ for other values of C_1 between 0.1 pF and 1 pF.

cascode design, assume the clamp transistors Q_{11} and Q_{12} of Fig. 6.13 have not been included. Also, assume the current sources biasing the output stages do not change during transients.

6.19 Derive the slew rate of the fully differential opamp of Fig. 6.15 in terms of C_L, K, and I_{bias}.

6.20 Derive an equation for the unity-gain frequency of the fully differential opamp shown in Fig. 6.16 in terms of K and I_{total}.

6.21 Assume the CMFB circuit of Fig. 6.19 has ±2.5 V power supply voltages, and that the current sources require 0.5 V across them in order to have all transistors remain in the active regions. Ignoring the body effect, what V_{eff} bias voltage should be used for the p-channel transistors to maximize signal swing? What is the maximum single-ended signal swing before the gain of the common-mode feedback circuitry goes to zero and why?

6.22 Repeat Problem 6.21, but do not ignore the body effect. Assume the n-wells of the p-channel transistors are connected to V_{DD}.

6.23 Derive the low-frequency gain of the current-feedback amplifier of Fig. 6.23 without assuming that the impedance looking into the emitters of Q_1 and Q_2 is zero.

6.24 Consider the circuit shown in Fig. P6.24, where a high-output-impedance opamp is used. Capacitor C_p is the parasitic capacitance at the opamp input, while C_o is the output capacitance. Assuming linear settling, show that the optimal value for minimizing the time constant of this circuit is given by

$$C_{1,opt} = \sqrt{MC_oC_p}$$

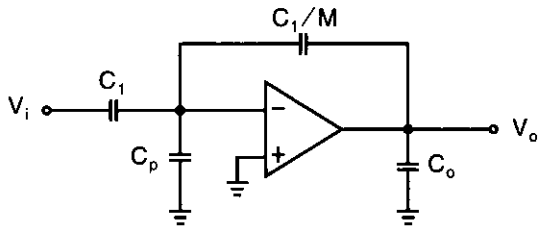


Fig. P6.24

6.25 Based on the results of Problem 6.24, derive τ_{min} for $M = 1$, $C_o = 1$ pF, and $C_p = 0.05$ pF. Sketch τ for other values of C_1 between 0.1 pF and 1 pF.

CHAPTER

7 Comparators

Perhaps the second most widely used electronic components (after amplifiers) are comparators. A comparator is used to detect whether a signal is greater or smaller than zero, or to compare the size of one signal to another. As we will see in Chapter 13, comparators are used in large abundance in A/D converters. They also find widespread use in many other applications, such as data transmission, switching power regulators, and others. In this chapter, we look at comparator design and practical limitations, where a number of different approaches are discussed. First, we examine a simplistic approach of using an open-loop opamp for a comparator. Although this approach is too slow for practical applications, it is a good example to use when we discuss several design principles for minimizing input-offset voltage and charge-injection errors. A number of other approaches are also described, such as multiple-stage comparators, positive-feedback track-and-latch comparators, and fully differential comparators. Finally, recent examples of both CMOS and bipolar comparator circuits are also described.

7.1 USING AN OPAMP FOR A COMPARATOR

A simplistic approach for realizing a comparator is to use an open-loop opamp, as shown in Fig. 7.1. The main drawback to this approach is the slow response time since the opamp output has to slew a large amount of output voltage and settles too slowly. However, temporarily ignoring this slow response time, we will first investigate another problem due to input-offset voltage.

The simplistic opamp approach shown in Fig. 7.1 has a resolution limited to the input-offset voltage of the opamp. This offset might be on the order of 2 mV to 5 mV for typical MOS processes, which is inadequate for many applications. An alternative architecture that can resolve signals with accuracies much less than the input-offset voltages of opamps is shown in Fig. 7.2 [McCreary, 1975; Yee, 1978]. Although this circuit has been used many times in early analog-to-digital converters, as we will see,

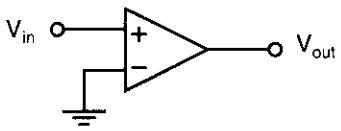


Fig. 7.1 A simplistic approach of using an open-loop opamp for a comparator.

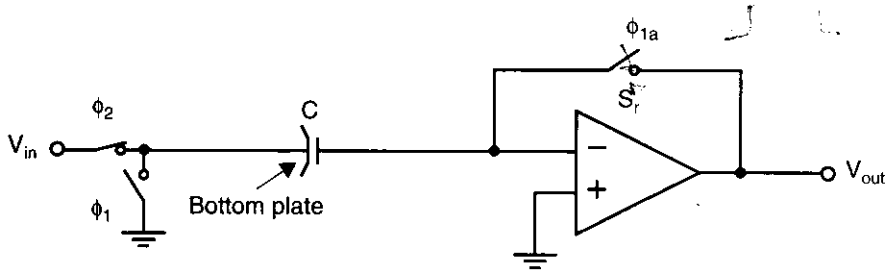


Fig. 7.2 Cancelling the offset voltage of a comparator—the comparator here must be stable, with unity-gain feedback during ϕ_{1a}. (ϕ_{1a} is a slightly advanced version of ϕ_1 so that charge-injection effects are reduced.)

it is not preferable nowadays. However, it is a simple example that can be used to illustrate many important design principles. The circuit of Fig. 7.2 operates as follows: During ϕ_1, known as the *reset phase,* the bottom plate[1] of the capacitor C (i.e., the left side of capacitor C) is connected to ground, and the top plate is connected to the inverting input of the opamp. At the same time, the output of the opamp is also connected to the inverting input of the opamp by closing switch S_r. Assuming the opamp is ideal, this connection causes the capacitor to be charged to zero volts. Next, during the *comparison phase*, the reset switch, S_r, is turned off, and the bottom plate of the capacitor is connected to the input voltage. The opamp is now in an open-loop configuration. If the input signal is greater than zero, the output of the opamp swings to a large negative voltage. If the input signal is less than zero, the output of the opamp swings to a large positive voltage. These two cases are easily resolved and the decision can be stored using a simple digital latch.

The limitations of this approach become apparent when one considers nonideal opamps, which have finite gains and require compensation to be stable during the reset phase.

EXAMPLE 7.1

Consider the case in which a 0.5-mV signal must be resolved using the circuit shown in Fig. 7.2, where the opamp's output should be 5 V. Assuming the opamp's unity-gain frequency is 10 MHz, find the maximum clocking rate of the comparator circuit if reset and comparison phases are equal and if six time constants are allowed for settling.

Solution

After the comparison phase, the output of the opamp should have a 5-V difference between the cases in which the input signal is either –0.25 mV or

1. The bottom plate of a capacitor has significant parasitic capacitance between it and ground. Therefore, it is almost always connected to the less sensitive input side rather than to the critical amplifier side (see Chapter 10).

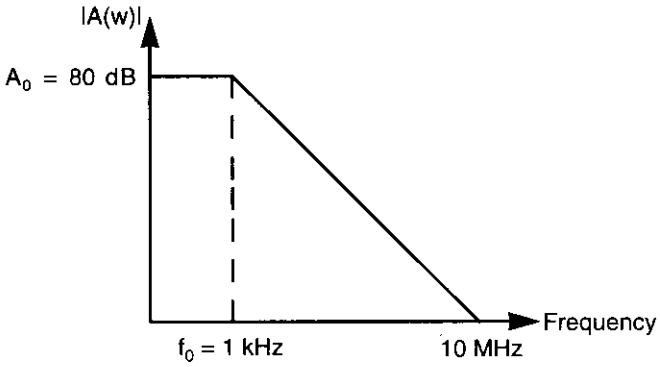


Fig. 7.3 The open-loop transfer function of the opamp used to realize the comparator.

+0.25 mV.[2] As a result, the opamp gain must be at least 10,000 V/V. By assuming that the dominant-pole compensation is used to guarantee stability during the reset phase, we obtain an open-loop transfer function for the opamp similar to that shown in Fig. 7.3. Here, the –3-dB frequency of the opamp is given by

$$f_{-3\ dB} = \frac{f_t}{A_0} = 1\ \text{kHz} \tag{7.1}$$

During the comparison phase, the output of the opamp will have a transient response similar to that of a first-order system that has a time constant given by

$$\tau = \frac{1}{2\pi f_{-3\ dB}} \cong 0.16\ \text{ms} \tag{7.2}$$

If six time constants are allowed for settling during the comparison phase, then approximately 1 ms is needed for the comparison phase. Assuming the reset time is the same as the comparison time, the clock frequency can be no greater than 500 Hz—a frequency that is much too slow for most applications.

One possibility for speeding up the comparison time is to disconnect the compensation capacitor during the comparison phase. For example, a simplified opamp schematic is shown in Fig. 7.4. In this opamp, transistor Q_1 is used to achieve lead compensation when it is on during the reset phase, ϕ_1. During the comparison phase, Q_1 is turned off, which disconnects the compensation capacitor, C_C, thereby greatly speeding up the opamp during this phase. Using this technique, it is possible to use

2. This assumes that ±2.5-V power supplies are being used and that digital output values are the same as the power-supply voltages, as is typical for CMOS. If a single power supply voltage of 5 V is used, then analog ground can be considered to be at 2.5 V.

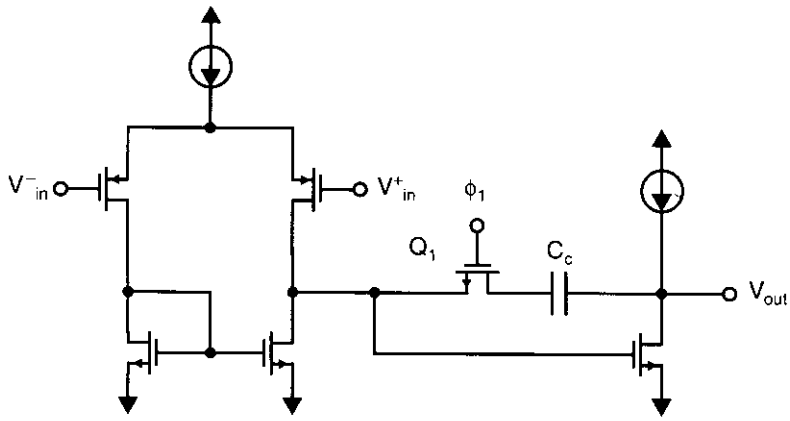


Fig. 7.4 An opamp that has its compensation capacitor disconnected during the comparison phase.

clock frequencies ten to fifty times greater than would otherwise be the case—perhaps as high as 25 or 50 kHz in our example. If this is adequate, then the approach is reasonable. Unfortunately, this often isn't adequate, and other approaches, to be described shortly, are necessary.

One superior aspect of the approach just described is that the input capacitor, C in Fig. 7.2, is never charged or discharged during operation. Specifically, in phase ϕ_1, C is always charged to 0 V. In phase ϕ_2, the top plate, connected to the inverting input of the opamp, is open-circuited at that time (assuming the parasitic capacitors are ignored), and the voltage across capacitor C remains at 0 V (in other words, the top-plate voltage follows V_{in}). This approach greatly minimizes the charge required from the input when V_{in} changes. If the switches attached to the bottom plate had their phases interchanged, then the comparison operation would be noninverting, but now capacitor C must be charged or discharged during the reset phase since the bottom plate follows V_{in}, whereas the top plate remains at virtual ground. Normally, we like to use a reasonably large input capacitor to minimize clock-feedthrough effects (described in Section 7.2). This charging/discharging requirement puts severe constraints on the input signal source, and, thus, the clock phasing shown in Fig. 7.2 should be used. Besides, it is usually possible to tolerate an inverting comparison since it can be made noninverting through the use of a digital inversion somewhere else in the system.

Input-Offset Voltage Errors

One source of error is due to the input-offset voltage of the opamp. This input offset might be caused by device mismatches or might be inherent in the design of the comparator. In switched-capacitor comparators, such as that shown in Fig. 7.2, input offset is not a problem since it is stored across the capacitor during the reset phase, and then the error is cancelled during the comparison phase. To appreciate this cancella-

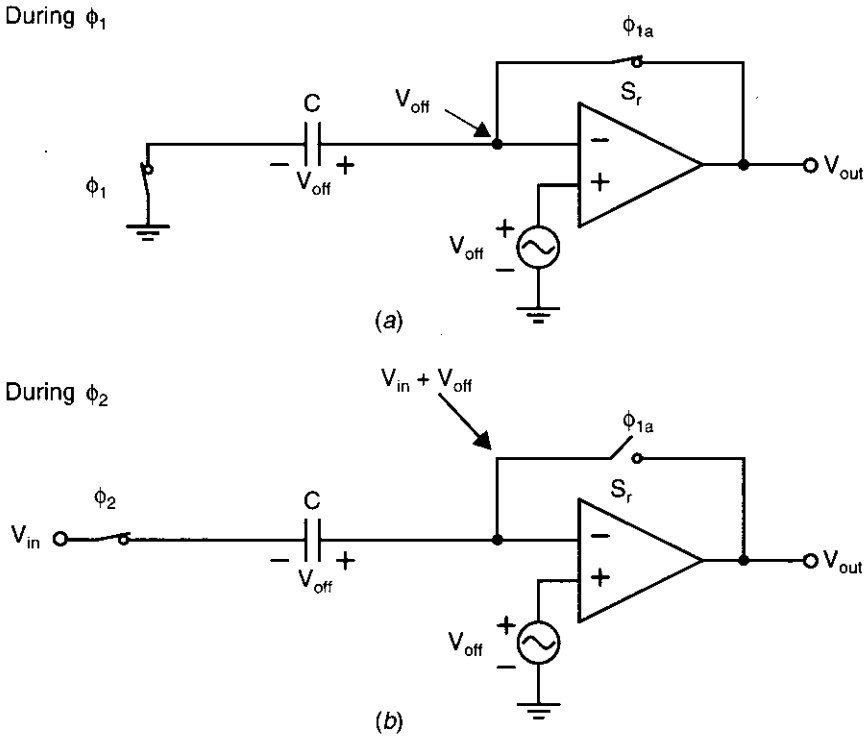


Fig. 7.5 The circuit configuration (*a*) during the reset phase, and (*b*) during the comparison phase, assuming the opamp has an input-offset voltage given by V_{off}.

tion, assume the opamp has an input-offset voltage error that is modelled as a voltage source in series with one of the opamp's inputs. The circuit configuration during the reset phase is shown in Fig. 7.5(*a*). Assuming the opamp gain is very large, then the inverting input of the opamp is at the voltage V_{off}, which implies that the input capacitor is charged to V_{off} during this phase. Next, during ϕ_2, as shown in Fig. 7.5(*b*), the left side of capacitor C is connected to the input voltage. The right side of the capacitor has a voltage given by $V_{in} + V_{off}$, which results in the comparator output becoming negative if V_{in} is greater than zero or positive if V_{in} is less than zero, regardless of the value of V_{off}. Not only does this technique eliminate input-offset voltage errors, but it also minimizes errors caused by low-frequency 1/f noise, which can be large in CMOS microcircuits.

7.2 CHARGE-INJECTION ERRORS

Perhaps the major limitation on the resolution of comparators is due to what is referred to as *charge injection,* also commonly called *clock feedthrough.* This error is due to unwanted charges being injected into the circuit when the transistors turn off. For the comparator in Fig. 7.2, the switches are normally realized by either n-channel transistors alone or CMOS transmission gates (which consist of n-channel transistors

in parallel with p-channel transistors, both of which must turn off). When MOS switches are on, they operate in the triode region and have zero volts between their drain and their source. When MOS switches turn off, charge errors occur by two mechanisms. The first is due to the channel charge, which must flow out from the channel region of the transistor to the drain and the source junctions.[3] The channel charge of a transistor that has zero V_{DS} is given by

$$Q_{CH} = WLC_{ox}V_{eff} = WLC_{ox}(V_{GS} - V_t) \tag{7.3}$$

This charge often dominates. The second charge (typically smaller, unless V_{eff} is very small) is due to the overlap capacitance between the gate and the junctions.

Figure 7.6 shows the comparator of Fig. 7.2, where the switches have been realized as n-channel transistors. Also shown in Fig. 7.6 are the parasitic capacitances due to gate-drain and gate-source overlap capacitors. Finally, although it is not shown in Fig. 7.6, one should remember to include the channel charge dispersion of any transistors that change from being on to off.[4] In the circuit shown, all transistors are n-channel, implying that the channel charge is negative.

Consider first when Q_3 turns off. If the clock waveform is very fast, the channel charge due to Q_3 will flow equally out through both junctions [Shieh, 1987]. The $Q_{ch}/2$ charge that goes to the output node of the opamp will have very little effect other than causing a temporary glitch. However, the $Q_{ch}/2$ charge that goes to the inverting-input node of the opamp will cause the voltage across C to change, which introduces an error. Since this charge is negative, for an n-channel transistor, the node

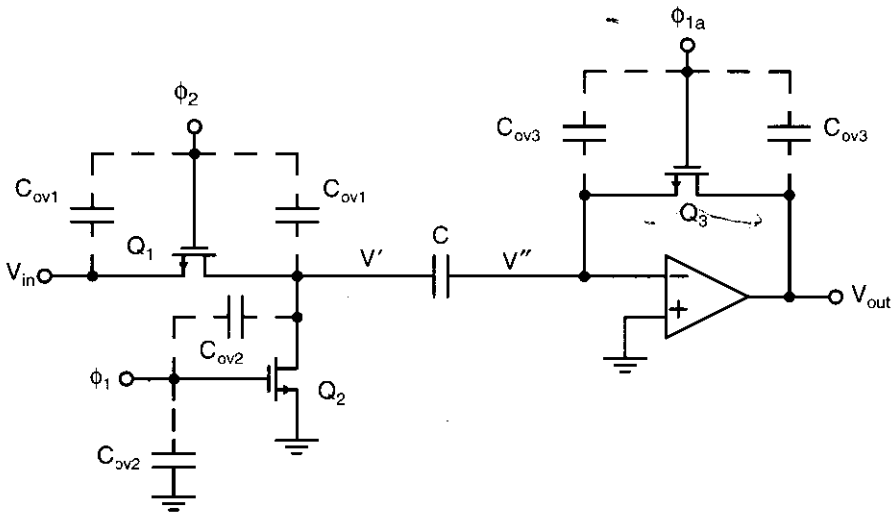


Fig. 7.6 The comparator in Fig. 7.2, with n-channel switches and overlap parasitic capacitances shown. Transistors turned on also have channel charge.

3. For a transistor in the triode region that has zero V_{DS}, it doesn't matter which junctions are called the drain and the source, since both junctions are at the same potential.
4. Transistors also accumulate channel charge when turning on, but this does not typically affect circuit performance since the charge comes from low-impedance nodes.

voltage V'' will become negative. The voltage change due to the channel charge is given by

$$\Delta V'' = \frac{(Q_{ch}/2)}{C} = -\frac{V_{eff3}C_{ox}W_3L_3}{2C} = -\frac{(V_{DD}-V_{tn})C_{ox}W_3L_3}{2C} \tag{7.4}$$

since the effective gate-source voltage of Q_3 is given by $V_{eff3} = V_{GS3} - V_{tn} = V_{DD} - V_{tn}$. The preceding voltage change in V'' is based on the assumption that Q_2 turns off slightly after Q_3 turns off. More will be said about this assumption shortly.

To calculate the change in voltage due to the overlap capacitance, it is first necessary to introduce the capacitor-divider formula. This formula is used to calculate the voltage change at the internal node of two series capacitors, when the voltage at one of the end terminals changes. This situation is shown in Fig. 7.7, where it is assumed that V_{in} is changing and we want to calculate the change in $V_{out} = V_{C2}$. The series combination of C_1 and C_2 is equal to a single capacitor, C_{eq}, given by

$$C_{eq} = \frac{C_1C_2}{C_1 + C_2} \tag{7.5}$$

When V_{in} changes, the charge flow into this equivalent capacitor is given by

$$\Delta Q_{eq} = \Delta V_{in}C_{eq} = \Delta V_{in}\frac{C_1C_2}{C_1 + C_2} \tag{7.6}$$

All of the charge that flows into C_{eq} is equal to the charge flow into C_1, which is also equal to the charge flow into C_2. Thus, we have

$$\Delta V_{out} = \Delta V_{C2} = \frac{\Delta Q_{C2}}{C_2} = \frac{\Delta V_{in}C_1}{C_1 + C_2} \tag{7.7}$$

This formula is often useful when calculating charge flow in integrated circuits. It can be applied to the circuit of Fig. 7.6 to calculate the change in V'' due to the overlap capacitance of Q_3 when it turns off. For this case, we have $C_1 = C_{ov}$, $C_2 = C$, and $\Delta V_{in} = -(V_{DD} - V_{SS})$. This assumes the clock signals change from V_{DD} to V_{SS}. The change in V'' due to the overlap capacitance is now found to be given by

$$\Delta V'' = \frac{-(V_{DD} - V_{SS})C_{ov}}{C_{ov} + C} \tag{7.8}$$

This change is normally less than that due to the change caused by the channel charge since C_{ov} is small.

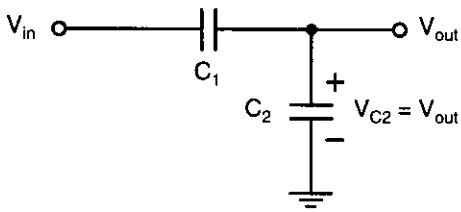


Fig. 7.7 A capacitor divider.

EXAMPLE 7.2

Assume that a resolution of ±2.5 mV is required from the circuit of Fig. 7.6. The following values are given: $C = 1$ pF, $C_{ox} = 1.92$ fF/(μm)^2, $(W/L)_3 = 10\ \mu m/0.8\ \mu m$, $L_{ov} = 0.1\ \mu m$, $V_{tn} = 0.8$ V, $V_{DD} = 2.5$ V, and $V_{SS} = -2.5$ V. What is the change in $\Delta V''$ when Q_3 turns off?

Solution

Using (7.4), we have $\Delta V'' = -13.1$ mV. The overlap capacitance is given by $C_{ov} = W_3 L_{ov} C_{ox} = 1.92$ fF. Using (7.8), this gives a voltage change in $\Delta V''$ of $\Delta V'' = -9.6$ mV, which is smaller than the change due to the channel charge, but not insignificant. The total change is found by adding the two effects so that $\Delta V'' = -22.7$ mV.

This result is on the order of 10 times greater than the value of ±2.5 mV that should be resolved. Clearly, additional measures should be taken to minimize the effects of charge injection if this resolution is required.

Making Charge-Injection Signal Independent

The charge injection due to transistors Q_1 and Q_2 may cause temporary glitches, but it will have much less effect than the charge-injection due to Q_3 if we assume that Q_2 turns off slightly after Q_3 does [Haigh, 1983]. This time difference is the reason why the clock voltage of Q_2 is denoted ϕ_1, whereas the clock voltage of Q_3 is denoted ϕ_{1a} (ϕ_1 advanced). The argument for this arrangement is slightly complicated, but goes as follows: When Q_2 turns off, its charge injection causes a negative glitch at V', but this will not cause any change in the charge stored in C since the right side of C is connected to an effective open circuit, assuming that Q_3 has already turned off. Later, when Q_1 turns on, the voltage V' will settle to V_{in} regardless of the previous charge injection of Q_2. Thus, the voltage at the left of C is unaffected by the charge injection of Q_2, and the voltage across C is also unaffected by the charge injection of Q_2. Therefore, the voltage V'' is unaffected by the charge injection of Q_2. This is not the case if Q_2 is turned off at the same time or before Q_3 is turned off. The charge injection of Q_1 has no effect for a simpler reason—the comparison has already been made when it turns off. In addition, its charge injection has no effect when it turns on because the right side of C is connected to an open circuit at that time, assuming the clocks do not overlap. *In summary, by turning off ϕ_{1a} first, the circuit is affected only by the charge injection of Q_3 and not by any of the other switches.*

Fig. 7.8 shows a simple digital circuit, originally developed by one of the authors [Martin, 1980], that is capable of generating the desired clock waveforms. The waveforms ϕ_1 and ϕ_{1a} do not overlap with ϕ_2 and ϕ_{2a}. Also, ϕ_{1a} will be advanced slightly (by two inverter delays), compared to ϕ_1.

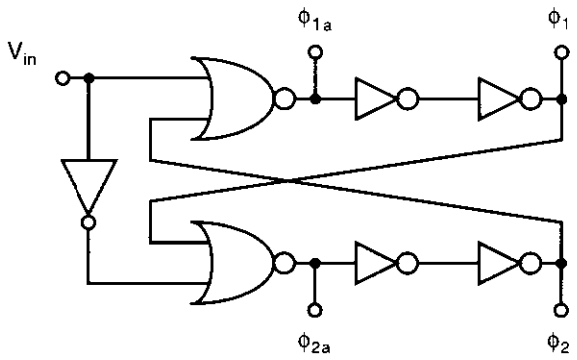


Fig. 7.8 A clock generator suitable for generating the desired clock waveforms for the comparator in Fig. 7.2.

Minimizing Errors Due to Charge Injection

The simplest way to reduce errors due to charge injection is to use larger capacitors. For our previous example, capacitors on the order of about 100 pF guarantee that the clock-feedthrough errors are less than 0.5 mV. Unfortunately, this large amount of capacitance would require a large amount of silicon area. Also, integrated capacitors have parasitic capacitances between the bottom plate and the substrate that might be about 20 percent of the size of the realized capacitor. Thus, a bottom plate capacitance of about 20 pF would have to be driven by the input circuits, which would greatly slow down the circuits.

An alternative approach for minimizing errors due to charge injection is to use fully differential design techniques for comparators, similar to what is often done for opamps. A simple example of a one-stage, switched-capacitor, fully differential comparator is shown in Fig. 7.9. Ideally, when the comparator is taken out of reset mode,

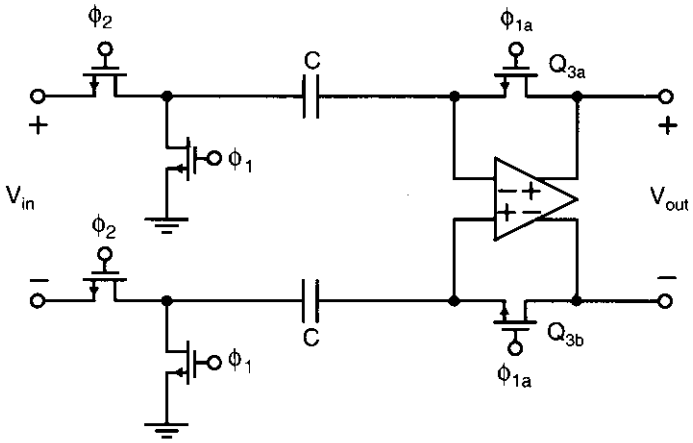


Fig. 7.9 A fully differential, single-stage, switched-capacitor comparator.

the clock feedthrough of reset switch Q_{3a} matches the clock feedthrough of Q_{3b}. In this case, the common-mode voltage may be slightly affected, but the differential input voltage is unaffected. The only errors now are due to mismatches in the clock feedthrough of the two switches, which will typically be at least ten times smaller than in the single-ended case. For this reason, virtually all modern integrated comparators utilize fully differential design techniques.

A third alternative that can be used along with fully differential design techniques is to realize a multi-stage comparator [Poujois, 1978; Vittoz, 1985], where the clock feedthrough of the first stage is stored on coupling capacitors between the first and second stage, thereby eliminating its effect. The clock feedthrough of the second stage can be stored on the coupling capacitors between the second and the third stages, and so on. Although this technique is almost always used along with fully differential configurations, we will explain the technique using single-ended configurations, for simplicity. Consider the three-stage comparator shown in Fig. 7.10(*a*), along with the appropriate clock waveforms shown in Fig. 7.10(*b*). Consider the time when ϕ_1' drops and the switch of the first stage has charge injection. Figure 7.11 illustrates this case, along with the parasitic capacitances of the first-stage switch. When ϕ_1' drops, Q_1 injects charge into both the inverting input and the output of the first stage through parasitic capacitors C_{p1} and C_{p2}, respectively. The charge injected at the first stage

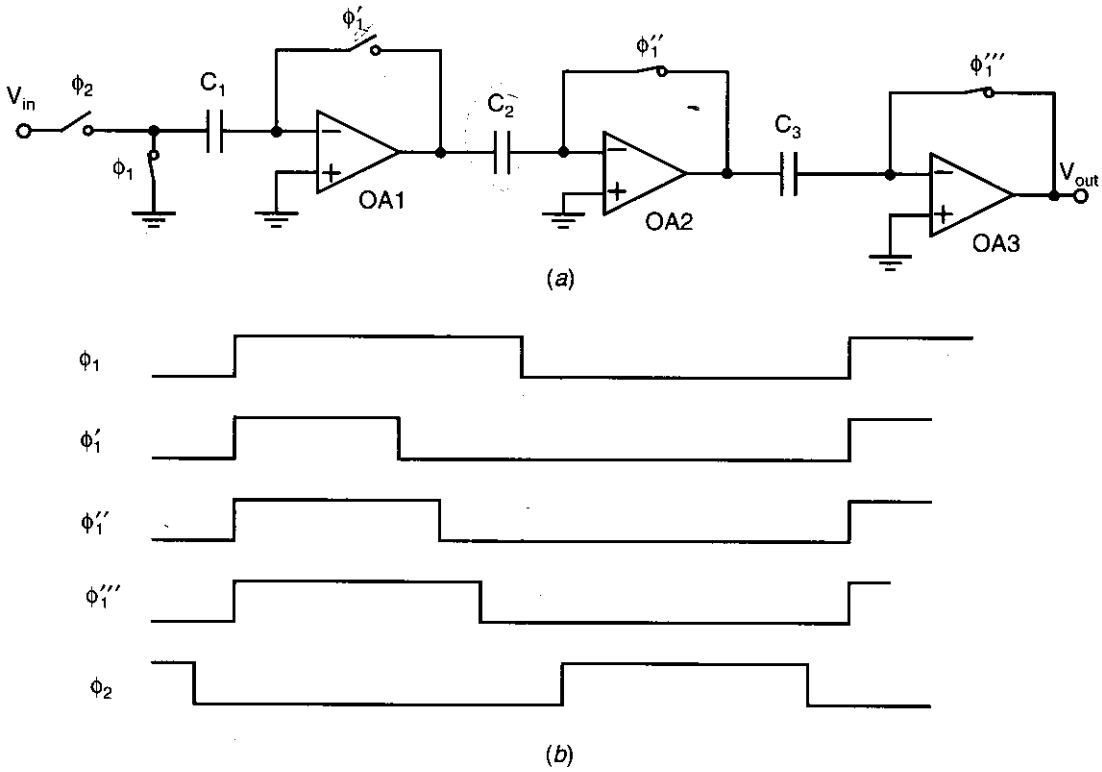


Fig. 7.10 (*a*) A multi-stage comparator used to eliminate clock feedthrough errors, with (*b*) the appropriate clock waveforms.

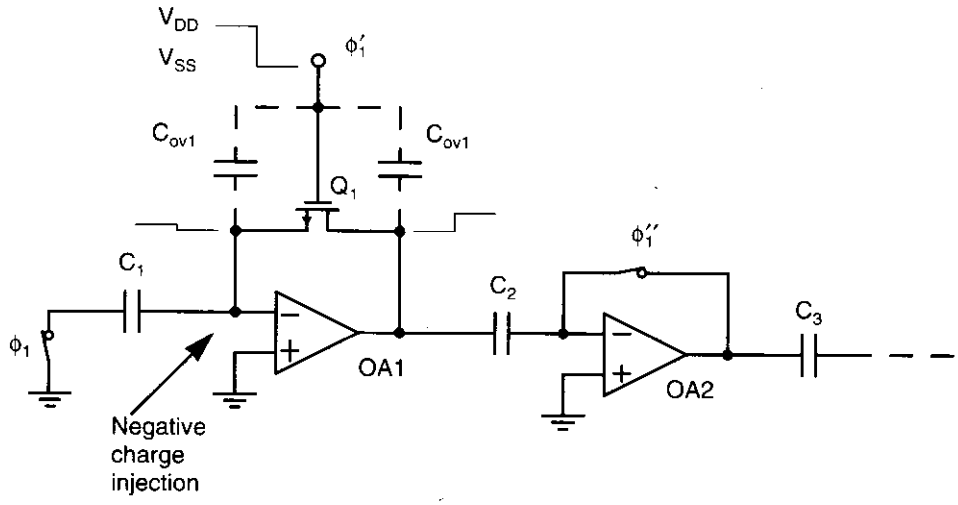


Fig. 7.11 The first stage of the comparator in Fig. 7.10, when the first stage is injecting charge.

output causes only a temporary glitch there. The charge injected at the inverting input causes this node to become negative. The amount by which this node voltage becomes negative is calculated using the analysis method described previously. In any practical case, this amount will be in the range of tens of millivolts, assuming that coupling capacitors of around 1 pF are used. After the inverting input becomes negative, the output of the first stage becomes positive by an amount equal to the negative transition of the inverting input multiplied by the first stage's gain. However, at this time, ϕ_1'' is still high. Therefore, the second stage is still being reset and C_2 is charged up to the output error caused by the clock feedthrough of the first stage, thereby eliminating its effect. In a similar manner, when ϕ_1'' turns off and the second stage goes from closed-loop reset mode to open-loop comparison mode, the third stage is still in reset mode and the clock feedthrough of the second stage is stored on coupling capacitor C_3. Finally, when the third stage turns off, its charge injection is not cancelled. However, the error it causes in resolving an input voltage to all three stages is equal to the voltage transition at the inverting input of the third stage divided by the negative of the gains of the first two stages (i.e., this is the input voltage needed to cancel the effect of the clock feedthrough of the third stage coming out of reset mode).

In a variation on this multi-stage approach [Poujois, 1978], the first stage was not reset. However, the input-offset voltage of the first stage was still cancelled by storing it on coupling capacitors between the first and second stage during a reset mode. This approach is described Section 7.4.

EXAMPLE 7.3

Assume all capacitors are 1 pF, transistor sizes are $W/L = 10\ \mu m/0.8\ \mu m$, $C_{ox} = 1.92\ fF/(\mu m)^2$, $L_{ov} = 0.1\ \mu m$, $V_{tn} = 0.8\ V$, $V_{DD} = 2.5\ V$, and

$V_{SS} = -2.5$ V, and each stage has a gain of 20. What is the input-offset voltage error caused by the clock feedthrough of the third stage when the third stage comes out of reset?

Solution

The values used are identical to those used in Example 7.2. Therefore, the charge injection at the inverting input of the third stage is the same as that found in the solution to Example 7.2, or −22.7 mV. The input signal that can overcome this 22.7 mV value is given by

$$\Delta V_{in} = \frac{22.7 \text{ mV}}{A_1 A_2} = 57 \ \mu\text{V} \tag{7.9}$$

Thus, the equivalent input-offset voltage is 57 μV, which is much less than the resolution found in Example 7.2 of 22.7 mV.

Speed of Multi-Stage Comparators

The approach just described can be used to realize very-high-resolution comparators, especially when it is combined with fully differential circuit design techniques. However, this approach does have the limitation that it requires multiple-phase clock waveforms, which slow down the circuits.

Although the multi-stage approach is limited in speed due to the need for the signal to propagate through all the stages, it can still be reasonably fast because each of the individual stages can be made to operate fast. Typically, each stage consists of a single-stage amplifier that has only a 90° phase shift and therefore does not need compensation capacitors (i.e., each stage has a 90° phase margin without compensation). To see why this multi-stage approach can be fast, consider the simplified case of a cascade of first-order, uncompensated inverters, as shown in Fig. 7.12. This circuit is a very rough approximation of a multi-stage comparator when it is not being reset. The parasitic load capacitance at the output of the ith stage is approximately given by

$$C_{pi} \cong C_{0\text{-}i} + C_{gs\text{-}i+1} \tag{7.10}$$

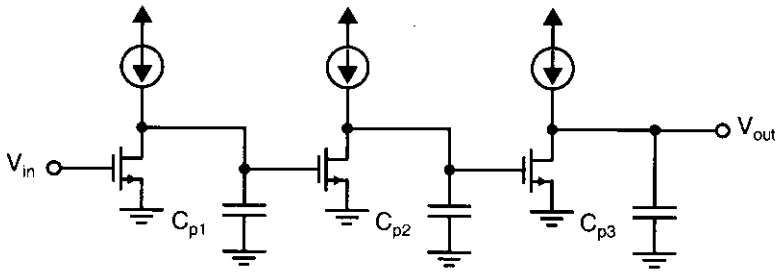


Fig. 7.12 Realizing a comparator by using a cascade of first-order gain stages.

where $C_{0\text{-}i}$ is the output capacitance of the ith stage, which would normally be due to junction capacitance, and $C_{gs\text{-}i+1}$ is the gate-source capacitance of the input transistor of the succeeding stage. Equation (7.10) will not necessarily be true for the last stage, but the load of the last stage should be of the same magnitude. If one assumes the stages are matched, then normally $C_{0\text{-}i} < C_{gs\text{-}i+1}$ since junction capacitances are usually less than gate-source capacitances, implying that one can usually assume

$$C_{pi} < 2C_{gs\text{-}i} \tag{7.11}$$

The unity-gain frequency of a single stage is then on the order of

$$\omega_{ti} \sim \frac{g_{mi}}{2C_{gs\text{-}i}} \tag{7.12}$$

or larger, where g_{mi} is the transconductance of the input capacitor of the ith stage. Thus, the unity-gain frequency of a single gain-stage is on the order of one-half the unity-gain frequency of a single transistor. If one assumes the ith stage has a dc gain $A_{0\text{-}i}$ and that the ith stage is well described by a first-order transfer function, then the transfer function of a single stage is approximately given by

$$A_i(s) = \frac{A_{0\text{-}i}}{1 + s/\omega_{p\text{-}i}} \tag{7.13}$$

where

$$\omega_{p\text{-}i} \cong \frac{\omega_{t\text{-}i}}{A_{0\text{-}i}} \sim \frac{g_{mi}}{2A_{0\text{-}i}C_{gs\text{-}i}} \tag{7.14}$$

Thus, the –3-dB frequency of a single stage is on the order of one-half the unity-gain frequency of a transistor divided by the gain of the stage. If one has a cascade of n stages, then the overall transfer function is given by

$$A_{total}(s) = \prod A_i(s) \tag{7.15}$$

This result can be approximated by a first-order transfer function, where all higher-order terms are ignored, resulting in

$$A_{total}(s) \cong \frac{\prod A_{0\text{-}i}}{1 + s\sum 1/\omega_{p\text{-}i}} \cong \frac{A_0^n}{1 + ns/\omega_{p\text{-}i}} \tag{7.16}$$

Thus, a cascade of n stages has a time constant approximately given by

$$\tau_{total} \cong \frac{2nA_0C_{gs}}{g_m} \cong 2nA_0\tau_T \tag{7.17}$$

where $\tau_T = C_{gs}/g_m$ is the approximate transit time of a single transistor (i.e., 1 over the unity-gain frequency of a single transistor). In other words, the time constant of the cascade of first-order stages is approximately equal to n times the time constant of a single stage, which is roughly given by $2A_0$ times the transit time of a single transistor. This result should be compared to using a single opamp, which, for the same overall gain, will have a time constant considerably greater than $2A_0^n$ times the transit

time of a single transistor. Although the preceding speed estimate is valid for simple single-ended stages only, the general principles apply to more complicated, fully differential stages as long as each stage is first order and doesn't require compensation.

Equation (7.17) can be simplified further. Recall from Chapter 2 that

$$C_{gs} = \frac{2}{3}C_{ox}WL \tag{7.18}$$

for a transistor in the active region, and that

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D} \tag{7.19}$$

We can substitute (7.18) and (7.19) into (7.17), and, after simple manipulation, arrive at

$$\tau_{total} \cong \frac{2nA_0C_{gs}}{g_m} \cong \frac{4nA_0L^2}{3\mu_n V_{eff}} \tag{7.20}$$

where, again from Chapter 2,

$$V_{eff} = \sqrt{\frac{2I_D}{\mu_n C_{ox}(W/L)}} \tag{7.21}$$

Equation (7.20) is very useful in quickly accessing the speed and resolution capabilities of a given technology. It also gives some insight into designing comparators. First, the effective gate-source voltages of input drivers of each stage should be as large as possible. Second, the widths of the transistors are relatively unimportant, assuming they are large enough so that C_{gs} dominates parasitic capacitances due to interconnect and external loading.

EXAMPLE 7.4

Assume for a 0.8-μm technology that $A_0 = 20$, $n = 3$, $V_{eff} = 0.5$ V, and $\mu_n = 0.05\ M^2/V \cdot s$. What is the maximum clocking frequency of the comparator?

Solution

Using (7.20), we have $\tau_{total} = 4.1$ ns. If we assume that 3τ is required for each half period, then a complete period requires 6τ, which corresponds to a clocking frequency of 40.1 MHz. The resolution is on the order of $5\ V/A_0^n = 0.625$ mV. This corresponds fairly closely to state-of-the-art realized comparators in a 0.8-μm technology.

7.3 LATCHED COMPARATORS

Modern high-speed comparators typically have one or two stages of preamplification followed by a *track-and-latch stage*, as Fig. 7.13 [Yukawa, 1985] shows in simplified

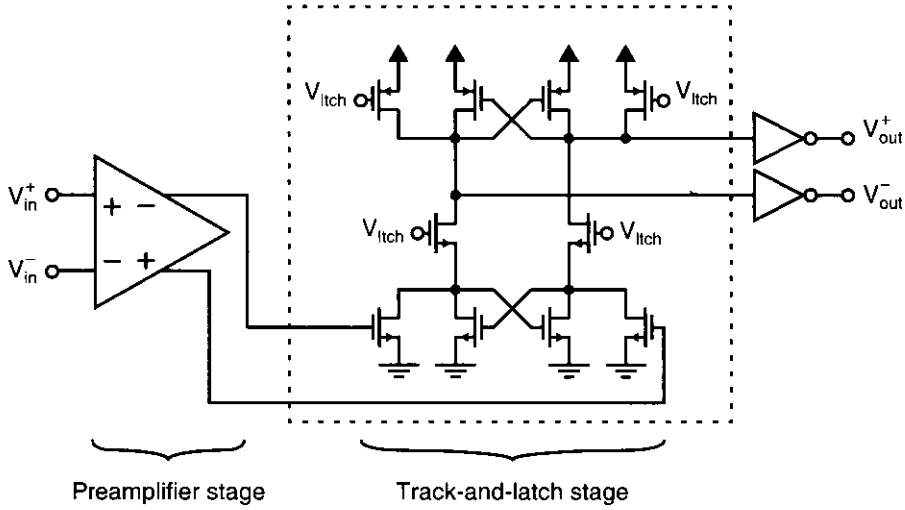


Fig. 7.13 A typical architecture for a high-speed comparator.

form for a CMOS implementation. The rationale behind this architecture is as follows: The preamplifier or preamplifiers are used to obtain higher resolution and to minimize the effects of kickback (which will be explained shortly). The output of the preamplifier, although larger than the comparator input, is still much smaller than the voltage levels needed to drive digital circuitry. The track-and-latch stage then amplifies this signal further during the track phase, and then amplifies it again during the latch-phase, when positive feedback is enabled. The positive feedback regenerates the analog signal into a full-scale digital signal. The track-and-latch stage minimizes the total number of gain stages required, even when good resolution is needed, and, thus, is faster than the multi-stage approach just described.

The preamplifier typically has some gain, perhaps 4 to 10, although sometimes it may simply be a unity-gain buffer, if very high speed but only moderate resolution is required. The preamplifier usually does not have gains much greater than 10; otherwise its time constant is too large and its speed is limited. It is not good practice to eliminate the preamplifier altogether since kickback into the driving circuitry will then result in very limited accuracy. *Kickback* denotes the charge transfer either into or out of the inputs when the track-and-latch stage goes from track mode to latch mode. This charge transfer is caused by the charge needed to turn the transistors in the positive-feedback circuitry on and by the charge that must be removed to turn transistors in the tracking circuitry off. Without a preamplifier or buffer, this kickback will enter the driving circuitry and cause very large glitches, especially in the case when the impedances seen looking into the two inputs are not perfectly matched.

In high-resolution applications, capacitive coupling and reset switches are also typically included to eliminate any input-offset-voltage and clock-feedthrough errors, in a manner similar to that described in Sections 7.1 and 7.2.

One very important consideration for comparators is to ensure that no memory is transferred from one decision cycle to the next. For example, if a comparator toggles

in one direction, it might have a tendency to stay in that direction. This tendency is sometimes called *hysteresis*. In order to eliminate it, one can reset the different stages before entering track mode. This might be achieved by connecting internal nodes to one of the power supplies or by connecting differential nodes together using switches before entering track mode. For example, the comparator shown in Fig. 7.13 has internal nodes of the latch reset to both V_{DD} and ground when the V_{ltch} signal is low. Not only does this eliminate memory, but it also sets the comparator to its *trip point*, which speeds up operation when the comparator resolves small input signals.

Another consideration for comparators is that the gain should not be too large; otherwise the time constants are too large, which limits the speed, especially in track mode.

The track-and-latch stage has many variations. The circuitry shown in Fig. 7.13 should be considered only symbolic, although it will suffice in many applications.

Latch-Mode Time Constant

The time constant of the latch, when it is in its positive-feedback (i.e., latch) phase, can be found by analyzing a simplified circuit consisting of two back-to-back inverters, such as those shown in Fig. 7.14. If we assume that the output voltages of the inverters are close to each other at the beginning of the latch phase, and that the inverters are in their linear range, then each of the inverters can be modelled as a voltage-controlled current source driving an RC load, as shown in Fig. 7.15, where A_V is the low-frequency gain of each inverter, which has a transconductance given by $G_m = A_V/R_L$. For this linearized model, we have

$$\frac{A_V}{R_L}V_y = -C_L\left(\frac{dV_x}{dt}\right) - \left(\frac{V_x}{R_L}\right) \tag{7.22}$$

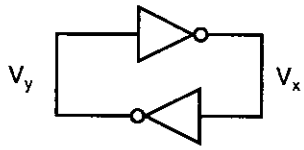


Fig. 7.14 Two back-to-back inverters used as a simplified model of a track-and-latch stage in its latch phase.

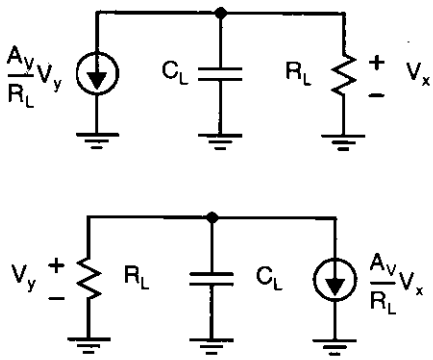


Fig. 7.15 A linearized model of the track-and-latch stage when it is in its latch phase.

and

$$\frac{A_V}{R_L}V_x = -C_L\left(\frac{dV_y}{dt}\right) - \left(\frac{V_y}{R_L}\right) \tag{7.23}$$

Multiplying (7.22) and (7.23) by R_L and rearranging gives

$$\tau\left(\frac{dV_x}{dt}\right) + V_x = -A_V V_y \tag{7.24}$$

and

$$\tau\left(\frac{dV_y}{dt}\right) + V_y = -A_V V_x \tag{7.25}$$

where $\tau = R_L C_L$ is the time constant at the output node of each inverter. Subtracting (7.25) from (7.24) and rearranging terms gives

$$\left(\frac{\tau}{A_V - 1}\right)\left(\frac{d\Delta V}{dt}\right) = \Delta V \tag{7.26}$$

where $\Delta V = V_x - V_y$ is the voltage difference between the output voltages of the inverters. Equation (7.26) is a first-order differential equation with no forcing function. Its solution is given by

$$\Delta V = \Delta V_0 e^{(A_V - 1)t/\tau} \tag{7.27}$$

where ΔV_0 is the initial voltage difference at the beginning of the latch phase. Thus, the voltage difference increases exponentially in time with a time constant given by

$$\tau_{ltch} = \frac{\tau}{A_V - 1} \cong \frac{R_L C_L}{A_V} = \frac{C_L}{G_m} \tag{7.28}$$

where, again, $G_m = A_V/R_L$ is the transconductance of each inverter. Note that τ_{ltch} is roughly equal to the inverse of the unity-gain frequency of each inverter.

In the case of MOS devices, normally the output load is proportional to the gate-source capacitance of a single transistor, or specifically,

$$C_L = K_1 W L C_{ox} \tag{7.29}$$

Here, K_1 is a proportionality constant between 1 and 2. Also, the inverter transconductance is proportional to the transconductance of a single transistor, and is given by

$$G_m = K_2 g_m = K_2 \mu_n C_{ox} \frac{W}{L} V_{eff} \tag{7.30}$$

where K_2 might be between 0.5 and 1. Substituting (7.29) and (7.30) into (7.28) gives

$$\tau_{ltch} = \frac{K_1}{K_2}\frac{L^2}{\mu_n V_{eff}} = K_3 \frac{L^2}{\mu_n V_{eff}} \tag{7.31}$$

where K_3 might be between 2 and 4. Note that (7.31) implies that τ_{ltch} depends primarily on the technology and not on the design (assuming a reasonable design is used

that maximizes V_{eff} and minimizes C_L). Note also the similarity between (7.31) and (7.20), the equation for the time constant of a cascade of gain stages. For a given technology, (7.31) is very useful in determining a rough estimate for the maximum clock frequency of a latch-and-track comparator.

If it is necessary for a voltage difference of ΔV_{logic} to be obtained in order for succeeding logic circuitry to safely recognize the correct output value, then, by using (7.27), we find the time necessary for this to happen is given by

$$T_{ltch} = \frac{C_L}{G_m} \ln\left(\frac{\Delta V_{logic}}{\Delta V_0}\right) = K_3 \frac{L^2}{\mu_n V_{eff}} \ln\left(\frac{\Delta V_{logic}}{\Delta V_0}\right) \tag{7.32}$$

If ΔV_0 is small, this latch time can be large, perhaps larger than the allowed time for the latch phase. Such an occurrence is often referred to as *metastability*. In other words, because its initial value is too small, the differential output voltage of the latch does not increase enough to be recognized as the correct logic value by succeeding circuitry. Sometimes, even when the initial voltage difference is large enough, it is possible that circuit noise can cause the initial voltage difference to become small enough to cause metastability.

EXAMPLE 7.5

Assume for a 0.8-μm technology that $K_3 = 2.5$, $\Delta V_0 = 10$ mV, $\Delta V_{logic} = 2$ V, $V_{eff} = 0.5$ V, and $\mu_n = 0.05$ $M^2/V\cdot s$. What is the maximum clocking frequency of the comparator?

Solution

Using (7.32), we have $T_{ltch} = 0.34$ ns. Assuming this is half the total period (which might be quite an optimistic assumption), the smallest possible period for a track-and-latch might be 0.68 ns, which corresponds to a clocking frequency of 1.5 GHz. In a typical comparator, the maximum clock rate would probably be limited by the frequency response of the preamplifiers and the track-and-latch during the track phase, rather than by the speed of the latch during the latch phase.

7.4 EXAMPLES OF CMOS AND BiCMOS COMPARATORS

This section describes a number of high-speed comparators that are currently popular.

The literature on integrated circuit technology has many examples of latched comparators. For example, Fig. 7.16 shows a comparator presented in [Song, 1990]. This comparator has the positive feedback of the second stage always enabled. In track mode, when the two diode-connected transistors of the gain stage are enabled, the gain around the positive-feedback loop is less than one, and the circuit is stable. The combination of the diode-connected transistors of the gain stage and the transistors of the

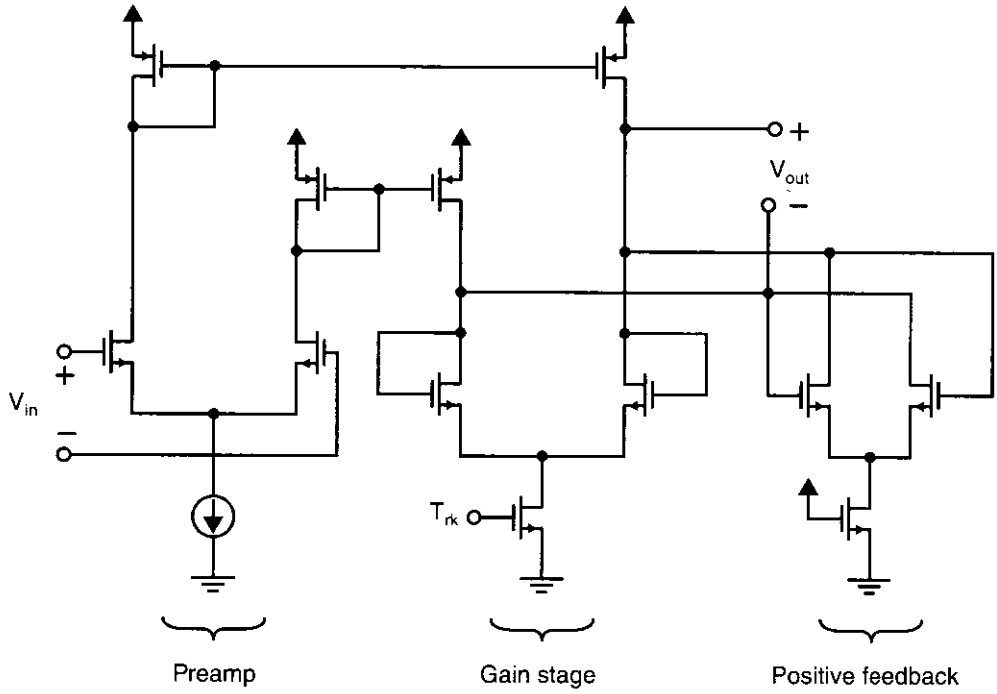


Fig. 7.16 A two-stage comparator that has a preamplifier and a positive-feedback track-and-latch stage.

positive-feedback loop acts as a moderately large impedance and gives gain from the preamplifier stage to the track-and-latch stage. The diode-connected loads of the preamplifier stage give a limited amount of gain in order to maximize speed, while still buffering the kickback from the input circuitry.

A second comparator is shown in Fig. 7.17 [Norsworthy, 1989]. This design also uses diode-connected loads to keep all nodes at relatively low impedance (similar to current-mode circuit-design techniques), thereby keeping all node time constants small and giving fast operation. This design also uses precharging to eliminate any memory from the previous decision. For example, the positive-feedback stage is precharged low, whereas the digital-restoration stage is precharged high, somewhat similar to what is done in Domino CMOS logic [Krambeck, 1982].

A third comparator was designed, fabricated, and characterized by K. Martin in 1984 (unpublished) and is shown in Fig. 7.18, with appropriate clock waveforms shown in Fig. 7.19. This design eliminates any input-offset voltages from both the first and second stages by using capacitive coupling. It also has common-mode feedback circuitry for the first preamplifier stage, which allows for input signals that have large common-mode signals. Unlike in fully differential opamps, in fully differential comparators, the linearity of the common-mode feedback circuitry is noncritical since, whenever large signals are present (and the common-mode feedback circuitry becomes nonlinear), there is no ambiguity in resolving the sign of the input signal. This allows a simple differential pair to be used for CMFB circuitry. Measured (but

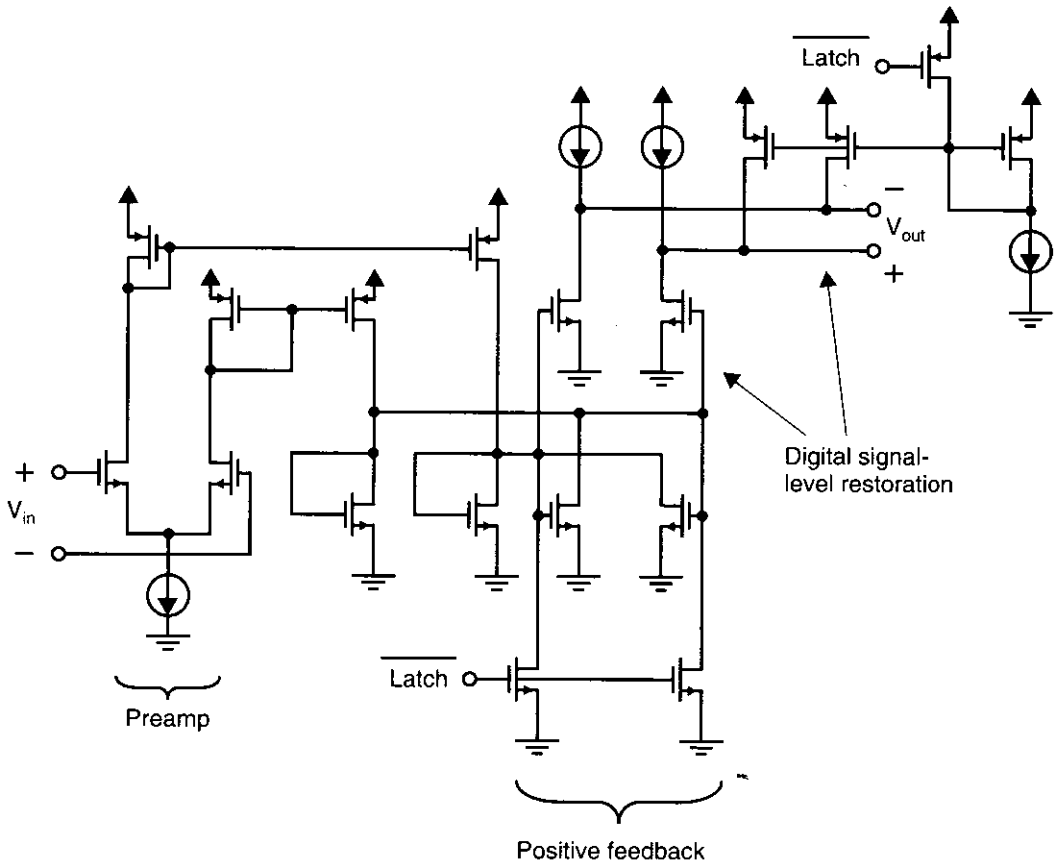


Fig. 7.17 A two-stage comparator.

unpublished) performance of this circuit resulted in a 0.1-mV resolution at a 2-MHz clock frequency, despite a very old 5-μm technology. The performance measured was limited by the test set-up rather than by the circuitry.

A fourth comparator that is realized in a BiCMOS technology and that exhibits very good performance, is described in [Razavi, 1992]. Indeed, this comparator is one of the better BiCMOS designs to date. It is based on the idea that it is not necessary to reset the first stage; the input-offset errors of the first stage can still be eliminated by resetting the right side of the coupling capacitors between the first stage and the second stage, assuming that the gain of the first stage is not too large [Poujois, 1978], [Vittoz, 1985]. A simplified schematic of this comparator is shown in Fig. 7.20. During the reset phase, the inputs to the preamplifier are connected directly to ground (or to a reference voltage), and the outputs of the coupling capacitors are connected to ground as well. This stores any offset voltages of the first stage on the capacitors. When the comparator is taken out of reset phase, then the effect of the clock feedthrough of S_5 and S_6 on the input resolution is divided by the gain of the first stage. In the realization described in [Razavi, 1992], the first stage is a BiCMOS

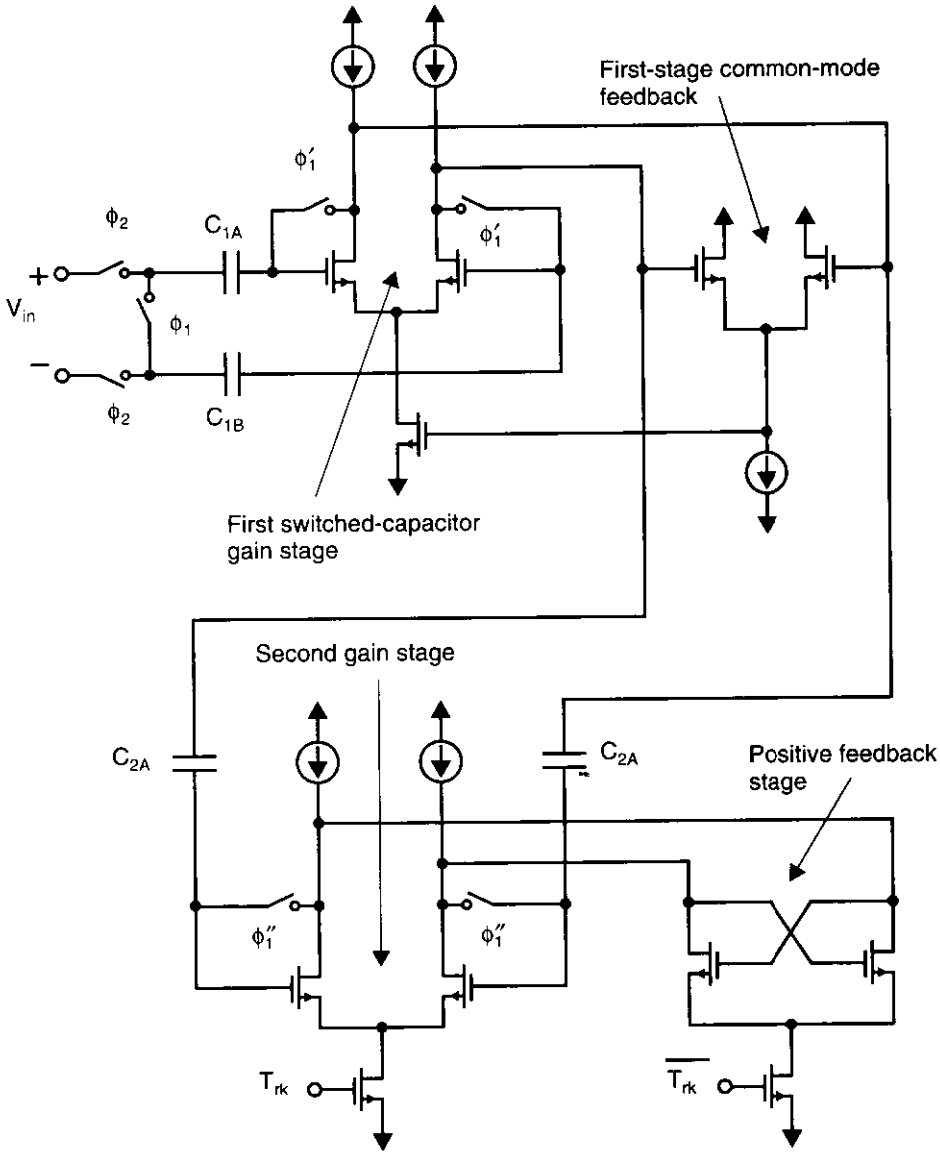


Fig. 7.18 A two-stage comparator with capacitive coupling to eliminate input-offset voltage and clock-feedthrough errors, along with positive feedback for fast operation.

preamplifier, consisting of MOS source followers followed by a bipolar differential amplifier and emitter-follower output buffers, as shown in Fig. 7.21. Notice that the circuit operates between ground and a negative voltage supply. Note also that the switches are realized using p-channel transistors. The track-and-latch stage has both a bipolar latch and a CMOS latch, as shown in Fig. 7.22. During the reset phase, ϕ_1 is low and ϕ_2 is high, X_1 and Y_1 are connected to ground through switches M_1 and M_2, C_3 is discharged to the minus supply through M_{11}, and X_2 and Y_2 are discharged to

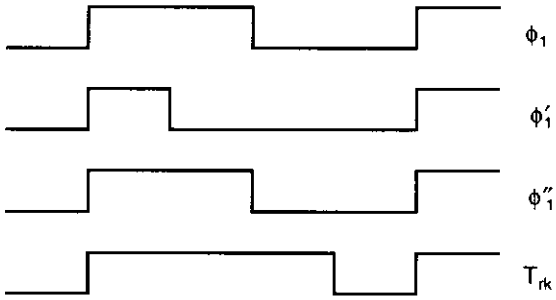


Fig. 7.19 The clock waveforms required by the comparator in Fig. 7.18.

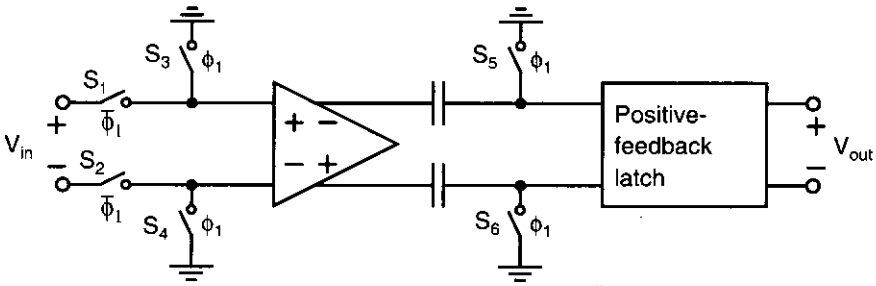


Fig. 7.20 A simplified schematic of the comparator described in [Razavi, 1992].

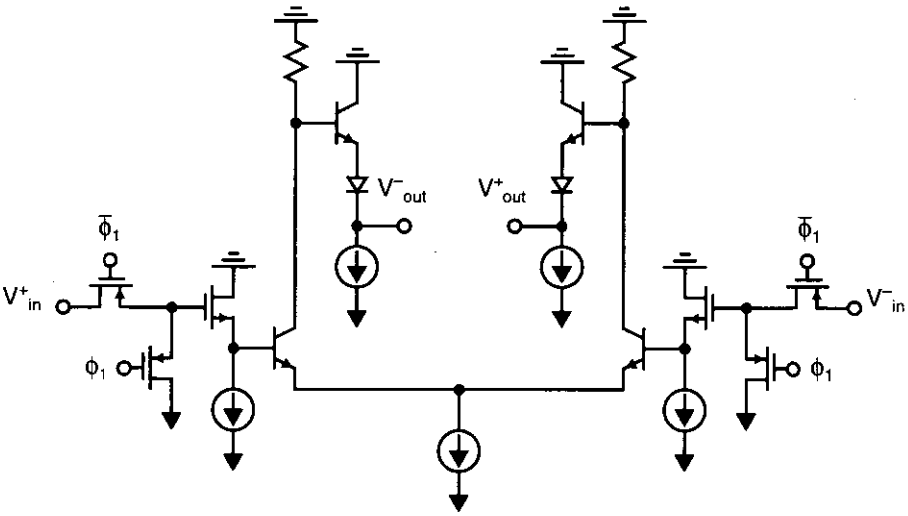


Fig. 7.21 A BiCMOS preamplifier.

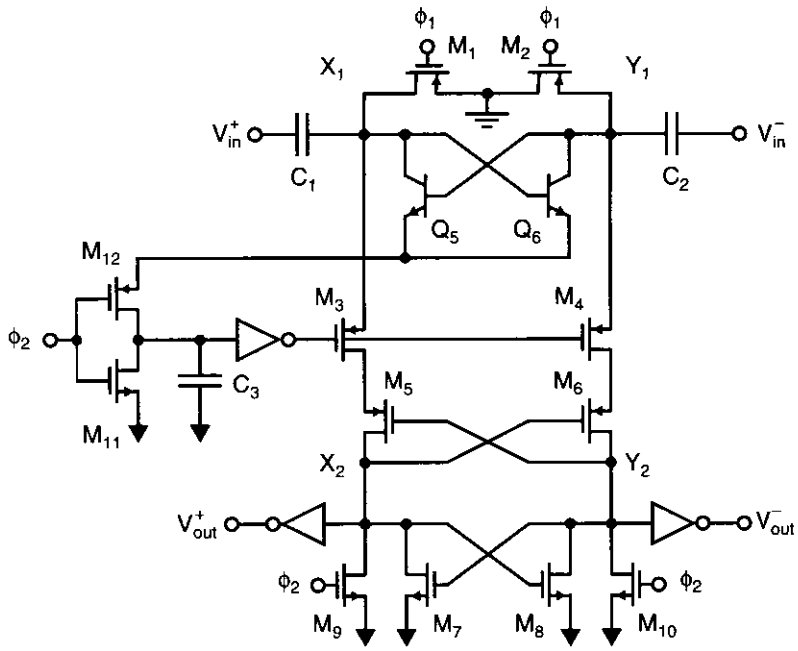


Fig. 7.22 The BiCMOS track-and-latch stage in [Razavi, 1992].

the minus supply through M_9 and M_{10}. Also, at this time, M_3 and M_4 are off. Next, ϕ_1 becomes high, which leaves X_1 and Y_1 floating and connects the outputs of the preamplifier to its input signal. After a short delay, which is needed for the transient response of the preamplifier, ϕ_2 becomes low, which turns on M_{12}, thereby activating the positive feedback action of Q_5 and Q_6. This develops a differential voltage between X_1 and Y_1 of about 200 mV, since C_3 is about one-fifth the size of C_1 and C_2. The offset of this bipolar latch is quite small, typically on the order of one millivolt or less. Also note that, when activated, the only power dissipated is that required to charge C_3—there is no dc power dissipation. A short time after the bipolar latch is activated, the inverter connected to C_3 changes state, and its output voltage will drop. This low voltage turns on M_3 and M_4, which activates the CMOS part of the latch consisting of cross-coupled M_5 and M_6 and cross-coupled M_7 and M_8. The reasons for including cross-coupled M_5 and M_6 are to prevent nodes X_1 and Y_1 from being discharged very far from ground toward the minus power supply and at the same time to amplify the 200-mV difference signal developed by the bipolar latch to almost a full-level CMOS voltage change. Thus, the latch is accurate, relatively fast, and low power. Also described in [Razavi, 1992] is an interesting CMOS-only comparator, which is not described here, but which interested readers may investigate.

Input-Transistor Charge Trapping

When realizing very accurate CMOS comparators, an error mechanism that must be considered is charge trapping in the gate oxide of the input transistors of the MOS

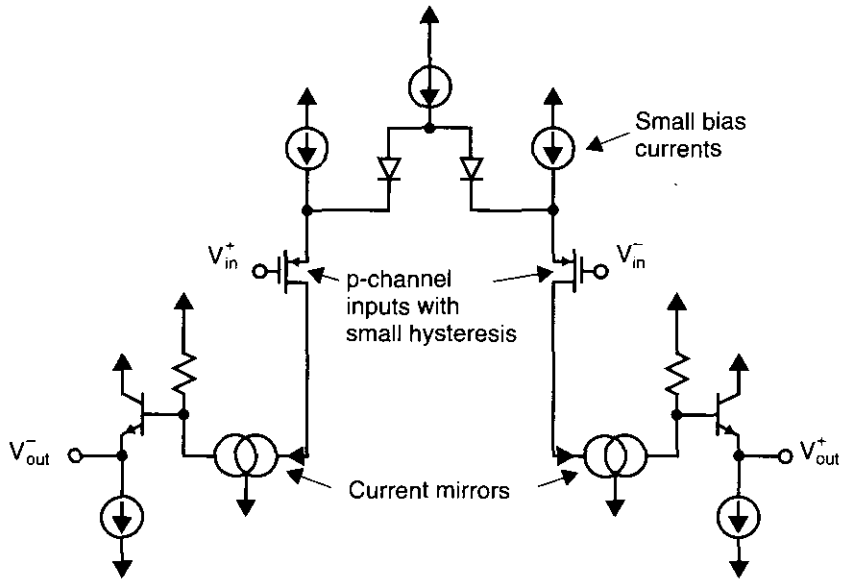


Fig. 7.23 A comparator with little hysteresis. (Hysteresis is caused by charge trapping.)

comparators [Tewksbury, 1989]. When n-channel transistors are stressed with large positive gate voltages, electrons can become trapped via a tunneling mechanism in which electrons tunnel to oxide traps close to the conduction band. The time constant for the release of these trapped electrons is on the order of milliseconds and is much longer than the time it takes for them to become trapped. During the time they are trapped, the effective transistor threshold voltage is increased. This leads to a comparator hysteresis on the order of 0.1 to 1 mV. This effect correlates well with transistor 1/f noise and is much smaller in p-channel transistors. When very accurate comparators are needed, charge trapping must be considered and minimized. One possible means of minimizing this effect, assuming we have a BiCMOS technology, is described in [Miller, 1990] and is illustrated in Fig. 7.23. A BiCMOS technology allows diodes to be used in the input stage, which allows the inclusion of two additional small-current sources in the input stage; this guarantees that the p-channel input transistors never turn off.[5] Also, by using p-channel input transistors, the charge trapping is greatly minimized because p-channel transistors exhibit much less hysteresis than n-channel transistors.

An alternative approach is to flush the input transistors after each use where the junctions and wells of n-channel transistors are connected to a positive power supply whereas the gates are connected to a negative power supply. This effectively eliminates the trapped electrons [Swanson, 1993]. Alternatively, or in addition, two input stages can be used for the comparator—a rough stage can be used during times when

5. p-channel transistors exhibit trapping for positive gate voltages since only electrons, and not holes, are trapped. Therefore, to prevent trapping, they should never be turned off.

large signals with overloads are possible, whereas a fine stage can be used during times when accurate comparisons are necessary and it can be guaranteed that no large signals are present [Swanson, 1993; Tan, 1990].

7.5 EXAMPLES OF BIPOLAR COMPARATORS

High-speed bipolar comparators are typically latched comparators, as shown previously in Fig. 7.13, where preamplifier and track-and-latch stages are used. A typical example is shown in Fig. 7.24 [Wakimoto, 1988]. The preamplifier is a simple differential amplifier that might have a gain of around 5. Besides increasing the resolution, this preamplifier helps to eliminate kickback from the track-and-latch stage. The track-and-latch stage is very similar to a current-mode digital latch. During track phase, the differential pair consisting of transistors Q_3 and Q_4 is enabled and operates as a second differential amplifier. When latch mode is enabled, I_2 is diverted to the differential pair consisting of Q_5 and Q_6. The inputs to this third differential pair come from the outputs of the second differential pair through emitter-follower buffers consisting of Q_9 and Q_{10}. The outputs of the third differential pair are connected back to the high-impedance output nodes in a cross-coupled manner so that positive feedback results. When this positive feedback is enabled, it takes the initially-small differential voltage and quickly amplifies it to between 0.4 and 0.6 V (depending on the logic levels being used). There are two reasons for using the emitter-follower buffers. These buffers isolate the

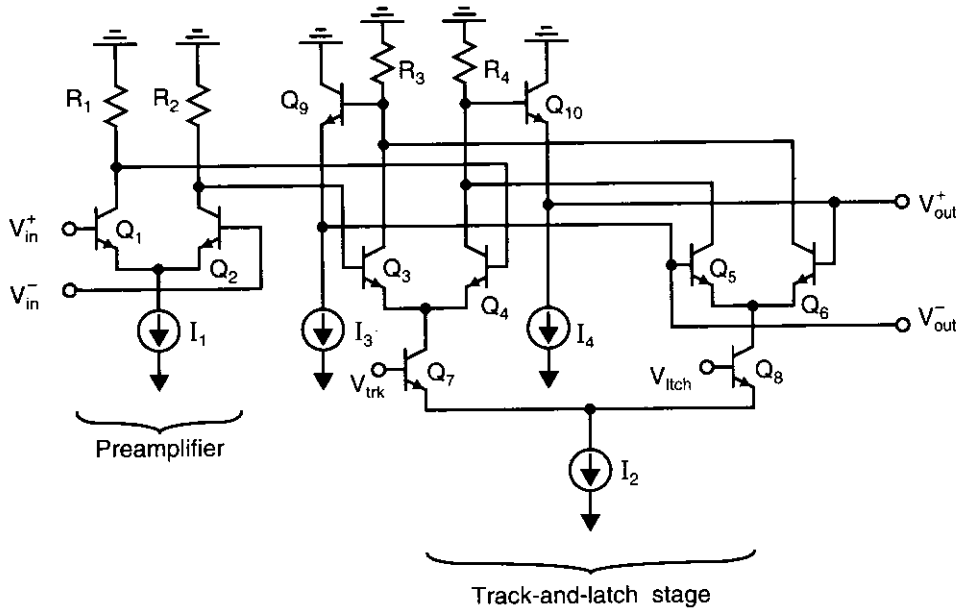


Fig. 7.24 A typical bipolar comparator.

high-impedance nodes at the collectors of the second and third differential pairs from the base-emitter capacitances of the third differential pair. They also isolate the same high impedance nodes from the output load capacitance.

Another architecture for a bipolar comparator is shown in Fig. 7.25 [Van de Plassche, 1988]. In this comparator, the input signals go to a differential pair that is always turned on (even during the latch phase). During the track phase, Q_3 and Q_4 are on. They operate as common-base transistors, and therefore the comparator operates as a cascode amplifier during this phase. When the comparator goes into latch mode, the collector current from the input transistors is redirected from the common-base transistors to transistors Q_7 and Q_8, which are connected in a positive-feedback loop. Since the bias current of the input transistors is never turned off, the kickback from them is greatly reduced without adding additional preamplifier stages. This maximizes the bandwidth during the tracking phase.

A third bipolar comparator is shown in Fig. 7.26. Although this comparator has never been experimentally tested, it is used as an example to illustrate many comparator design principles. A small current source is added to the tracking differential pair so that this differential pair doesn't turn off completely when the comparator is in hold mode. This makes the comparator faster when it goes from hold mode into track mode. Also, a cascode common-base stage is added just below the output voltages. This stage is used to isolate the outputs from the positive-feedback loop. This in turn allows for much smaller resistors to be used in the positive-feedback loop, which speeds this loop up and eliminates the need for emitter-follower buffers in the loop. Finally, a larger resistor is added between the emitters of the cascode transistors. This resistor prevents the cascode transistors from turning off completely during the latch phase, which also speeds up the transition from latch mode to tracking mode.

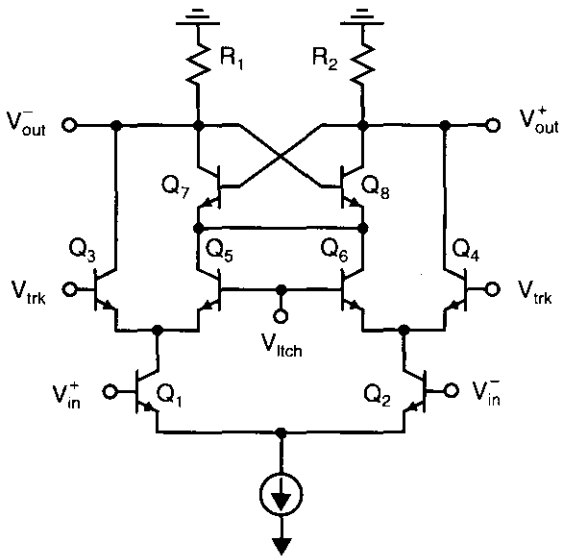


Fig. 7.25 A second bipolar comparator.

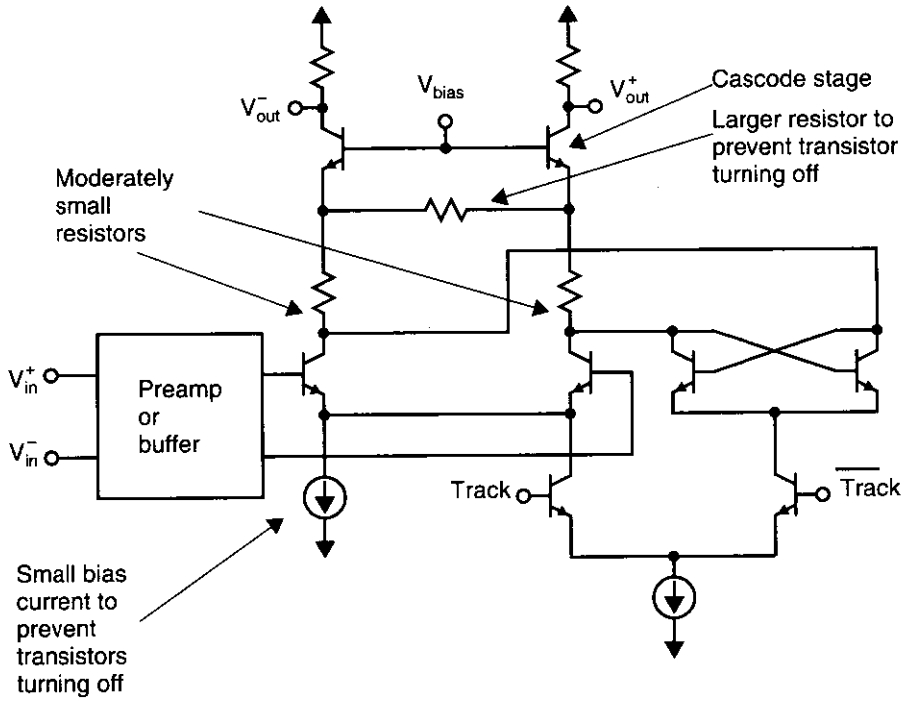


Fig. 7.26 A third bipolar comparator.

7.6 REFERENCES

D. G. Haigh and B. Singh. "A Switching Scheme for Switched-Capacitor Filters, Which Reduces Effect of Parasitic Capacitances Associated with Control Terminals," *Proc. IEEE Int. Symp. on Circuits and Systems*, Vol. 2, pp. 586–589, June 1983.

R. H. Krambeck, C. Lee, and H. S. Law. "High-Speed Compact Circuits with CMOS," *IEEE J. of Solid-State Circuits*, Vol. 17, no. 3, pp. 614–619, June 1982.

K. Martin, "Improved Circuits for the Realization of Switched-Capacitor Filters," *IEEE Trans. Circuits and Systems*, Vol. CAS-27, no. 4, pp. 237–244, April 1980.

———. "New Clock-Feedthrough Cancellation Technique for Analog MOS Switched Capacitor," *Electron. Lett.*, no. 18, 1992.

K. Martin, L. Ozcolak, Y. S. Lee, and G. C. Temes. "A Differential Switched-Capacitor Amplifier," *IEEE J. of Solid-State Circuits*, Vol. SC-22, no. 1, pp. 104–106, February 1987.

L. L. McCreary and P. R. Gray. "All-MOS Charge Redistribution Analog-to-Digital Conversion Techniques—Part I, "*IEEE J. of Solid-State Circuits*, Vol. SC-10, pp. 371–379, December 1975.

G. Miller, M. Timko, H. S. Lee, E. Nestler, M. Mueck, and P. Ferguson. "An 18-b 10-μs Self-Calibrating ADC," *IEEE Intern. Solid-State Circuits Conf.*, pp. 168–169, February 1990.

S. Norsworthy, I. Post, and S. Fetterman. "A 14-bit 80-KHz Sigma-Delta A/D Converter: Modeling, Design, and Performance Evaluation," *IEEE J. of Solid-State Circuits*, Vol. 24, no. 2, pp. 256–267, April 1989.

R. Poujois and J. Borel. "A Low-Drift Fully Integrated MOSFET Operational Amplifier," *IEEE J. of Solid-State Circuits*, Vol. SC-13, pp. 499–503, August 1978.

B. Razavi, *Principles of Data Conversion System Design*, IEEE Press, Piscataway, New Jersey, 1995.
B. Razavi and B. Wooley. "Design Techniques for High-Speed, High-Resolution Comparators," *IEEE J. of Solid-State Circuits*, Vol. 27, no. 12, pp. 1916–1926, December 1992.
J. Shieh, M. Patil, and B. L. Sheu. "Measurement and Analysis of Charge Injection in MOS Analog Switches," *IEEE J. of Solid-State Circuits*, Vol. 22, no. 2, pp. 277–281, April 1987.
B. Song, H. S. Lee, and M. Tompsett. "A 10-b 15-MHz CMOS Recycling Two-Step A/D Converter," *IEEE J. of Solid-State Circuits*, Vol. 25, no. 6, pp. 1328–1338, December 1990.
E. J. Swanson, "Method and Circuitry of Decreasing the Recovery Time of a MOS Differential Voltage Comparator," U.S. patent no. 5,247,210, September 1993.
K. Tan, et al. "Error-Correction Techniques for High-Performance Differential A/D Converters," *IEEE J. of Solid-State Circuits*, Vol. 25, no. 6, pp. 1318–1327, December 1990.
T. L. Tewksbury, H. S. Lee, and G. Miller. "The Effects of Oxide Traps on the Large-Signal Transient Response of Analog MOS Circuits," *IEEE J. of Solid-State Circuits*, Vol. 24, no. 2, pp. 542–543, April 1989.
R. Van de Plassche and P. Baltus. "An 8-bit 100-MHz Full-Nyquist Analog-to-Digital Converter," *IEEE J. of Solid-State Circuits*, Vol. 23, no. 6, pp. 1334–1344, December 1988.
E. A. Vittoz, "Dynamic Analog Techniques," in *Design of MOS VLSI Circuits for Telecommunications*, ed. Y. Tsividis and P. Antognetti, Prentice Hall, Englewood Cliffs, New Jersey, 1985.
T. Wakimoto, Y. Akazawa, and S. Konaka. "Si Bipolar 2-GHz 6-bit Flash A/D Conversion LSI," *IEEE J. of Solid-State Circuits*, Vol. 23, no. 6, pp. 1345–1350, December 1988.
Y. S. Yee, L. M. Terman, and L. G. Heller. "A 1-mV MOS Comparator," *IEEE J. of Solid-State Circuits*, Vol. SC-13, pp. 294–297, June 1978.
Yubawa, A. "A CMOS 8-bit High-Speed A/D Converter," *IEEE J. of Solid-State Circuits*, Vol. SC-20, no. 3, pp. 775–779, June 1985.

7.7 PROBLEMS

Unless otherwise stated, assume the following:

- n-channel MOS transistors:
  $\mu_n C_{ox} = 92\ \mu A/V^2$
  $V_{tn} = 0.8\ V$
  $\gamma = 0.5\ V^{1/2}$
  $r_{ds}\ (\Omega) = 8000\ L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 2.4 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.0 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$
- p-channel MOS transistors:
  $\mu_p C_{ox} = 30\ \mu A/V^2$
  $V_{tp} = -0.9\ V$
  $\gamma = 0.8\ V^{1/2}$
  $r_{ds}\ (\Omega) = 12{,}000\ L\ (\mu m)/I_D\ (mA)$ in active region
  $C_j = 4.5 \times 10^{-4}\ pF/(\mu m)^2$
  $C_{j\text{-}sw} = 2.5 \times 10^{-4}\ pF/\mu m$
  $C_{ox} = 1.9 \times 10^{-3}\ pF/(\mu m)^2$
  $C_{gs(overlap)} = C_{gd(overlap)} = 2.0 \times 10^{-4}\ pF/\mu m$

7.1 A simple CMOS inverter is used as a comparator, as shown in Fig. P7.1. Assume ϕ_{1a} drops before ϕ_1 drops, so that only the charge injection from Q_1 need be considered. What is the change in V_{out} due to this charge injection?

Assume the transistors in the inverter have sizes chosen such that the inverter threshold voltage is at $V_{DD}/2 = 2.5$ V (i.e., $W_5/W_4 = \mu_n/\mu_p$), and that all transistor lengths are 0.8 μm. Also assume the inverter gain is –24 at $V_{out} = 2.5$ V, and ignore overlap capacitance. Do not ignore the body effect on the threshold voltage of Q_1.

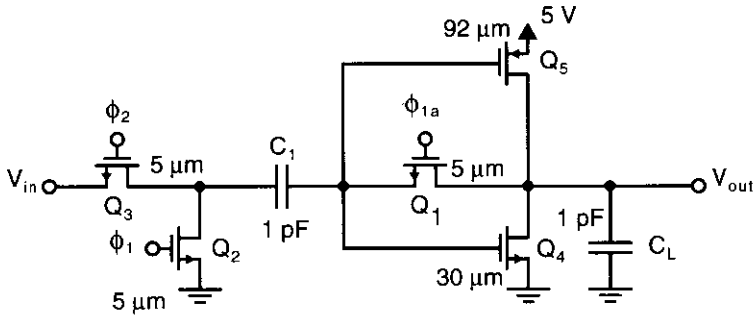


Fig. P7.1

7.2 Assume that three stages of inverters identical to the inverter used in Problem 7.1 are cascaded as shown in Fig. 7.10. What is the input-voltage offset error due to the charge injection of the reset switch of the last stage only?

7.3 Estimate the time constant for the inverter used in Problem 7.1 during its reset phase and also during its comparison phase.

7.4 For the circuit shown in Fig. 7.10, find the voltages at each of the opamp inputs at the end of ϕ_1', ϕ_1'', ϕ_1''', ϕ_1, and ϕ_2, where V_i is the input and $V_{err\text{-}i}$ is the error due to the reset switch of the ith opamp. Assume opamps are offset free and that they remain in their linear region after their reset switch is opened. Show that the overall input-referred clock feedthrough is given by $|V_{err\text{-}3}|/(A_1A_2)$.

7.5 Repeat Problem 7.4 where offsets of the opamps are not ignored. What is the input-referred offset and clock feedthrough?

7.6 Assume a MOS transistor's unity-gain frequency is approximately given by $\omega_u = g_m/C_{gs}$. Find an expression for ω_t in terms of V_{eff} and C_{ox}. Assume a single gain stage without compensation has a unity-gain frequency given by $\omega_t = \omega_u/2$. Further assume that a comparator is realized by a cascade of three stages, each having a gain of –15. If three time constants are required for settling during the open-loop comparison phase, what is the maximum frequency of operation for the comparator? What is the resolution? You may assume that $V_{eff} = 0.25$ V, $\mu_n = 0.05\ M^2/V \cdot s$, and that there are only n-channel transistors in the signal path.

7.7 Give reasonable device widths of the comparator in Fig. 7.16 so that the first stage has a gain of 5 and so that the latch-and-track stage has a gain of 10 when

in track mode and a loop gain of 4 when in latch mode. Assume the input transistors are 25-μm wide and that the bias current of the first stage is 0.1 mA.

7.8 Assume the comparator of Fig. 7.16 has the device sizes found in Problem 7.7. Estimate the equivalent time constant when the comparator is in track mode. If the input signal changes from 0 to 5 mV at time 0, how long would it take the differential output voltage to change by 50 mV?

7.9 Consider the comparator in Fig. 7.16 with the device sizes found in Problem 7.7. Estimate the equivalent time constant of the latch-and-track stage when the comparator is in latch mode. If the differential output voltage is 50 mV at the beginning of the latch phase, how long would it take for the output voltage to change to a differential voltage of 2 V?

7.10 What is the minimum power-supply voltage that can be used for the comparator in Fig. 7.25 if the current source and the resistive loads both require 0.4 V across them? You may assume that for the bipolar transistors $V_{CE\text{-}sat} = 0.3$ V.

7.11 The CMOS transmission gate shown has an input voltage of 2.5 V when it turns off. The W/L of the n-channel is 5 μm/0.8 μm and the W/L of the p-channel is 15 μm/0.8 μm. Assume the total parasitic capacitance between the output node and ground is 50 fF, $V_{DD} = 5$ V, and that the clock signals change very fast. Also ignore changes due to overlap capacitance. Estimate the change in output voltage due to charge injection. What will the final output voltage be?

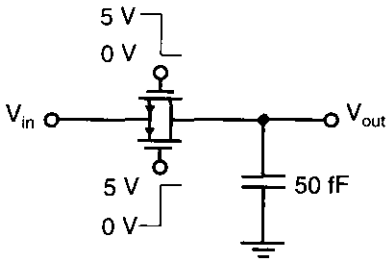


Fig. P7.11

CHAPTER

8 Sample and Holds, Voltage References, and Translinear Circuits

In this chapter, some common analog building blocks are described other than those already presented. Specifically, circuits to realize sample and holds, bandgap voltage references, and translinear gain cells and multipliers are discussed. A *sample and hold* (S/H) is used to sample an analog signal and to store its value for some length of time. A *bandgap voltage reference* is often used to bias circuits or to supply a reference to which other voltages are compared. A *translinear gain cell* is commonly used to create an amplifier whose gain can be adjusted through the use of a controlling input current. Finally, a *translinear multiplier* is realized by interconnecting two gain cells and finds a variety of uses in applications such as a modulator (in a communication system) or a phase detector (in a phase-locked loop).

8.1 PERFORMANCE OF SAMPLE-AND-HOLD CIRCUITS

An important analog building block, especially in data-converter systems, is the sample-and-hold-circuit. Before proceeding, it is worthwhile to mention that sample-and-hold circuits are also often referred to as "track-and-hold" circuits. Normally, these two terms are synonymous except for a few particular switched-capacitor sample-and-hold circuits that do not have a phase where the output signal is tracking the input signal. Sample-and-hold circuits are necessary components in many data-acquisition systems such as A/D converters. In many cases, the use of a sample and hold (at the front of the data converter) can greatly minimize errors due to slightly different delay times in the internal operation of the converter.

Before discussing the basic principles of sample-and-hold circuits, it is necessary to mention some performance parameters used in characterization.

1. The first of these parameters is a *sampling pedestal* or a *hold step*. This is an error that occurs each time a sample and hold goes from sample mode to hold mode. During this change in operation, there is always a small error in the voltage being held that makes it different from the input voltage at the time of sampling. Obviously, this error should be as small as possible. *Perhaps more importantly, this error should be signal independent; otherwise it can introduce nonlinear distortion.*

2. Another parameter is a measure of how isolated the sampled signal is from the input signal during hold mode. Ideally, the output voltage will no longer be affected by changes in the input voltage. In reality, there is always some signal feedthrough, usually through parasitic capacitive coupling from the input to the output. In well-designed sample and holds, this signal feedthrough can be greatly minimized.
3. A third important parameter is the speed at which a sample and hold can track an input signal, when in sample mode. In this mode, a sample and hold will have both small-signal and large-signal limitations due to its –3-dB bandwidth and finite slew rate, respectively. Both the –3-dB bandwidth and slew rate should be maximized for high-speed operation.
4. Yet another limitation (somewhat less important in high-speed designs) is the *droop rate* in hold mode. This error is a slow change in output voltage, when in hold mode, caused by effects such as leakage currents due to the finite base currents of bipolar transistors and reverse-biased junctions. In most CMOS designs, this droop rate is so small it can often be ignored.
5. A fifth limitation is *aperture jitter* or *aperture uncertainty*. This error is the result of the effective sampling time changing from one sampling instance to the next and becomes more pronounced for high-speed signals. Specifically, when high-speed signals are being sampled, the input signal changes rapidly, resulting in small amounts of aperture uncertainty causing the held voltage to be significantly different from the ideal held voltage.

Other performance parameters are also important when realizing sample and holds. These include parameters such as dynamic range, linearity, gain, and offset error. Some of the mechanisms whereby these errors arise in particular sample and holds are described in the next section using specific S/H examples. In addition, a number of design principles whereby these errors can be minimized are explained.

Testing Sample and Holds

Before describing various S/H architectures, it is appropriate to first describe a popular method for testing S/Hs, referred to as the *beat test*. This test consists of clocking the S/H at its maximum clock frequency and applying a sinusoidal input signal that has a frequency slightly different than the clock frequency.[1] The output of this system is demodulated to a low frequency equal to the difference between the frequency of the clock signal and that of the input signal. This low-frequency signal is then characterized using a spectrum analyzer or by digitizing it using a high-accuracy A/D converter clocked at the difference frequency and then analyzed using a computer. The test setup for a beat test is shown in Fig. 8.1. Example waveforms for the input signal, the sampling signal, and the output signal of the sample and hold are shown in Fig. 8.2.

1. If the circuit being tested is, in fact, a track and hold, then the output of the track and hold would be resampled by a second track and hold clocked on the opposite phase as the original.