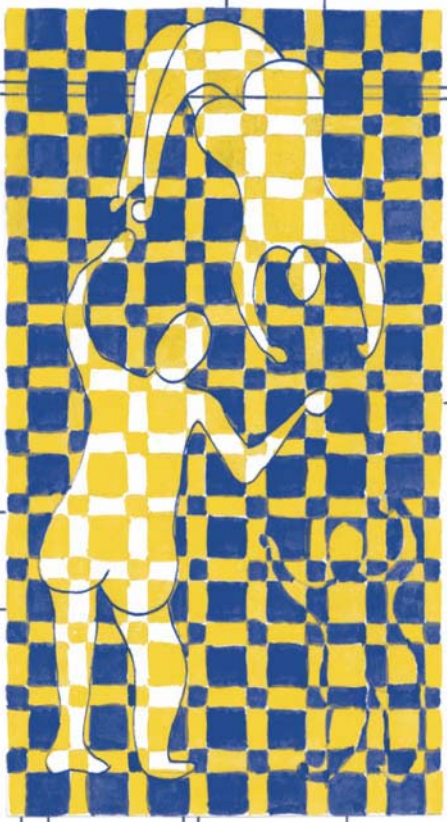
Analog VLSI: Circuits and Principles



Liu
Kramer
Indiveri
Delbrück
Douglas

foreword by
Carver A. Mead

Analog VLSI: Circuits and Principles

This page intentionally left blank

Analog VLSI: Circuits and Principles

Shih-Chii Liu, Jörg Kramer, Giacomo Indiveri, Tobias Delbrück, and Rodney Douglas
with contributions from Albert Bergemont, Chris Diorio, Carver A. Mead, Bradley A. Minch, Rahul Sarpeshkar, and Eric Vittoz.

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by the authors using the LaTeX document preparation system.
Printed on recycled paper and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Analog VLSI : circuits and principles / Shih-Chii Liu ... [et al.] with contributions from Albert Bergemont ... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-262-12255-3 (hc. : alk. paper)
1. Integrated circuits, Very large scale integration. 2. Linear integrated circuits. I. Liu, Shih-Chii.
TK7874.75 .A397 2002
621.39'5—dc21

2002021915

This book is dedicated to the memory of our creative colleague and friend, Misha Mahowald, who was a pioneer and an inspiration in this field.

This page intentionally left blank

Contents

Contents

This page intentionally left blank

Authors and Contributors

Albert Bergemont
Maxim Integrated Products
3725 North First Street,
San Jose, CA 95134–1350
U.S.A.

Tobias Delbrück
Institute of Neuroinformatics,
ETH/UNIZ
Winterthurerstrasse 190
8057 Zurich, Switzerland

Chris Diorio
Department of Computer Science
and Engineering
The University of Washington
114 Sieg Hall, Box 352350
Seattle, WA 98195
U.S.A.

Rodney Douglas
Institute of Neuroinformatics,
ETH/UNIZ
Winterthurerstrasse 190
8057 Zurich, Switzerland

Giacomo Indiveri
Institute of Neuroinformatics,
ETH/UNIZ
Winterthurerstrasse 190
8057 Zurich, Switzerland

Jörg Kramer
Institute of Neuroinformatics,
ETH/UNIZ
Winterthurerstrasse 190
8057 Zurich, Switzerland

Shih-Chii Liu
Institute of Neuroinformatics,
ETH/UNIZ
Winterthurerstrasse 190
8057 Zurich, Switzerland

Carver A. Mead
Department of Computation and
Neural Systems
California Institute of Technology
Pasadena, CA 91125
U.S.A.

Bradley A. Minch
Department of Electrical Engineering
Cornell University
405 Phillips Hall
Ithaca, NY 14853–5401
U.S.A.

Rahul Sarpeshkar
Research Laboratory of Electronics
Massachusetts Institute of
Technology
Cambridge, MA 02139
U.S.A.

Eric Vittoz
Chief Scientist
Advanced Microelectronics
Center for Electronics and
Microtechnology
Jaquet-Droz 1
2007 Neuchatel
Switzerland

This page intentionally left blank

Acknowledgments

This book was written by a small group of authors who represent the work of a far larger community. We would like to acknowledge our colleagues who have contributed to the advance of concepts and circuits in neuromorphic engineering; in particular, John Lazzaro, Massimo Silvilotti, John Tanner, Kwabena Boahen, Paul Hasler, Steve Deweerth, Ron Benson, Andre van Schaik, John Harris, Andreas Andreou, Ralph Etienne-Cummings, and many others. We especially wish to thank the following people for their help in the completion of this book: Andre Van Schaik, Regina Mudra, Elisabetta Chicca, and Ralph Etienne-Cummings for their constructive comments in earlier versions of the book; Samuel Zahnd for putting together the material for the example circuits on the website; Adrian Whatley for ensuring the integrity of the bibliography, David Lawrence for dealing with computer mishaps, Mietta Loi for entering some of the material in the book, Kathrin Aguilar-Ruiz for dealing with legal details, Claudia Stenger for her endless patience with all sorts of requests, and Donna Fox for always providing the answers for difficult requests. We also thank Sarah K. Douglas for the cover design of this book. The work in this fledging field has been supported by progressive funding organizations: National Science Foundation, Office of Naval Research, Gatsby Charitable Foundation, Swiss National Science Foundation, Whitaker Foundation, Department of Advanced Research Projects Agency, and our various home institutions. We also acknowledge Mike Rutter for his enthusiasm in starting this project, and Bob Prior for seeing the project to its completion.

This page intentionally left blank

Preface

The aim of this book is to present the collective expertise of the neuromorphic engineering community. It presents the central concepts required for creative and successful design of analog very-large-scale-integrated (VLSI) circuits. The book could support teaching courses, and provides an efficient introduction to new practitioners who have some previous training in engineering, physics, or computer science.

Neuromorphic engineers are striving to improve the performance of artificial systems by developing chips and systems that process information collectively using predominantly analog circuits. Consequently, our book biases the discussion of analog principles and design towards novel circuits that emulate natural signal processing. These circuits have been used in implementations of neural computational systems or neuromorphic systems and biologically-inspired processing systems. Unlike most circuits in commercial or industrial applications, our circuits are operated mainly in the subthreshold or weak inversion region. Moreover, their functionality is not limited to linear operations, but encompasses also many interesting nonlinear operations similar to those occuring in natural systems.

Although digital circuits are the basis for a large fraction of circuts in current VLSI systems, certain computations like addition, subtraction, expansion, and compression are natural for analog circuits and can be implemented with a small number of transistors. These types of computations are prevalent in the natural system which has an architecture which is not that of a conventional Turing machine. The mechanisms for signaling in the neural system which are governed by Boltzmann statistics can be captured by circuits comprising metal-oxide-semiconductor field effect transistors (MOSFETs) that operate in the subthreshold or weak inversion regime. Because the exponential dependence of charges on the terminal voltages of a MOSFET is similar to those of the bipolar junction transistor (BJT), current techniques for constructing a circuit which implements a given function using bipolar circuits, can be extended to MOSFET circuits[1]. Besides the advantage of the reduced power consumption of MOSFET circuits that operate in the weak inversion regime, this new circuit philosophy also translates to novel circuits and system architectures.

Local memory is an essential part of any artificial parallel distributed processor or neural network system. In this book, we show circuits for analog memory storage and for implementing local and global learning rules using floating-gate charge modulation techniques in conventional CMOS tech-

1 BJTs are traditionally used for analog circuits in industry.

nology. We also show how by using floating-gate circuits together with the translinear principle, we can develop compact circuits which implement a large class of nonlinear functions.

The first integrated aVLSI system that implemented a biological function was a silicon retina by Mead and Mahowald. This system used analog circuits that performed both linear and nonlinear functions in weak-inversion operation. Following this initial success, subsequent examples of simple computational systems and novel circuits have been developed by different labs. These examples include photoreceptor circuits, silicon cochleas, conductance-based neurons, and integrate-and-fire neurons. These different circuits form the foundation for a physical computational system that models natural information processing.

The material presented in this book has evolved from the pioneering series of lectures on aVLSI and principles introduced by Carver Mead into the Physics of Computation curriculum at the California Institute of Technology in the mid 80s. Today, similar courses are taught at many institutions around the world; and particularly, at the innovative annual Telluride Neuromorphic Workshop (funded by the US National Science Foundation, and others). Many of the people who teach these courses are colleagues who were trained at Caltech, or who have worked together at Telluride.

We have been fortunate to obtain the enthusiastic participation of the many authors who provided material and the primary text for this book. Their names are associated with each chapter. However, the result is not simply an edited collection of papers. Each of the authors named on the cover made substantial contributions to the entire book. Liu and Douglas have edited the text to provide a single voice.

We have attempted to make the material in this book accessible to readers from any academic background by providing intuition for the functionality of the circuits. We hope that the book will prove useful for insights into novel circuits and that it will stimulate and educate researchers in both engineering and interdisciplinary fields, such as computational neuroscience, and neuromorphic engineering.

Foreword

Carver Mead

In the beginning of any new technology, the applications of the technology cannot be separated from the development of the technology itself. With vacuum tubes, discrete transistors, or integrated circuits, the first circuit topologies were invented by those closest to the device physics and fabrication process. As time passes, the knowledge of how to *make* the devices gradually separates from the knowledge of how to *use* the devices. Abstractions are developed that greatly simplify our conceptual models of the underlying technology. Canonical circuit forms are encapsulated into symbols that have meaning in the field of use, rather than in the space of implementation. Familiar examples are logic gates, operational amplifiers, and the like. As more time passes, the abstractions become entrenched in university courses and industrial job descriptions. It is common to hear phrases like "we need to hire a system architect, two logic designers, a circuit designer, and a layout specialist" or "who will we get to teach the op-amp course next year?" This inexorable trend toward specialization fits well with the myth that putting more engineers on a design task will make it happen faster. In my personal experience, I have never seen this myth reflected in reality. In many cases, I have seen chips designed by large groups fail to converge at all. Chip designs that converge rapidly and perform well have always been done by small, cohesive groups. For the truly great chip designs, there has always been a single person that could keep the entire design in their head. The lesson from these experiences is unmistakable: leading contributors are able to lead because they move past the limitations of the prevailing paradigm.

The second myth is that digital techniques are displacing analog techniques throughout modern electronics. In fact, an explosion of new analog applications is occurring as this sentence is being written. Cellular telephones, fiberoptic transmitters and receivers, wideband wired and wireless data modems, electronic image capture devices, digital audio and video systems, "smart" power control, and many others all have explicitly analog circuits. But, and perhaps even more important, *all* circuits are analog, even if they are used as digital building blocks. Rise times, delays, settling times, and metastable behaviors are analog properties of nominally digital circuits. As the speed of operation approaches the limits of the technology, analog considerations increasingly dominate the process of digital design.

Modern integrated circuit technology is extremely sophisticated; it is capable of realizing a vast array of useful device and circuit structures. Designers who avail themselves of the richness of the technology have a powerful advantage over those who are limited to logic gates and operational amplifiers. This book contains precisely the information required by such renaissance designers. Device physics, process technology, linear and nonlinear circuit forms, photodetectors, floating-gate devices, and noise analysis are all described in clear, no-nonsense terms. It stands in refreshing contrast to the litany of operational amplifier and switched capacitor techniques that are widely mistaken for the whole of analog design.

1 Introduction

This book presents an integrated circuit design methodology that derives its computational primitives directly from the physics of the used materials and the topography of the circuitry. The complexity of the performed computations does not reveal itself in a simple schematic diagram of the circuitry on the transistor level, as in standard digital integrated circuits, but rather in the implicit characteristics of each transistor and other device that is represented by a single symbol in a circuit diagram. The main advantage of this circuit-design approach is the possibility of very efficiently implementing certain 'natural' computations that may be cumbersome to implement on a symbolic level with standard logic circuits. These computations can be implemented with compact circuits with low power consumption permitting highly-parallel architectures for collective data processing in real time. The same type of approach to computation can be observed in biological neural structures, where the way that processing, communication, and memory have evolved has largely been determined by the material substrate and structural constraints. The data processing strategies found in biology are similar to the ones that turn out to be efficient within our circuit-design paradigm and biology is thus a source of inspiration for the design of such circuits.

The material substrates that will be considered for the circuits in this book are provided by standard integrated semiconductor circuit technology and more specifically, by **C**omplementary **M**etal **O**xide **S**ilicon (CMOS) technology. The reason for this choice lies in the fact that integrated silicon technology is by far the most widely used data processing technology and is consequently commonly available, inexpensive, and well-understood. CMOS technology has the additional advantages of only moderate complexity, cost-effectiveness, and low power consumption. Furthermore it provides basic structures suitable for implementation of short-term and long-term memory, which is particularly important for adaptive and learning structures as found ubiquitously in biological systems. Although we will specifically consider CMOS technology as a physical framework it turns out that various fundamental relationships are quite similar in other frameworks, such as in bipolar silicon technology, in other semiconductor technologies and to a certain extent also in biological neural structures. The latter similarities form the basis of *neuromorphic emulation* of biological circuits on an electrical level that led to such structures as silicon neurons and silicon retinas.

The book is divided into four sections: Silicon and Transistors; Statics; Dynamics; and Special Topics. The first section (Silicon and Transistors) provides a short introduction into the underlying physics of the devices that are discussed in the rest of the book; in particular the operation of the MOSFET in the subthreshold region and a discussion of analog charge storage using floating-gate technology. Chapter 2 discusses useful equations that can be derived from modeling the physics of the basic devices in the silicon substrate. These device models provide a foundation for the derivation of the equations governing the operation of the MOSFET as described in Chapter 3. From results discussed in Chapters 2 and 3 we show in Chapter 4 how MOS technology can be used to build analog charge storage elements. Readers who are more interested in circuits at the transistor level description may omit Chapters 2 and 4 and continue to the Statics section.

The Statics section comprises three chapters. These chapters describe examples of linear and nonlinear static functions that can be implemented by simple circuits. Chapter 5 presents some basic circuits which show the richness of the processing that can be performed by the transistor. Chapter 6 introduces an analog circuit design concept where currents represent the signal and state variables in a circuit. As examples, some current-mode circuits are described that implement nonlinear functions that are prevalent in natural systems, for example, a winner-take-all circuit. Chapter 7 derives a methodology for implementing a large class of linear and nonlinear functions using a particular building block called a multiple-input translinear element.

The Dynamics section describes circuits which process time-varying signals. Chapter 8 reviews the basics of linear systems theory, which is a useful tool for the small-signal analysis of circuits both in the time and space domain. We apply this theory in Chapter 9 to selected examples of simple circuits for first-order and second-order filters. Chapter 10 provides a brief introduction into semiconductor photosensors and focuses on circuits that model prominent properties of biological photoreceptors. It also gives an overview of common image sensing principles.

The last section (Special Topics) contains chapters which expound further on the basics of semiconductor technology. These chapters cover topics on noise in transistors, the flow from design to layout to fabrication of an integrated circuit, and the issue of scaling semiconductor technology into the future. Chapter 11 describes the different noise sources in a transistor and how these sources can be measured. It also presents a novel way of demonstrating the equivalence of thermal noise and shot noise. The circuit layout masks

needed to specify a layout to a fabrication house are listed in Chapter 12 along with useful layout tips for good circuit performance. Chapter 13 describes the processing steps executed in a 0.25 μm process and Chapter 14 projects how transistors will scale in future technologies.

This book is directed towards students from a variety of backgrounds. The students who are more interested in circuits can omit Section I and should still be able to follow the chapters in Sections II and III. Most of the material in this book has been taught in classes at the Institute for Neuroinformatics, University of Zurich/ ETH Zurich, Switzerland, and also at the Telluride Neuromorphic Engineering Workshop.

Examples of simulation and layout files for simple circuits are available from the Institute of Neuroinformatics website. These circuits were fabricated and used in our laboratory courses. They have been developed using the integrated circuit design tools available from Tanner Research, Inc. Students who are interested in simulating these circuits can also use the free public domain software *AnaLOG* by John Lazzaro and David Gillespie.

This page intentionally left blank

I SILICON AND TRANSISTORS

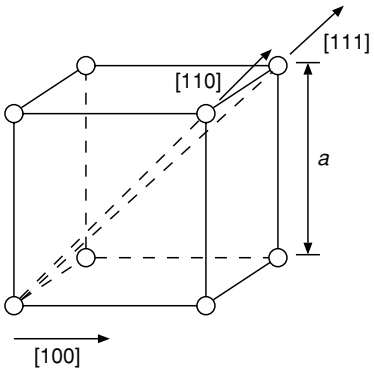
This page intentionally left blank

2 Semiconductor Device Physics

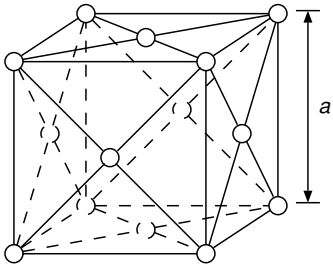
The purpose of this chapter is to provide an introduction to the basics of the semiconductor physics needed for the understanding of the devices described in this book. Most of this introduction pertains to semiconductors in general. Where general statements are not possible we focus on silicon. The values of material constants, and the typical values of other parameters, are for silicon. It is not intended to provide a detailed step-by-step derivation of the formulas describing device behavior. Often we limit ourselves to stating the necessary conditions for the derivation to hold and the important results without formal derivation. More extended summaries of solid-state and semiconductor physics can be found in standard texts (Grove, 1967; Sze, 1981; Kittel, 1996; Singh, 2001). Detailed analyses of the subject fill entire books (Dunlap, 1957; Smith, 1979; Moss, 1980).

2.1 Crystal Structure

In semiconductors and other materials the atoms are arranged in regular structures, known as *crystals*. These structures are defined and held together by the way the valence (outermost) electrons of the atoms are distributed, given that electrons tend to form pairs with antiparallel spin. Figure 2.1 shows the crystal structures of some important semiconductors; *silicon (Si)* and *gallium arsenide (GaAs)*. A silicon atom, for example, has four unpaired *valence electrons* that can form *covalent bonds* with a tetrahedral spatial characteristic (Fig. 2.1(b)). Pure silicon naturally crystallizes in a *diamond structure*. The diamond lattice is based on the face-centered cubic (fcc) arrangement shown in Fig. 2.1(a), which means that the atoms are located at the corners and face centers of cubes with a given side length a, called the lattice constant. The diamond structure consists of two interleaved fcc lattices that are displaced by $a/4$ in each dimension. Silicon has a lattice constant of $a = 5.43$ Å. Gallium arsenide has the same structure as silicon, except that one of the interleaved fcc lattices holds the gallium atoms and the other the arsenic atoms. This arrangement is known as *zincblende structure*.

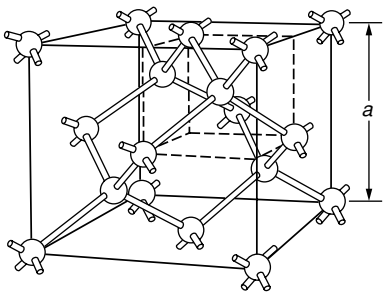


Simple cubic

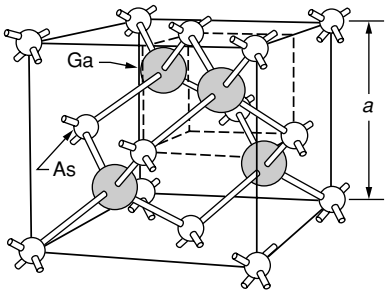


Face-centered cubic

(a)



Diamond
(C, Ge, Si, etc)



Zincblende
(GaAs, GaP, etc)

(b)

Figure 2.1
Various crystal lattice structures, with lattice constant a. (a) Simple cubic lattice with atoms at the cube corners, and face-centered cubic (fcc) lattice with additional atoms on the cube faces. The most important crystal directions [100], [110], and [111] of the simple cubic lattice are indicated. (b) Structures with two interleaved fcc lattices: Diamond structure, consisting of one kind of atom, and zincblende structure, consisting of two kinds of atoms. Figure adapted from S. M. Sze (1981), Physics of Semiconductor Devices, 2nd Edition.

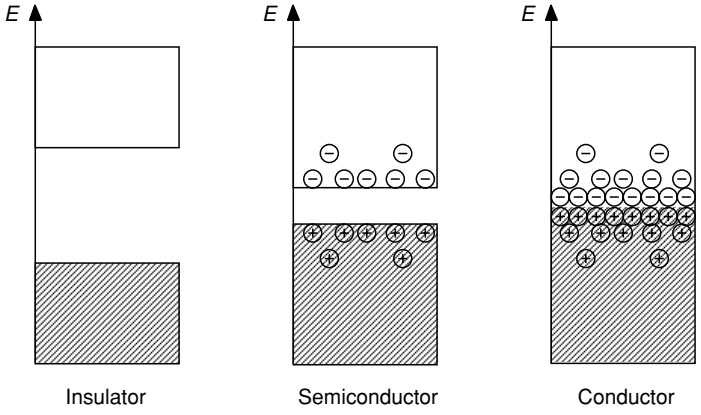


Figure 2.2
Schematic representation of electron energy bands in a crystal for an insulator, a semiconductor, and a conductor (or metal). The energy bands are represented as boxes. The hatched areas symbolize the states in the energy bands that are occupied by electrons at zero temperature. At non-zero temperatures some electrons (denoted by circled minus signs) occupy higher-energy states, leaving "holes" (denoted by circled plus signs) in the unoccupied lower-energy states.

2.2 Energy Band Diagrams

Crystals and other solids are classified according to their electrical conductivity into *insulators*, *semiconductors*, and *conductors* or *metals* in the order of increasing conductivity. Electrical processing structures, such as transistors, are mostly fabricated from semiconductors because they operate at an intermediate conductivity level, which can be modulated by varying the electrical boundary conditions and by introducing atoms of foreign elements into the crystal structure. This latter process is called *impurity doping*. Conductors and insulators also play an important role in electrical circuits, because they connect and separate respectively, the nodes of the processing structures. For example, in integrated silicon technologies *silicon dioxide* (SiO_2) (often referred to as *oxide*) is commonly used as an insulator, while polycrystalline silicon, also known as *polysilicon*, and aluminum are used as conductors.

The physical basis for the above classification of the materials lies in the properties of the atoms and their arrangement. Electrical currents in solids are carried by the motion of valence electrons, which are attracted to the fixed positively charged ion cores. A valence electron can thus either be bound to a particular ion core by electromagnetic forces or it can be mobile and

contribute to the current flow. In order to be mobile, an electron must acquire a certain minimum energy to break free of its ion core. This energy is called *ionization energy*. While free electrons can be in any energy state, the energies of electrons in solids lie in certain ranges of values, which are separated by forbidden zones due to the interaction of the valence electrons with the ion cores. The allowed ranges of energy values are called *energy bands* and the forbidden zones are called *energy gaps* or *bandgaps*. Each energy band in a crystal consists of closely-spaced discrete levels. This discretization of energy levels is a quantum-mechanical effect that is due to the spatial confinement of bound electrons and to the spatial periodicity of the potential energy for mobile electrons. An energy level supports a limited number of states, each of which is either occupied by an electron or empty at any given time. Simplified energy band diagrams of an insulator, a semiconductor, and a conductor are schematically shown in Fig. 2.2. Under normal conditions, the energy bands of insulators are either completely filled or completely empty, that is all electrons are bound. A metal has a partly filled bands within which electrons can move. Some metals have two or more overlapping energy bands that are partly filled. They are often referred to as *semimetals*. In a semiconductor, one or more bands are either almost filled or almost empty. Current flow is then influenced by certain physical parameters and boundary conditions.

The energy bands in semiconductors are called *valence bands* and *conduction bands* according to whether they are almost filled or almost empty. In a conduction band, electrons are essentially free to move, while in valence bands the electrons are bound to the atom bodies and can only move from one of the few unoccupied states to another one, if they cannot acquire enough energy to bridge the bandgap to a conduction band. An elegant mathematical concept, which is commonly used throughout the semiconductor literature, allows a more symmetric view of this mechanism. This concept considers unoccupied energy states in a valence band as *holes*. The term derives from the notion that an electron is absent from a state that is usually occupied. An electron valence band can thus be considered as a hole conduction band, since the holes are quite free to move around, that is electrons in the valence band in the vicinity of an unoccupied state may hop to that state and leave their initial state free for another electron to hop in, and so on. In common terminology the definition of valence and conduction bands is related to electrons and the holes are thus said to move in the valence band. In mathematical expressions the holes can be treated as positively charged particles that within a semiconductor acquire as much physical reality as electrons. Consequently, they are also attributed

other physical parameters used to characterize particles, such as mass and mobility. However, it is important to note that holes are not just positively charged electrons, but have different characteristic parameter values. Furthermore, you should keep in mind that the symmetry between electrons and holes breaks down as soon as the charge carriers leave the semiconductor, as we shall see, for example, in the chapters dealing with floating-gate structures.

Energy bands in a crystal have certain properties that are closely related to the crystal structure and thus vary significantly between different types of crystals. Graphic representations of the energy-band diagrams of Si and GaAs are shown in Fig. 2.3. The allowed electron energies are plotted as a function of the electron momentum for two sets of directions (cf. Fig. 2.1(a)), namely the [100] directions along the edges of the crystal lattice and the [111] directions along the lattice diagonals. By convention, energy-band diagrams are drawn such that electron energy increases in the upward direction. Energy values are usually specified in units of *electron volts (eV)*. This unit is convenient for the conversion of an energy-band diagram into an electrostatic potential distribution, which is obtained by dividing the energy values by the electron charge, that is the negative value of the elementary charge $q = 1.60218 \times 10^{-19}$ C. An electron volt is the energy corresponding to a potential change of an electron of one volt.

For simplicity, only the valence band and the conduction band with the smallest energy separation are shown in Fig. 2.3. The lines represent the *band edges*, that is the highest-energy states of the valence band and the lowest-energy states of the conduction band. The band edges tell us the minimum amount of energy an electron has to acquire or lose to bridge the bandgap for a given change in momentum. The difference between the lowest conduction band energy and the highest valence band energy is called *bandgap energy* E_g. The bandgap energy of silicon at room temperature is 1.12 eV. The valence band edge appears at zero momentum and is degenerate, that is common to several valence bands, for the most widely-used semiconductors. The momentum associated with the conduction band edge may be zero or not, depending on the semiconductor, and the conduction band edge is not degenerate. If the minimum of the conduction band edge is at the same momentum as the maximum of the valence band edge we speak of a *direct bandgap*, otherwise of an *indirect bandgap*. As we can see from Fig. 2.3, gallium arsenide has a direct bandgap and silicon has an indirect bandgap.

Electron energy and momentum changes can be induced by different physical processes, the most important of which are interactions with lattice vibra-

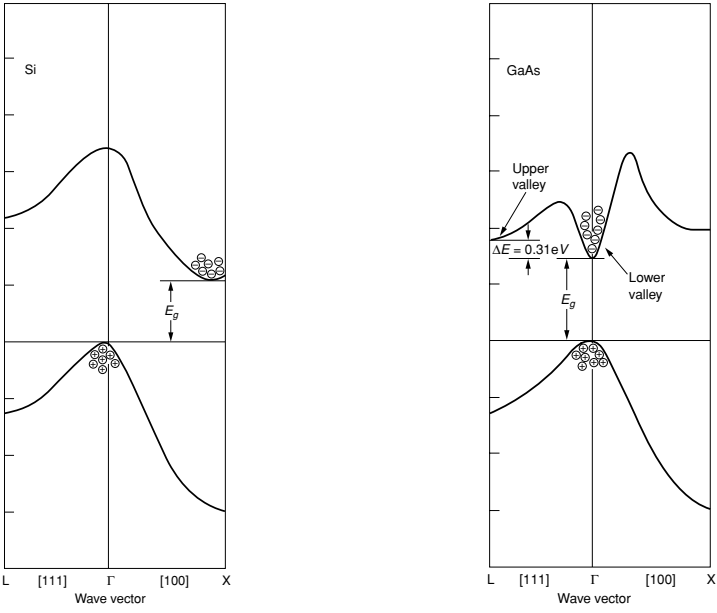


Figure 2.3
Energy-band diagrams of (a) silicon (Si) and (b) gallium arsenide (GaAs). Only the edges of the uppermost valence band and of the lowermost conduction band are shown as a function of the wave vector for two sets of directions in the crystal. The Γ point corresponds to charge carriers being at rest. The [111] set of directions is along the diagonals of the crystals, while the [100] set is oriented along the edges of the crystals, as shown in Fig. 2.1(a). The L point stands for wave vectors $(\pi/a)(\pm 1, \pm 1, \pm 1)$; and the X point stands for wave vectors $(2\pi/a)(\pm 1, 0, 0)$, $(2\pi/a)(0, \pm 1, 0)$ and $(2\pi/a)(0, 0, \pm 1)$. The momentum $\mathbf{p}$ of a charge carrier is computed from its wave vector $\mathbf{k}$, as $\mathbf{p} = \overline{h}\mathbf{k}$ where $\overline{h}$ is the reduced Planck constant. The bandgap energy E_g is the separation between the top of the topmost valence band and the bottom of the bottommost conduction band. These extrema appear at different momenta for Si and at the same momentum for GaAs. The bandgap of Si is thus called *indirect* and the bandgap of GaAs is called *direct*. Figure adapted from J. R. Chelikowsky and M. L. Cohen (1976), Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors, *Phys. Rev.*, **B14**, 556-582.

tions, that is collisions with the ions in the crystal, and with electromagnetic waves. The energies that are transferred during these interactions are quantized. The energy quantum of a crystal lattice vibration is called a *phonon* and the energy quantum of an electromagnetic wave is called a *photon*. Absorption

or emission of a phonon changes mainly the momentum of an electron while the energy change is typically small ($\sim$0.01 eV to 0.03 eV) compared to the bandgap energy. On the other hand, the momentum transfer by a photon with an energy of the order of the bandgap energy is negligible. This means that in the case of an indirect bandgap the transitions between valence and conduction bands with the smallest energies typically involve phonons and photons, while direct bandgaps can, for example, be bridged by photons alone. This is the reason why materials with direct bandgaps can be used as efficient sources of electromagnetic radiation, while those with indirect bandgaps are inefficient, because the minimum energy transitions, which are the most probable ones, depend on the availability of a phonon with the proper momentum.

Most semiconductor devices do not make use of the electromagnetic properties of the material and the circuits are therefore shielded from high-energy electromagnetic radiation. Changes in electron momentum are then much more easily induced than changes in energy that are large enough to make the electron bridge the bandgap. The majority of electrical properties of semiconductors can thus be sufficiently accurately described without considering the momentum space at all. In the energy band diagrams the electron energies of the maximum of the valence band edge and the minimum of conduction band edge is therefore usually plotted as a function of position in one- or two-dimensional space, while the structure of the bands in momentum space does not appear anymore. Such a simplified energy band diagram is shown in Fig. 2.4.

2.3 Carrier Concentrations at Thermal Equilibrium

Thermal energy expresses itself in vibrations of the crystal lattice. Energy transfer from the lattice to the electrons is thus established through absorption or emission of phonons. At zero temperature, all low-energy states are filled and all high-energy states are empty. At higher temperatures some electrons will leave their lower-energy states in favor of higher-energy states. The occupancy of energy states is statistically described by a probability distribution. This distribution is known as *Fermi-Dirac distribution*. The probability that an energy state with value E is occupied is given by

$$F(E) = \left(1 + e^{(E-E_F)/kT}\right)^{-1} \tag{2.3.1}$$

where E_F denotes the energy at which the occupation probability is 0.5, called *Fermi level* or *chemical potential*, $k = 1.38066 \times 10^{-23}$ J/K is the *Boltzmann*

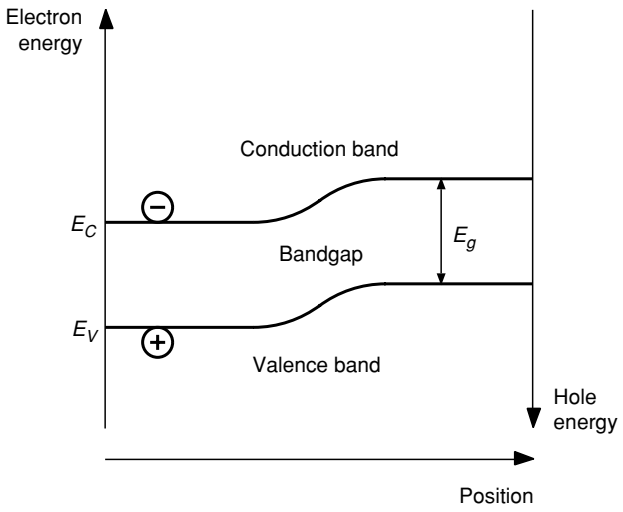


Figure 2.4
Simplified semiconductor energy-band diagram. The energy is plotted as a function of position in one dimension. Mobile charge carriers are symbolized by the signed circles.

constant, and T is the absolute temperature. For *intrinsic* (undoped) semiconductor crystals E_F is very close to the center of the bandgap. For typical impurity doping concentrations and bandgaps of commonly used semiconductors E_F is well-separated from the valence and conduction band edges, such that $|E-E_F| >> kT$ for all allowed energy states. This simplifies the Fermi-Dirac distribution in the conduction band to the *Boltzmann distribution*

$$\boxed{F(E) = e^{-(E-E_F)/kT}.} \tag{2.3.2}$$

This probability distribution is the reason for the exponential characteristics of diodes and transistors that will be described in this and the next chapter; and these devices will determine the characteristics of most circuits in this book.

The electron density $dn(E, dE)$ within an energy interval dE around an energy E is given by

$$dn(E, dE) = N(E)F(E)dE \tag{2.3.3}$$

where $N(E) \propto \sqrt{E - E_C}$ near the bottom of the conduction band. In *thermal equilibrium*, that is if no external voltage is applied to the semiconductor and no net current flows, the total electron density in the conduction band is

obtained by integrating Eq. 2.3.3 with respect to energy from the conduction band edge to infinity, resulting in

$$n = N_C e^{-(E_C - E_F)/kT} \tag{2.3.4}$$

where N_C denotes the effective density of states in the conduction band near its edge, and E_C is the energy of the conduction band edge. A corresponding equation can be derived for the hole density near the top of the valence band:

$$p = N_V e^{-(E_F - E_V)/kT} . \tag{2.3.5}$$

For intrinsic semiconductors n and p are equal. We define an *intrinsic carrier density* n_i as

$$n_i^2 = np \tag{2.3.6}$$

It follows from Eqs. 2.3.4, 2.3.5, and 2.3.6 that

$$n_i = \sqrt{np} = \sqrt{N_C N_V} e^{-E_g/2kT} \tag{2.3.7}$$

and that the Fermi level E_i for an intrinsic semiconductor is given by

$$E_i = \frac{E_C + E_V}{2} + \frac{kT}{2} \log\left(\frac{N_V}{N_C}\right) . \tag{2.3.8}$$

For silicon, $N_C = 2.80 \times 10^{19}\,\mathrm{cm}^{-3}$, $N_V = 1.04 \times 10^{19}\,\mathrm{cm}^{-3}$, and $n_i = 1.45 \times 10^{10}\,\mathrm{cm}^{-3}$ at room temperature. The concentration of Si atoms in a crystal is $5 \times 10^{22}\,\mathrm{cm}^{-3}$, which means that only one out of 3×10^{12} atoms is ionized at room temperature and therefore conductivity is very low.

2.4 Impurity Doping

The conductivity of a semiconductor can be increased significantly by doping it with impurities. In the doping process a small fraction of the semiconductor atoms in the crystal structure are replaced by atoms of a different element. As illustrated in Fig. 2.5, a *donor impurity* is an atom with a valence electron more than the semiconductor atom and an *acceptor impurity* is an atom with a valence electron less than the semiconductor atom. Note that impurity atoms are electrically neutral: The difference in number of electrons is balanced by an equal difference in number of protons in the nucleus. Since the additional valence electron of a donor does not fit into the crystal bond structure it is only loosely bound to its nucleus by electromagnetic forces. In the energy-band diagram donors form an energy level in the bandgap that is typically

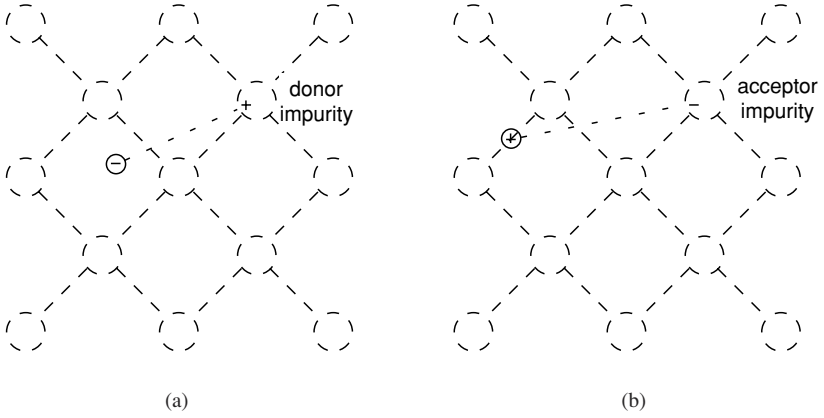


Figure 2.5
Illustration of a semiconductor with (a) donor and (b) acceptor impurity doping. The ion cores (dashed circles) are bound into the crystal structure by covalent bonds (dashed lines). The excess charge carrier (solid circle) that is introduced with the impurity does not fit into the covalent bond structure. This excess charge carrier is only loosely bound to the ion core of the impurity atom by electromagnetic forces and so is mobile. The hole introduced by an acceptor is a missing electron in a covalent bond.

close to the conduction-band edge. As long as the surplus electron is bound to its nucleus the corresponding state on that level is occupied and defined to be neutral. Conversely, acceptor atoms have loosely bound holes that appear as neutral states on an energy level in the bandgap that is typically close to the valence-band edge.

Figure 2.6 shows the energy-band diagrams, the state densities, the Fermi-Dirac distributions, and the carrier concentrations for differently-doped semiconductors. The introduction of donors moves the Fermi level from near the center of the bandgap further towards the conduction-band edge, whereas the introduction of acceptors moves it closer to the valence-band edge. If the donor-doping density is larger than the acceptor-doping density there

are more electrons in the conduction band than holes in the the valence band ($n > p$) under charge-neutrality conditions and the semiconductor is said to be n-*type*. In the reverse case, $p > n$ and the semiconductor is said to be p-*type*. The doping strength is often indicated by plus or minus signs. For example, a weak *p*-type doping is denoted by p^- and a very strong *n*-type doping by n^{++}. If the semiconductor is so strongly doped that the Fermi

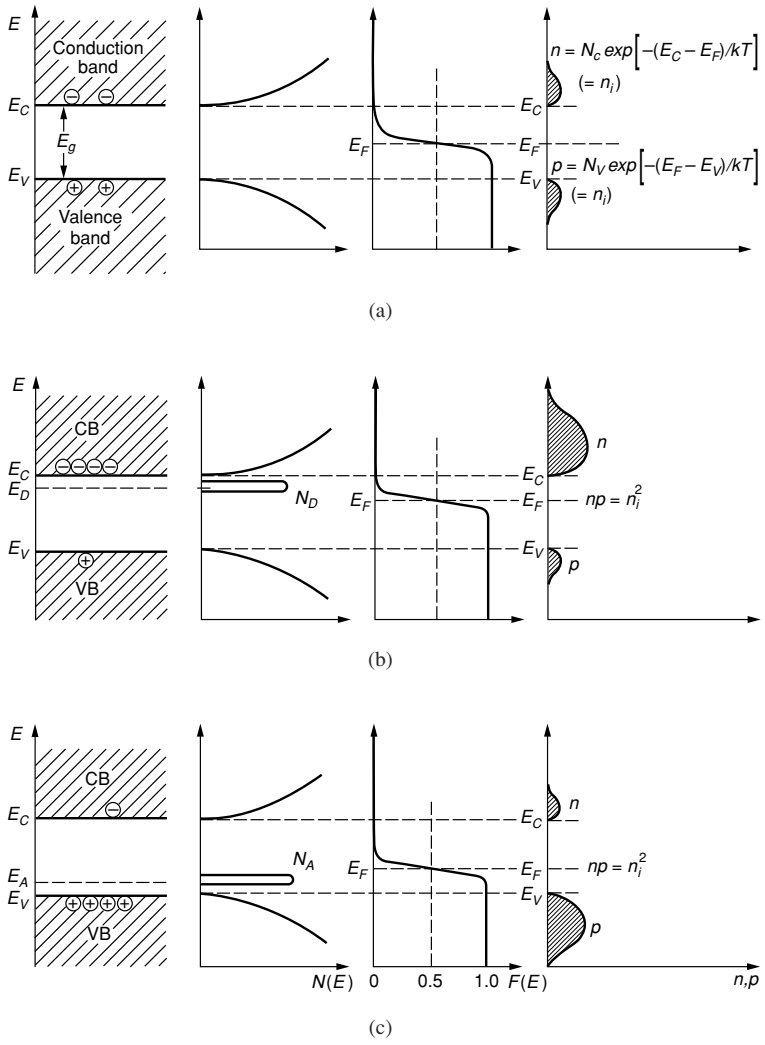


Figure 2.6
Energy-band diagram, density of states, Fermi-Dirac distribution, and carrier concentrations for (a) intrinsic, (b) *n*-type, and (c) *p*-type semiconductors at thermal equilibrium. The concentrations of mobile electrons and holes are indicated by the hatched areas in the plots on the right. Figure adapted from S. M. Sze (1981), Physics of Semiconductor Devices, 2nd Edition.

level is within the conduction or valence band or very near the edge of one of these bands, such that a large fraction of the states at the band edge are occupied, its properties become similar to those of a metal and we speak of a *degenerate semiconductor*. This happens for acceptor doping concentrations of around N_C and donor doping concentrations of around N_V. Commonly-used doping elements for silicon are phosphorus (P) and arsenic (As) as donors and boron (B) as an acceptor. The ionization energy of such an impurity atom, that is the energy required to remove the loosely-bound charge carrier from its ionic core, is on the order of 0.05 eV. This is only a small fraction of the bandgap energy and most donors and acceptors are thermally ionized at room temperature. The condition of charge neutrality in the crystal can then be stated as

$$n + N_A = p + N_D \tag{2.4.1}$$

where N_A denotes the acceptor impurity concentration and N_D the donor impurity concentration. Furthermore, Eq. 2.3.7 is also valid for doped semiconductors.

The mobile charge carriers that are more abundant in a semiconductor in thermal equilibrium are called *majority carriers*, whereas the sparser ones are called *minority carriers*. Using Eqs. 2.3.7 and 2.4.1 the concentration of majority electrons in the conduction band of an *n*-type semiconductor can be approximated by

$$n_{no} = \frac{1}{2}\left(N_D - N_A + \sqrt{(N_D - N_A)^2 + 4n_i^2}\right) \tag{2.4.2}$$

and the concentration of majority holes in the valence band of a *p*-type semiconductor by

$$p_{po} = \frac{1}{2}\left(N_A - N_D + \sqrt{(N_A - N_D)^2 + 4n_i^2}\right) . \tag{2.4.3}$$

For strongly doped *n*-type material with $N_D >> N_A$ and $N_D - N_A >> n_i$

$$n_{no} \approx N_D \tag{2.4.4}$$

and for strongly doped *p*-type material with $N_A >> N_D$ and $N_A - N_D >> n_i$

$$p_{po} \approx N_A . \tag{2.4.5}$$

The minority carrier concentrations can be computed from Eq. 2.3.6 as

$$p_{no} = \frac{n_i^2}{n_{no}} \approx \frac{n_i^2}{N_D} \tag{2.4.6}$$

and

$$n_{po} = \frac{n_i^2}{p_{po}} \approx \frac{n_i^2}{N_A} \,. \tag{2.4.7}$$

Using Eqs. 2.3.4 and 2.4.4 we can approximate the Fermi level of a highly-doped *n*-type semiconductor by

$$E_F = E_C - kT \log\left(\frac{N_C}{N_D}\right) \,. \tag{2.4.8}$$

Correspondingly, Eqs. 2.3.5 and 2.4.5 yield an approximation of the Fermi level of a strongly-doped *p*-type semiconductor:

$$E_F = E_V + kT \log\left(\frac{N_V}{N_A}\right) \,. \tag{2.4.9}$$

Hence, the Fermi level is near the conduction band edge for $N_D \approx N_C$ and near the valence band edge for $N_A \approx N_V$, as we noted before.

2.5 Current Densities

In the presence of external electric and magnetic fields the thermal equilibrium in the semiconductor is disturbed. The behavior of charged particles in such fields is described by the *Maxwell equations*. In normal semiconductor operation magnetic effects can be neglected. The most important consequence of the Maxwell equations, for our purposes, relates the charge density ρ (charge per volume) to the divergence of the electric field $\mathcal{E}$:

$$\nabla \cdot \mathcal{E} = \frac{\rho}{\varepsilon_s} \tag{2.5.1}$$

where ∇ is the Nabla operator[1], and $\varepsilon_s = \varepsilon_0 \varepsilon$ is the electrical permittivity of the semiconductor with $\varepsilon_0 = 8.85418 \times 10^{-12}$ F/m denoting the vacuum permittivity, and ε is the dielectric constant of the semiconductor. For silicon, $\varepsilon = 11.9$. This equation holds for homogeneous and isotropic materials under quasi-static conditions and is called the *Poisson equation*. The gradient of the

[1] $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$

electrostatic potential V is given by

$$\boxed{\nabla V = -\mathcal{E}\,.} \tag{2.5.2}$$

Hence, the Poisson equation can be rewritten as

$$\boxed{\Delta V = -\frac{\rho}{\varepsilon_s}} \tag{2.5.3}$$

where Δ is the Laplacian operator[2]. Since the potential energy of an electron is $-qV$, and the potential energy of a hole is qV, the differential spatial structure of the energy band edges can be computed from the electric field as

$$\nabla E_C = \nabla E_V = -q\nabla V = q\mathcal{E}\,. \tag{2.5.4}$$

The mobile charge carriers in a material that is not in thermal equilibrium give rise to current flow. The total current density $\mathbf{J}$ (charge flowing through a given cross-section during a given time interval) in a semiconductor is the sum of an electron current density $\mathbf{J}_n$ and a hole current density $\mathbf{J}_p$. By historical definition the electron is assigned the negative elementary charge $-q$, where we define q to be positive. However, the direction of the current density is defined as the direction of positive charge flow. Consequently, $\mathbf{J}_n$ is antiparallel to the electron flow and $\mathbf{J}_p$ is parallel to the hole flow. The average electron flow velocity can be expressed as

$$\mathbf{v}_n = -\frac{\mathbf{J}_n}{qn} \tag{2.5.5}$$

and the average hole flow velocity as

$$\mathbf{v}_p = \frac{\mathbf{J}_p}{qp}\,. \tag{2.5.6}$$

For each carrier type the current flow is due to two basic mechanisms, namely *diffusion* and *drift*. Diffusion is a term borrowed from gas dynamics. It describes the process by which a net particle flow is directed from a region of higher particle density to a region of lower particle density along the density gradient. This phenomenon is a direct consequence of the assumption of statistical isotropic motion of the particles. The electron and hole diffusion current

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

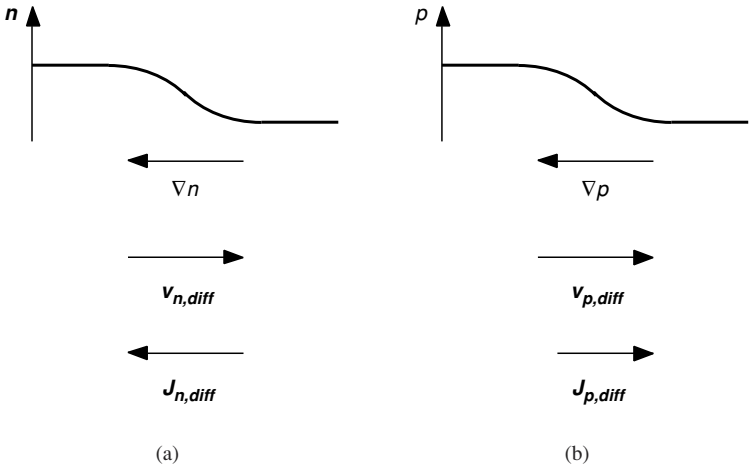


Figure 2.7
Diffusion of (a) electrons and (b) holes. The directions of the carrier concentration gradients, carrier motion, and electrical currents are shown.

densities are respectively given by

$$\mathbf{J}_{n,diff} = qD_n\nabla n \tag{2.5.7}$$

$$\mathbf{J}_{p,diff} = -qD_p\nabla p \tag{2.5.8}$$

where D_n and D_p are positive constants denoting the electron and hole *diffusion coefficient*, respectively. The average diffusion velocities are

$$\mathbf{v}_{n,diff} = -D_n\frac{\nabla n}{n} \tag{2.5.9}$$

$$\mathbf{v}_{p,diff} = -D_p\frac{\nabla p}{p}\,. \tag{2.5.10}$$

The relationships between the carrier concentrations, their gradients, the diffusion velocities, and the diffusion current densities are shown schematically in Fig. 2.7. As we shall see, diffusion determines the current flow in diodes and, within the operating range mainly considered in this book, in transistors. Diffusion also governs the ion flows in biological neurons.

Drift currents are caused by electric fields. For low electric fields the electron and hole drift current densities, respectively, are given by

$$\mathbf{J}_{n,drift} = q\mu_n n\mathcal{E} \tag{2.5.11}$$

$$\mathbf{J}_{p,drift} = q\mu_p p\mathcal{E} \tag{2.5.12}$$

where μ_n and μ_p are positive constants denoting the electron and hole *mobility*, respectively. The mobilities are the proportionality constants that relate the drift velocities of the charge carriers to the electric field according to

$$\mathbf{v}_{n,drift} = -\mu_n\mathcal{E} \tag{2.5.13}$$

$$\mathbf{v}_{p,drift} = \mu_p\mathcal{E}\,. \tag{2.5.14}$$

The mobilities decrease with increasing temperature as $\mu \propto T^{-n}$, where n=1.5 in theory, but empirically is found to be closer to n=2.5. The relationships between the different parameters are illustrated in Fig. 2.8. At sufficiently large electric fields the drift velocities saturate due to scattering effects and the term $\mu\mathcal{E}$ in the above equations must be replaced by a constant term $\mathbf{v_s}$, which is of the same order of magnitude as the thermal velocity. For intrinsic silicon at room temperature, approximate values of the mobilities are $\mu_n = 1500\,\mathrm{cm}^2/\mathrm{Vs}$ and $\mu_p = 450\,\mathrm{cm}^2/\mathrm{Vs}$, and the thermal velocity is 5×10^6 cm/s. Mobilities decrease with increasing impurity doping concentrations.

For non-degenerate semiconductors, there is a simple relation between diffusion constants and mobilities that was discovered by Einstein when he was studying Brownian motion, and is therefore known as the *Einstein relation*:

$$D_n = U_T\mu_n \tag{2.5.15}$$

$$D_p = U_T\mu_p \tag{2.5.16}$$

where $U_T = kT/q$ is the *thermal voltage*, and is the natural voltage scaling unit in the diffusion regime. Its value at room temperature is approximately 25 mV. From Eqs. 2.5.7−2.5.16 we then obtain the total electron and hole current densities

$$\mathbf{J}_n = \mathbf{J}_{n,drift} + \mathbf{J}_{n,diff} = q\mu_n(n\mathcal{E} + U_T\nabla n) \tag{2.5.17}$$

$$\mathbf{J}_p = \mathbf{J}_{p,drift} + \mathbf{J}_{p,diff} = q\mu_p(p\mathcal{E} - U_T\nabla p)\,. \tag{2.5.18}$$

In thermal equilibrium, diffusion and drift currents are balanced, that is $\mathbf{J}_n = \mathbf{J}_p = \mathbf{0}$ and the carrier concentration gradients can be computed by differentiating Eqs. 2.3.4 and 2.3.5. If we further use Eq. 2.5.4 to express the electric

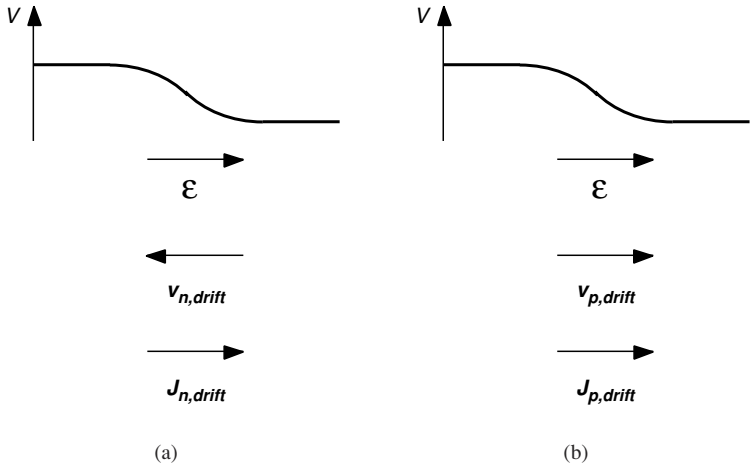


Figure 2.8
Drift of (a) electrons and (b) holes in an electrostatic potential V. The directions of the electric field $\mathcal{E}$, carrier motion, and electrical currents are shown.

field in terms of the gradient of the energy-band edges, we obtain the important result that in thermal equilibrium

$$\nabla E_F = 0\,. \tag{2.5.19}$$

That is, the Fermi level is constant. This result is intuitively clear, because otherwise a state of a given energy would more likely be occupied in one spatial position than in another. More mobile charge carriers would then move to this position than away from it, and so the energy states would be filled up until the probabilities would be matched everywhere.

The temporal dynamics of the the carrier density distributions are described by the *continuity equations*, which are a direct result of the Maxwell equations:

$$\frac{\partial n}{\partial t} = G_n - R_n + \frac{1}{q}\nabla \cdot \mathbf{J}_n \tag{2.5.20}$$

$$\frac{\partial p}{\partial t} = G_p - R_p - \frac{1}{q}\nabla \cdot \mathbf{J}_p\,. \tag{2.5.21}$$

where G_n and G_p denote the electron and hole *generation rate* and R_n and

R_p the electron and hole *recombination rate*, respectively. Generation and recombination effects account for the creation and annihilation of electron-hole pairs due to transitions between valence band and conduction band. Generation requires a certain amount of energy that can be supplied by thermal effects, optical excitation (discussed in Chapter 10) or impact ionization in high electric fields. Recombination counterbalances generation and is driven by the principle that a system tends towards a state of minimum energy. For a recombination process to take place an electron and a hole have to be present in close vicinity. Recombination is therefore limited by the availability of minority carriers. Approximations of the recombination rates under low injection conditions, where the majority carrier densities are much larger than the minority carrier densities are given by

$$R_n = \frac{n_p - n_{po}}{\tau_n} \tag{2.5.22}$$

$$R_p = \frac{p_n - p_{no}}{\tau_p} \tag{2.5.23}$$

where n_p and p_n are the minority carrier densities and n_{po} and p_{no} their values at thermal equilibrium. The minority carrier lifetimes τ_n and τ_p are equal if electrons and holes always recombine in pairs and no trapping effects occur.

2.6 *p-n* Junction Diode

The *p-n* junction diode is the fundamental semiconductor device. It is used as a basis for every transistor type. Furthermore, it is the dominant light-sensing device, and it will also become the most widely used sensor for electronic imaging applications. Light-sensing applications of diodes will be discussed in Chapter 10. The *p-n* junction has also found wide-spread use as the *light-emitting diode (LED)*. However, light-emitting diodes are very inefficient for semiconductors with indirect bandgaps, such as silicon, and will not be treated in this book.

Thermal Equilibrium

Consider what happens when an *n*-type semiconductor and a *p*-type semiconductor are brought into physical contact[3]. The diffusion processes described in Section 2.5 give rise to a net electron flow from the *n*-type region to the

3 In practice, surface oxidation of the semiconductor materials would prevent this.

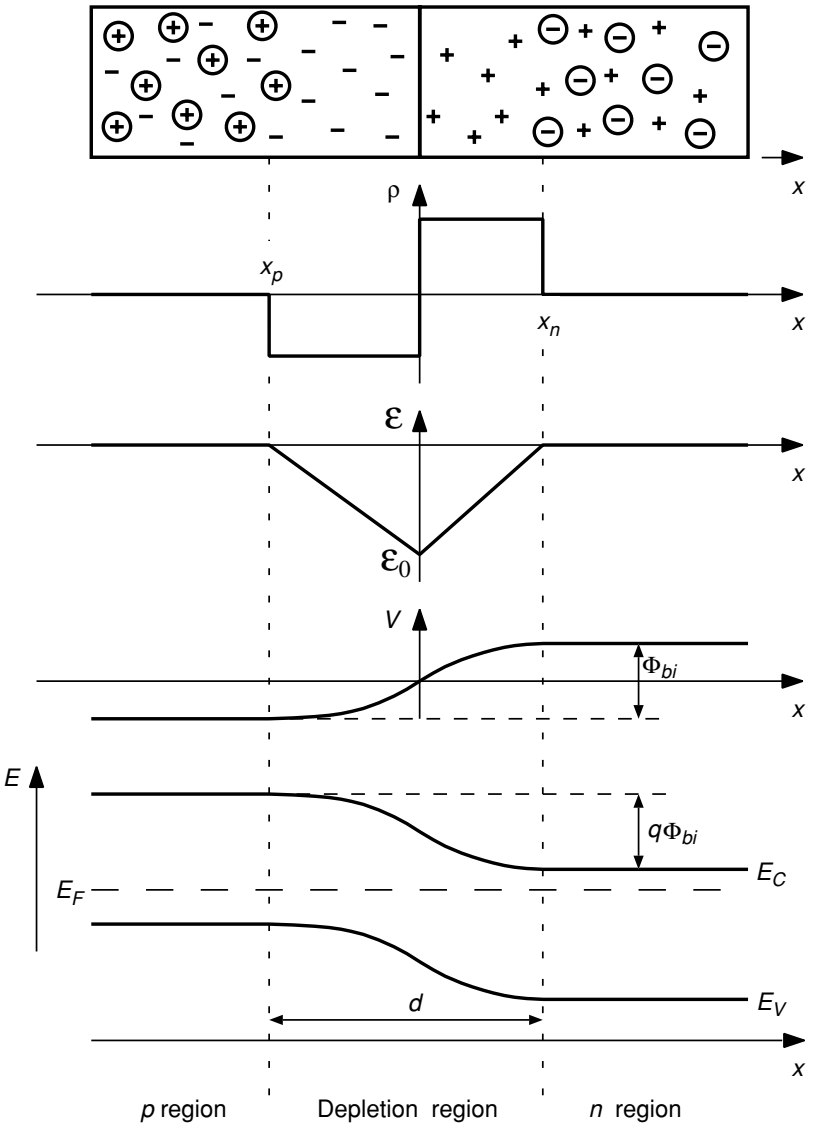


Figure 2.9
Characteristics of an abrupt *p-n* junction in thermal equilibrium with space-charge distribution ρ, electric field distribution $\mathcal{E}$, potential distribution V, and energy-band diagram E.

p-type region and a net hole flow from the *p*-type region to the *n*-type region. The combination of these two effects results in a diffusion current density $\mathbf{J}_{diff} = \mathbf{J}_{n,diff} + \mathbf{J}_{p,diff}$ from the *p*-type to the *n*-type region, as given by Eqs. 2.5.7 and 2.5.8. The diffusing minority carriers recombine with majority carriers in the vicinity of the junction. As a result, this diffusion region is largely devoid of mobile charge carriers and $np << n_i^2$. This region is therefore called a *depletion region*. This situation is schematically shown in the topmost graph of Fig. 2.9. While the *n*-type and *p*-type semiconductor are electrically neutral in isolation, this does not apply to the depletion region. On the *n*-type side of the junction the donor electrons are absent and the donors have a surplus proton; whereas on the *p*-type side the acceptor holes are filled with the donor electrons from the *n*-type side, and so have a surplus electron. Consequently, the depletion region is also called *space-charge region*. With a positive net charge on the *n*-type side and a negative net charge on the *p*-type side an electric field builds up in the depletion region that points from the *n*-type side to the *p*-type side, according to Eq. 2.5.1. This electric field in turn generates electron and hole drift currents with a combined current density of $\mathbf{J}_{drift} = \mathbf{J}_{n,drift} + \mathbf{J}_{p,drift}$ from the *n*-type to the *p*-type region, as given by Eqs. 2.5.11 and 2.5.12. In thermal equilibrium, diffusion and drift currents balance each other out and no net current flow is observed. Because of the electric field the electrostatic potential, and thus the energy band edges, vary across the space-charge region. Outside the space-charge region the energy bands flatten out, leaving a constant offset of the band edges between the neutral *n*-type and *p*-type regions. The electrical potential corresponding to this offset is called *diffusion potential* or *built-in potential* Φ_{bi}. Using the fact that the Fermi level is constant in thermal equilibrium (Eq. 2.5.19) and using Eqs. 2.3.4−2.3.7 we can compute the built-in potential to be

$$\Phi_{bi} = U_T \log\left(\frac{n_{no}p_{po}}{n_i^2}\right) = U_T \log\left(\frac{n_{no}}{n_{po}}\right) = U_T \log\left(\frac{p_{po}}{p_{no}}\right) . \qquad (2.6.1)$$

For highly-doped *n*-type and *p*-type regions Eqs. 2.4.4 and 2.4.5 are valid and the built-in voltage can also be expressed as

$$\Phi_{bi} = U_T \log\left(\frac{N_D N_A}{n_i^2}\right) . \qquad (2.6.2)$$

The simplest *p*-*n* junction to analyze is the *abrupt junction* where the *n*-type and *p*-type regions are each homogeneously doped and have a sharp boundary. This case is illustrated in Fig. 2.9. Assuming that the space-charge

region is fully depleted we obtain charge densities of

$$\rho_n = qN_D \tag{2.6.3}$$

$$\rho_p = -qN_A \tag{2.6.4}$$

within the depletion regions of the *n*-type and *p*-type material, respectively. The net charge density outside the depletion regions is zero, since the *n*-type and *p*-type bulks are electrically neutral. According to Eq. 2.5.1 the relationship between charge density distribution and electric field is given by

$$\frac{\partial \mathcal{E}_x(x)}{\partial x} = \frac{\rho(x)}{\varepsilon_s} \tag{2.6.5}$$

in the one-dimensional case, where x is the coordinate along an axis perpendicular to the junction plane with $\mathcal{E}_x$ pointing along that axis. Using the condition that the electric field is zero at the boundaries x_n and x_p of the depletion region in the *n*-type and *p*-type material, respectively; and the charge neutrality condition $N_D x_n = -N_A x_p$; Eq. 2.6.5 can be integrated to yield

$$\mathcal{E}_x(x) = \frac{qN_D(x - x_n)}{\varepsilon_s} = \mathcal{E}_0 + \frac{qN_D x}{\varepsilon_s} \tag{2.6.6}$$

in the *n*-type depletion region and

$$\mathcal{E}_x(x) = -\frac{qN_A(x - x_p)}{\varepsilon_s} = \mathcal{E}_0 - \frac{qN_A x}{\varepsilon_s} \tag{2.6.7}$$

in the *p*-type depletion region. Here

$$\mathcal{E}_0 = \mathcal{E}_x(x = 0) = -\frac{qN_D x_n}{\varepsilon_s} = \frac{qN_A x_p}{\varepsilon_s} \tag{2.6.8}$$

is the electric field at the junction; where it reaches its largest magnitude. The one-dimensional version of Eq. 2.5.2 states that the electric field is the partial derivative of the potential distribution along the x direction according to

$$\frac{\partial V(x)}{\partial x} = -\mathcal{E}_x(x)\,. \tag{2.6.9}$$

Since potentials are always measured with respect to a reference value the offset of the $V(x)$ curve is arbitrary. Choosing $V(x = 0) = 0$ we find

$$V(x) = -\mathcal{E}_0 x + \frac{\mathcal{E}_0}{2x_n} x^2 \tag{2.6.10}$$

in the *n*-type depletion region and

$$V(x) = -\mathcal{E}_0 x + \frac{\mathcal{E}_0}{2x_p} x^2 \tag{2.6.11}$$

in the *p*-type depletion region. The built-in potential can then be expressed in terms of the depletion region width

$$d = |x_n - x_p| \tag{2.6.12}$$

as

$$\Phi_{bi} = |V(x_n) - V(x_p)| = \frac{1}{2}\mathcal{E}_0 d\,. \tag{2.6.13}$$

Eliminating $\mathcal{E}_0$ from Eqs. 2.6.8 and 2.6.13 and solving for the depletion region width we obtain

$$d = \sqrt{\frac{2\varepsilon_s}{q}\frac{N_A + N_D}{N_A N_D}\Phi_{bi}}\,. \tag{2.6.14}$$

The *p*-*n* junctions fabricated with typical silicon processes are not abrupt, but have a more gradual profile. Their characteristics have to be determined numerically, but are qualitatively similar to those of the abrupt junction analyzed above.

Forward and Reverse Bias

Having characterized the *p*-*n* junction under thermal equilibrium conditions we now consider the cases where a net current flows through the diode. In practice, the most common way of generating such a current flow is by changing the boundary conditions for the *n*-type and *p*-type regions through the external application of a potential difference. If a positive voltage is applied to the *p*-type region relative to the *n*-type region the potential difference is called a *forward bias*, if the voltage applied to the *n*-type region is higher we speak of a *reverse bias*. Consequently, a current flowing from the *p*-type region to the *n*-type region is called *forward current* and a current flowing from the *n*-type region to the *p*-type region is called *reverse current*.

In steady state, the total current density must be constant throughout the diode. In the *n*-type and *p*-type bulk regions the current is made up of majority carriers. The electron current in the *n*-type region is thus transformed into a hole current in the *p*-type region. This transformation happens in the vicinity of the depletion region. The applied voltage V appears across the depletion region as a change in the built-in voltage and thus modifies the width of the

depletion region and the minority carrier densities outside the depletion region boundaries. The depletion region width can be computed from Eq. 2.6.14 by substituting Φ_{bi} with $\Phi_{bi} - V$, if we define V to be positive for a forward bias:

$$d = \sqrt{\frac{2\varepsilon_s}{q}\frac{N_A + N_D}{N_A N_D}(\Phi_{bi} - V)}\,. \tag{2.6.15}$$

According to the Boltzmann distribution (Eq. 2.3.2) the minority carrier densities grow exponentially with decreasing electrostatic potential, so that outside the depletion region boundaries they become

$$n_p = n_{po} e^{V/U_T} \tag{2.6.16}$$

$$p_n = p_{no} e^{V/U_T}\,. \tag{2.6.17}$$

Since the majority carrier distributions are approximately constant throughout the neutral regions, the np product is now given by

$$np = n_i^2 e^{V/U_T} \tag{2.6.18}$$

at the depletion region boundaries. The probability distributions for the occupancy of a given energy state are now centered around the so-called *quasi-Fermi levels* $q\Phi_n$ and $q\Phi_p$, where

$$V = \Phi_p - \Phi_n \tag{2.6.19}$$

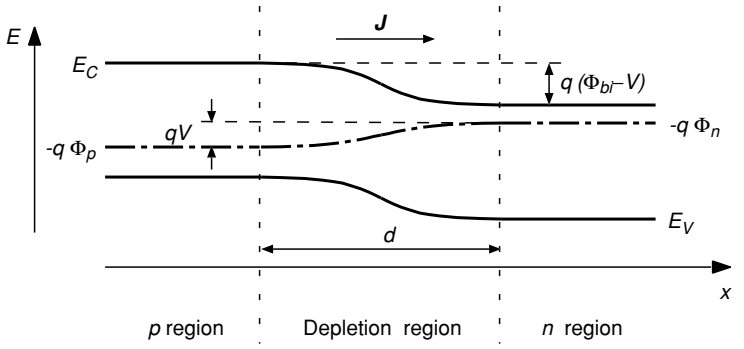
at the depletion region boundaries. The same argument that led to Eq. 2.5.19 for the thermal-equilibrium case now gives

$$\mathbf{J_n} = -q\mu_n n \nabla \Phi_n \tag{2.6.20}$$

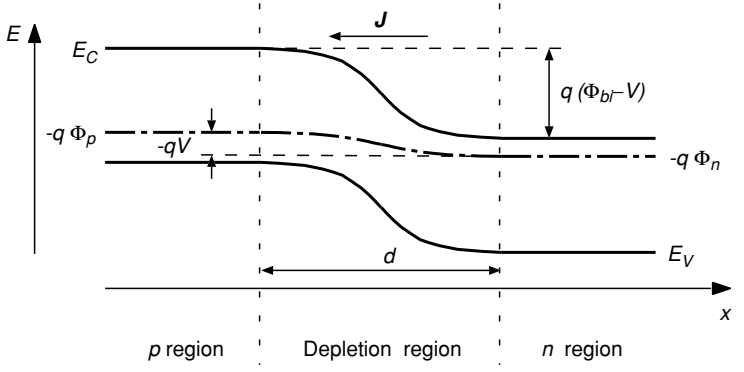
$$\mathbf{J_p} = -q\mu_p p \nabla \Phi_p\,. \tag{2.6.21}$$

The current densities are therefore proportional to the gradients of the quasi-Fermi levels. Outside the depletion region the electric field is small, as in the thermal-equilibrium case, and the current flows mainly by diffusion. Within the depletion region the concentration of mobile charge carriers is very low, and therefore no significant recombination effects take place there. Consequently, the electron and hole currents are almost constant throughout the depletion region. The energy band diagrams for the two different biasing conditions are shown in Fig. 2.10 and the carrier distributions and current densities in Fig. 2.11.

A forward bias diminishes the potential step across the junction. As a result, the minority carrier concentration and thus the np product on either side



(a)



(b)

Figure 2.10
Energy band diagram of a *p-n* junction diode for (a) forward bias, and (b) reverse bias.

of the depletion region are increased. This condition leads to an increase in the recombination rate outside the depletion region boundaries and thus to a minority carrier gradient that gives rise to a forward diffusion current. Since the minority carrier densities at the depletion region boundaries increase exponentially with applied forward bias (Eqs. 2.6.16 and 2.6.17) the recombination rate, and therefore the forward current density, increase exponentially.

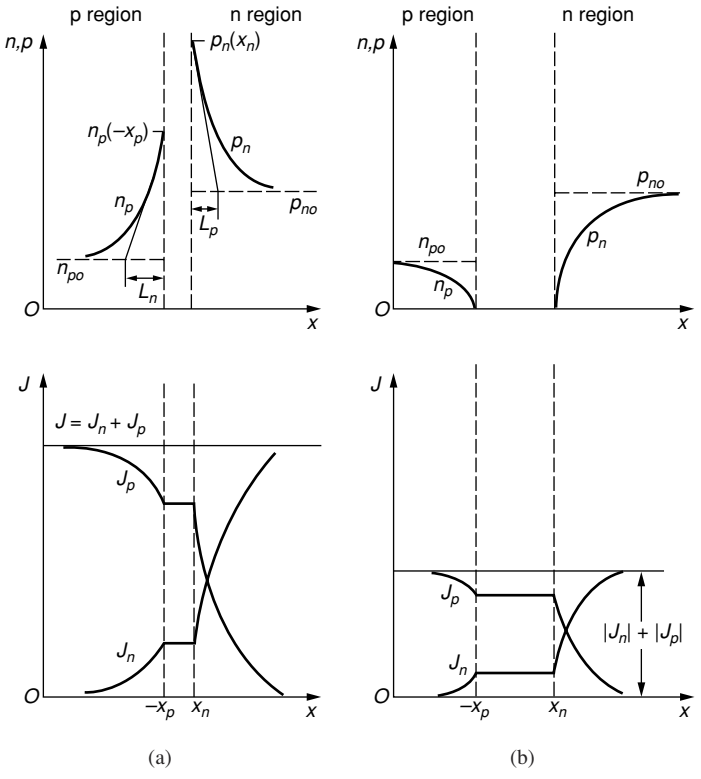


Figure 2.11
Minority carrier distributions and current densities in the vicinity of a *p-n* junction for (a) forward bias, and (b) reverse bias. Figure adapted from S. M. Sze (1981), Physics of Semiconductor Devices, 2nd Edition.

A reverse bias increases the potential step across the junction. The minority carrier concentrations, and the np products on both sides of the depletion region are decreased and therefore the recombination rate is decreased. The thermal generation rate now exceeds the recombination rate near the depletion region boundaries. This condition results in a small minority carrier gradient pointing away from the junction, and thus a small reverse diffusion current density occurs.

An approximation of the diode current-voltage relationship based on the above considerations is given by the *Shockley equation*:

$$\boxed{J = J_n + J_p = J_s \left(e^{V/U_T} - 1 \right)} \tag{2.6.22}$$

with

$$J_s = \frac{qD_n n_{po}}{L_n} + \frac{qD_p p_{no}}{L_p} \tag{2.6.23}$$

where $L_n = \sqrt{D_n \tau_n}$ is called the *electron diffusion length* and $L_p = \sqrt{D_p \tau_p}$ the *hole diffusion length*. This current-voltage relationship is illustrated in Fig. 2.12(a). Since the reverse current density is limited by J_s, which is much smaller than forward current densities at forward biases larger than about $4U_T$, diodes are often used as *rectifiers* with a large conductivity in the forward direction and a small conductivity in the reverse direction. This application is reflected in the circuit symbol for the diode (Fig. 2.12(b)), which is resembles an arrow pointing in the forward current direction.

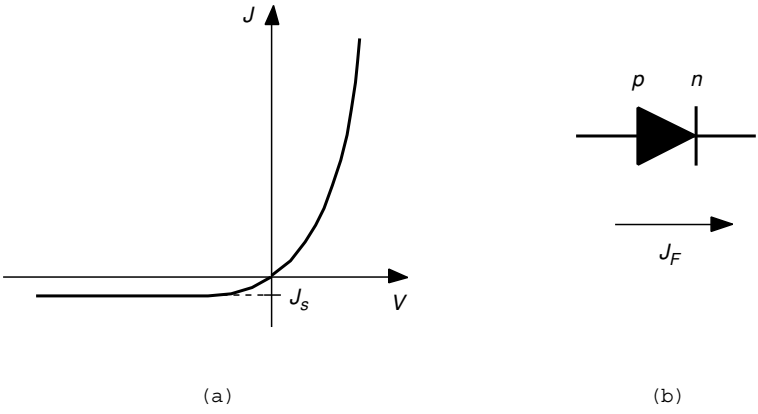


Figure 2.12
The *p-n* junction diode. (a) Current-voltage characteristic of an ideal diode according to the Shockley approximation. (b) Diode symbol; the arrow indicates the direction of a forward current density J_F.

The Shockley equation is derived from the diffusion current density equations 2.5.7 and 2.5.8, the continuity equations 2.5.20 and 2.5.21, as well as Eqs. 2.5.22 and 2.5.23 for the recombination rates. The underlying assump-

tions (*Shockley approximation*) are: Abrupt depletion layer boundaries; the validity of the Boltzmann approximation given by Eq. 2.3.2, and of the low-injection condition; negligible generation current in the depletion layer; and constant electron and hole currents within the depletion layer.

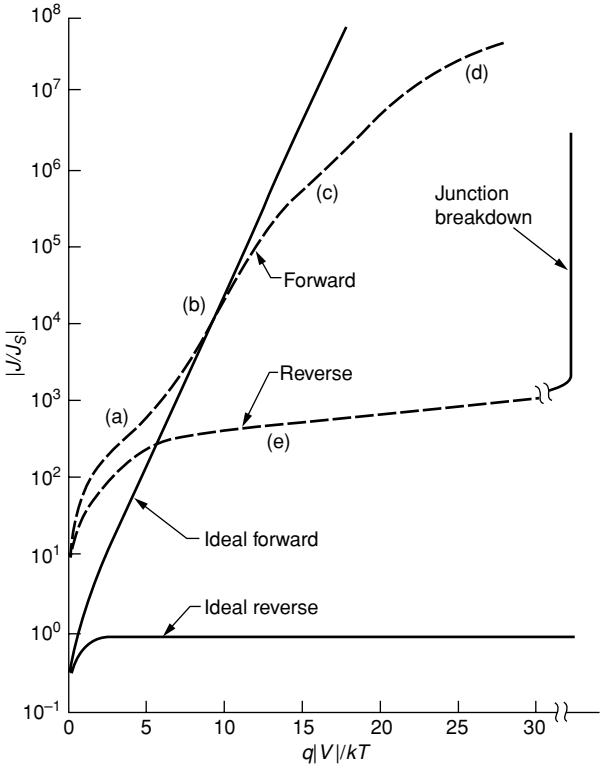


Figure 2.13
Comparison of the current-voltage characteristics of an ideal and a practical diode. (a) Generation-recombination current domain. (b) Diffusion current domain. (c) High-injection domain. (d) Series-resistance effect. (e) Reverse leakage current due to generation-recombination and surface effects. Figure adapted from J. L. Moll (1958), The evolution of the theory of the current-voltage characteristics of *p-n* junctions, *Proc. IRE*, **46**, 1076.

For silicon *p-n* junctions there is only a qualitative agreement between the observed behavior and the Shockley equation 2.6.22, because the above approximations are not completely justified; and because of surface effects at the semiconductor boundaries. Figure 2.13 compares the current-voltage

characteristics of an ideal diode and a real diode. In semiconductors with small intrinsic carrier concentrations n_i, such as silicon, the reverse diffusion current density (given by J_s for reverse biases larger than approximately $4U_T$) may be dominated by a superimposed reverse generation current density J_{gen}. The generation current is mainly due to *trapping centers* in the depletion region. Trapping centers are imperfections in the crystal, which capture and release mobile charge carriers. The generation current density due to trapping is given by

$$J_{gen} = \frac{qn_i d}{\tau_e} \tag{2.6.24}$$

where τ_e is the effective lifetime of the trapping. Similarly, under forward bias conditions there is a recombination current density component due to carrier capture processes mainly in the depletion region that exhibits an exponential behavior

$$J_{rec} \sim e^{V/2U_T} . \tag{2.6.25}$$

Empirically, the total forward current density can be fit with the function

$$J_F \sim e^{V/nU_T} \tag{2.6.26}$$

where n is a number between 1 and 2, depending on which current density component dominates.

For large forward biases, where the minority carrier concentrations approach the majority carrier concentrations near the depletion region boundaries, part of the applied voltage appears as linear potential drops outside the depletion region, which with increasing forward bias start to extend more and more into the semiconductor between the diode terminals. In this domain, the forward current-voltage characteristic is subexponential and finally asymptotes to a linear behavior given by the series resistance of the bulk regions.

For large reverse biases, a phenomenon called *junction breakdown* occurs that expresses itself in a sudden increase of reverse current at a certain reverse voltage. For silicon with typical impurity doping concentrations this effect is due to impact ionization: The generation of electron-hole pairs by collision with an electron or hole that has acquired sufficient kinetic energy in the electric field of the depletion region. A charge carrier may create multiple electron-hole pairs during its transition through the depletion region. The generated carriers can in turn create electron-hole pairs if they acquire sufficient energy, and so on. This effect is known as *avalanche multiplication*. It is characterized by a sharp onset and a high gain with respect to a reverse voltage change.

2.7 The Metal-Insulator-Semiconductor Structure

As its name implies, the **M**etal-**I**nsulator-**S**emiconductor (MIS) structure consists of a conductor and a semiconductor that are separated by a thin insulator layer. The most common version of the MIS structure is the **M**etal-**O**xide-**S**ilicon (MOS) structure, where the 'oxide' is in most cases silicon dioxide (SiO_2). The MOS structure and the *p-n* junction diode are the building blocks of today's most widely used type of transistor, commonly known as **M**etal-**O**xide-**S**emiconductor **F**ield-**E**ffect **T**ransistor (MOSFET) . In addition, the MOS structure is the basic building block of the **C**harge-**C**oupled **D**evice (CCD), which is currently the most widely used device for electronic imaging applications, and which is presented in Section 10.5.

Operation Domains

In a typical MIS structure the insulator layer is sufficiently thick that it cannot be crossed by charge carriers under normal operating conditions and sufficiently thin that the charge on the conductor can influence the charge distribution in the semiconductor via the electrostatic potential it induces. A positive charge on the conductor attracts mobile electrons from the semiconductor to the semiconductor-insulator interface and repels mobile holes away from the interface. Conversely, a negative charge on the conductor attracts holes and repels electrons. If the semiconductor is *n*-type, a positive charge on the conductor increases the majority carrier density near the semiconductor surface, an effect known as *accumulation*, while a negative charge on the conductor reduces the majority carrier density. With increasing negative charge on the conductor, most majority carriers are driven from the region near the surface, resulting in *depletion*, and eventually minority carriers start to accumulate at the semiconductor surface, an effect called *inversion*. The same effects are observed in *p*-type semiconductors, if the sign of the charge on the conductor is reversed.

The energy-band diagram is a helpful tool to visualize these effects. In order to be able to compare the energy levels and potentials in the conductor and the semiconductor it is helpful to define a few more parameters, as shown in the band diagram of Fig. 2.14 for the case of a *p*-type semiconductor. The basic concept is that of the *work function*, which is defined as the energy difference of an electron between the Fermi level in the material and the vacuum level in free space. The work function is denoted by $q\phi_m$, where ϕ_m is the electrostatic potential difference corresponding to the work function. In the

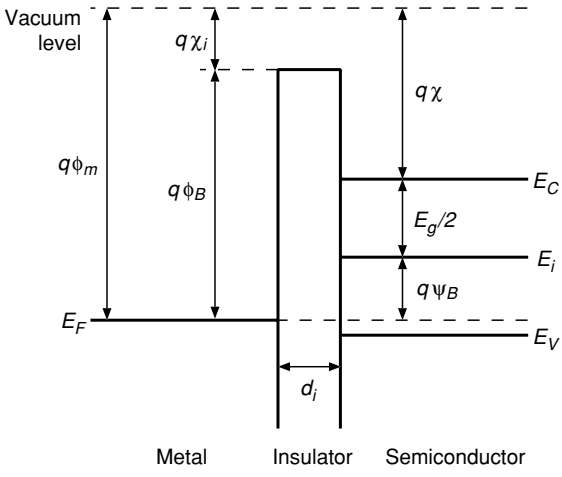


Figure 2.14
Energy-band diagram of an ideal MIS diode with no applied bias between the semiconductor and metal for a *p*-type semiconductor.

same context the *electron affinity* is defined as the energy difference between the bottom of the conduction band in the semiconductor or the insulator and the vacuum level. For the semiconductor we denote the electron affinity by $q\chi$, for the insulator by $q\chi_i$. Furthermore, the potential difference between the Fermi level in the metal and the insulator conduction-band edge is denoted by ϕ_B. The potential difference separating the Fermi level E_F and the intrinsic Fermi level E_i of the semiconductor can be computed from Eqs. 2.3.7, 2.3.8, and 2.4.9 as

$$\psi_B = U_T \log\left(\frac{N_A}{n_i}\right) . \tag{2.7.1}$$

An MIS diode is called ideal if it has the following properties: Firstly, when there is no applied bias, the work functions of the semiconductor and the metal are equal: the Fermi levels line up and the energy bands in the semiconductor are flat (*flat-band condition*). Secondly, the charge on the conductor plate is equal to the total charge in the semiconductor with opposite sign. Finally, the insulator is neither charged nor permeable to charge carriers. Note that for certain applications deviations from this ideal behavior may be desirable, as we will see in later chapters.

Steady-State Analysis

With the above terminology and assumptions we can now explain the effects of accumulation, depletion, and inversion with the bending of the semiconductor energy bands near the semiconductor-insulator interface, as shown in Fig. 2.15 for a *p*-type semiconductor. In thermal equilibrium, the semiconductor Fermi level is constant and separated from the conductor Fermi level by the energy corresponding to the applied potential difference. Inversion occurs when the intrinsic Fermi level crosses the Fermi level near the surface, corresponding to the situation where the minority carrier density is larger than the majority carrier density at the surface.

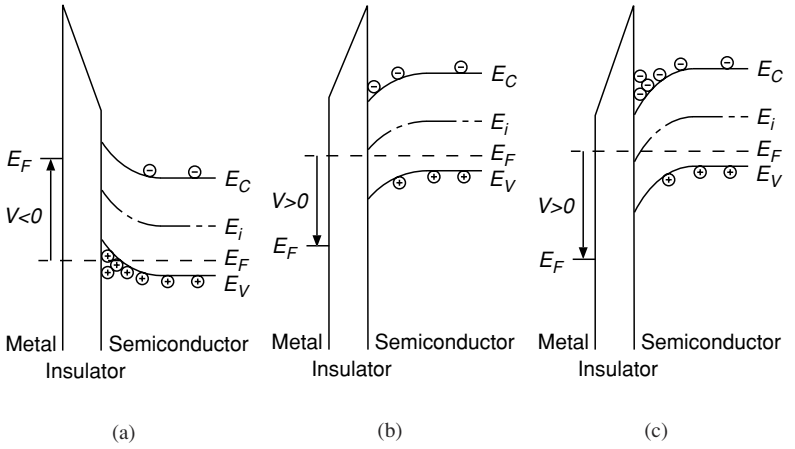


Figure 2.15
Energy-band diagrams of an ideal MIS diode with applied bias for a *p*-type semiconductor in (a) accumulation, (b) depletion, and (c) inversion.

For further analysis we define an electrostatic potential ψ corresponding to the difference between the local intrinsic Fermi level and the intrinsic Fermi level in the bulk. The value of this potential at the semiconductor surface is called *surface potential* ψ_s. For a *p*-type semiconductor, $\psi_s < 0$ corresponds to accumulation; $\psi_s = 0$ to the flat-band condition; $0 < \psi_s < \psi_B$ to depletion; and $\psi_B < \psi_s$ to inversion. The potential distribution and therefore the bending of the energy bands can be computed from the space-charge distribution using

the Poisson equation

$$\frac{d^2\psi}{dx^2} = -\frac{\rho(x)}{\varepsilon_s} \tag{2.7.2}$$

where the x axis is perpendicular to the semiconductor surface. The following analysis will only be made for a p-type semiconductor. For an n-type semiconductor the results are the same if the symbols p and n are interchanged, and the signs are appropriately changed. Using the flat-band charge-neutrality condition the space-charge density can be expressed as

$$\rho(x) = q(N_D^+ - N_A^- + p_p - n_p) = q(n_{po} - p_{po} + p_p - n_p) \tag{2.7.3}$$

where N_D^+ and N_A^- are the densities of ionized donors and acceptors, respectively. The carrier concentrations are given by

$$n_p = n_{po} e^{\psi/U_T} \tag{2.7.4}$$

$$p_p = p_{po} e^{-\psi/U_T} \,. \tag{2.7.5}$$

The Poisson equation can then be rewritten as

$$\frac{\partial^2\psi}{\partial x^2} = -\frac{q}{\varepsilon_s}\left(p_{po}\left(e^{-\psi/U_T} - 1\right) - n_{po}\left(e^{\psi/U_T} - 1\right)\right) \,. \tag{2.7.6}$$

Integration of this equation leads to

$$\mathcal{E}_x(x) = -\frac{\partial\psi}{\partial x} \tag{2.7.7}$$

$$= \pm\frac{\sqrt{2}U_T}{L_D}\sqrt{e^{-\psi/U_T} + \frac{\psi}{U_T} - 1 + \frac{n_{po}}{p_{po}}\left(e^{\psi/U_T} - \frac{\psi}{U_T} - 1\right)} \tag{2.7.8}$$

where $L_D = \sqrt{\varepsilon_s U_T/qp_{po}}$, and the electric field has the same sign as the potential. Integrating Eq. 2.6.5 from the bulk to the surface we can now define an area charge Q_s as the charge underneath a unit area of semiconductor surface and relate it to the surface potential using Eq. 2.7.7:

$$Q_s = -\varepsilon_s \mathcal{E}_s \tag{2.7.9}$$

$$= \mp\frac{\sqrt{2}\varepsilon_s U_T}{L_D}\sqrt{e^{-\psi_s/U_T} + \frac{\psi_s}{U_T} - 1 + \frac{n_{po}}{p_{po}}\left(e^{\psi_s/U_T} - \frac{\psi_s}{U_T} - 1\right)} \tag{2.7.10}$$

where $\mathcal{E}_s$ is the electric field at the surface. This relationship between area

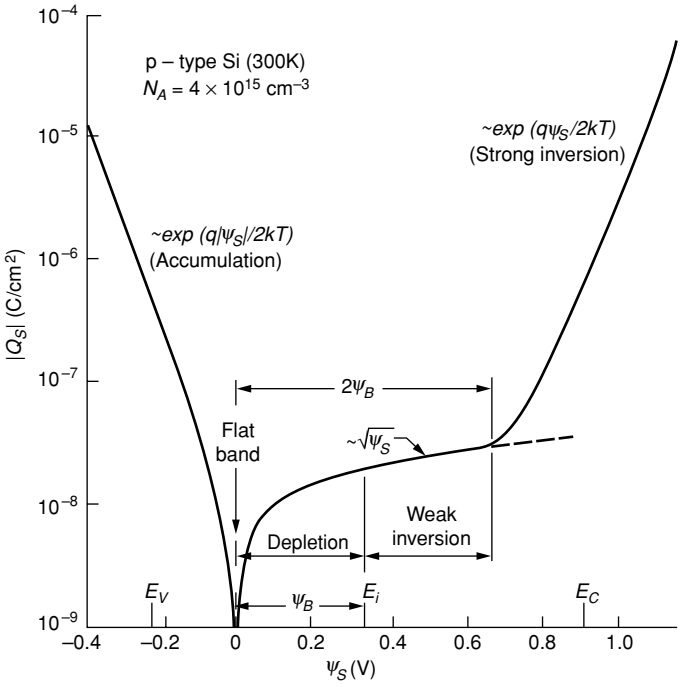


Figure 2.16
Dependence of the area charge Q_s on the surface potential for *p*-type silicon with acceptor density $N_A = 4 \times 10^{15}$ cm^{-3} at room temperature. Figure adapted from S. M. Sze (1981), Physics of Semiconductor Devices, 2nd Edition. ©1981 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

charge in the semiconductor and surface potential is plotted in Fig. 2.16. The different domains shown in Fig. 2.15 can be distinguished by the different dependencies of Q_s on ψ_s. In accumulation ($\psi_s < 0$) the first term under the square root dominates, and we obtain $Q_s \sim e^{-\psi_s/2U_T}$. In the flat-band situation ($\psi_s = 0$), $Q_s = 0$. In depletion ($0 < \psi_s < \psi_B$) the second term dominates, and $Q_s \sim -\sqrt{\psi_s/U_T}$. According to these characteristics the inversion domain is separated into two distinct sub-domains; *weak inversion* ($\psi_B < \psi_s < 2\psi_B$) which exhibits the same relationship between Q_s and ψ_s as depletion; and *strong inversion* ($\psi_s \gg 2\psi_B$) with $Q_s \sim -e^{\psi_s/2U_T}$. The transition region between weak and strong inversion ($\psi_s \geq 2\psi_B$) is called

moderate inversion. Strong inversion shows a relatively abrupt onset at

$$\psi_s \approx 2\psi_B \,. \tag{2.7.11}$$

In the case of inversion, the area charge consists of a contribution by mobile electrons close to the surface, denoted by Q_i, and a contribution by ionized acceptors in the depletion region, Q_d, and is equal with opposite sign to the charge per unit area on the conductor plate, Q_g:

$$Q_g = -Q_s = -Q_i - Q_d \tag{2.7.12}$$

where

$$Q_d = -qN_A d \tag{2.7.13}$$

and d is the depletion region width. Figure 2.17 shows the distributions of area charge, electric field and potential in the ideal MIS diode in inversion with an externally applied potential difference V, under the assumption that all mobile charge accumulates at the surface. Since the insulator is assumed to be neutral, the electric field is constant and the potential decreases linearly within the insulator. The total potential drop through the insulator is given by

$$V_i = \mathcal{E}_{xi} d_i = -\frac{Q_s d_i}{\varepsilon_i} = -\frac{Q_s}{C_i} \tag{2.7.14}$$

where E_{xi} denotes the electric field in the insulator, d_i the thickness of the insulator, ε_i the permittivity of the insulator, and

$$C_i = \varepsilon_i / d_i \tag{2.7.15}$$

is the insulator capacitance per unit area. The applied voltage is the sum of the voltage drop across the insulator and the surface potential:

$$V = V_i + \psi_s \,. \tag{2.7.16}$$

In the depletion and weak inversion case, the mobile charge can be neglected and the space-charge density is given by

$$\rho = -qN_A \,. \tag{2.7.17}$$

The potential distribution can be obtained as a function of the depletion region width using Eq. 2.7.2:

$$\psi = \psi_s (1 - \frac{x}{d})^2 \tag{2.7.18}$$

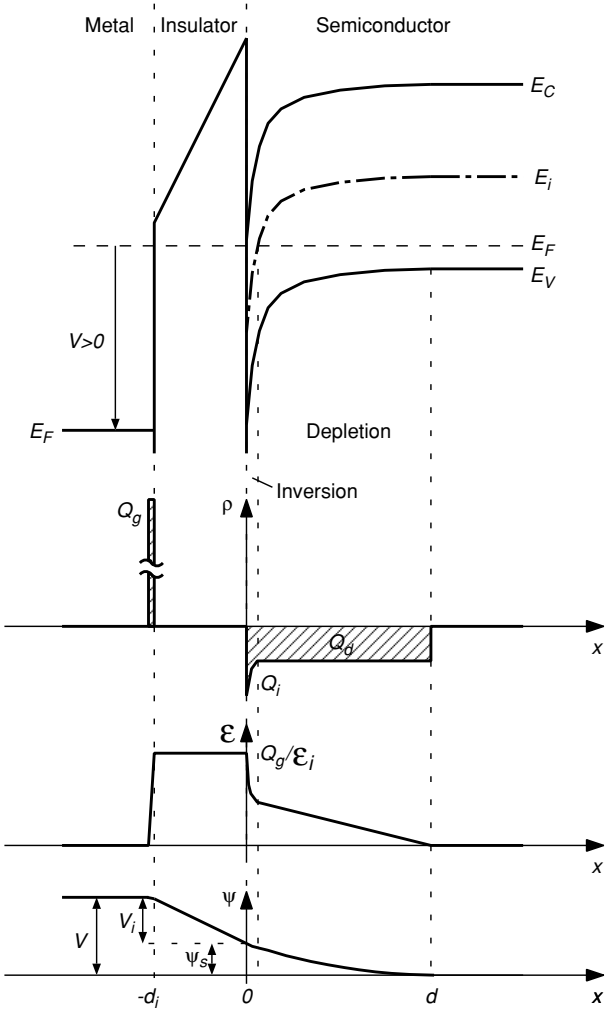


Figure 2.17
Ideal MIS diode with applied bias for a *p*-type semiconductor in inversion: Energy-band diagram, area charge distribution ρ, electric field distribution $\mathcal{E}$, and potential distribution ψ.

with

$$\psi_s = \frac{qN_A d^2}{2\varepsilon_s} . \tag{2.7.19}$$

The depletion region width reaches its maximum at the onset of strong inver-

sion, which can be computed from Eqs. 2.7.11 and 2.7.19 as

$$d_{max} = \sqrt{\frac{4\varepsilon_s \psi_B}{qN_A}}\,. \tag{2.7.20}$$

The minimum voltage that has to be applied to the MIS structure to obtain strong inversion is called *threshold voltage*. With the approximation that $Q_i \ll Q_d$ at the onset of strong inversion and using Eqs. 2.7.11, 2.7.13, 2.7.14, 2.7.16, and 2.7.20 the threshold voltage can be estimated to be

$$V_T \approx -\frac{Q_d}{C_i} + 2\psi_B = \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_i} + 2\psi_B\,. \tag{2.7.21}$$

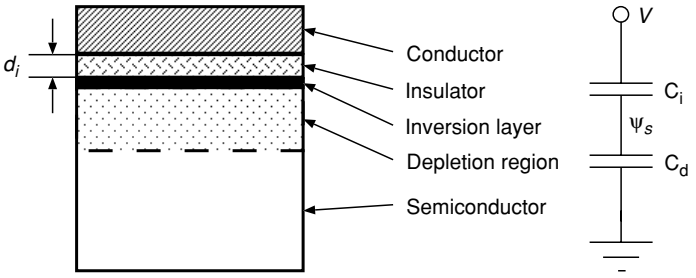


Figure 2.18
MIS diode in inversion with equivalent capacitive-divider circuit.

MIS Capacitance

Because the depletion layer does not contain any mobile charge, it can be regarded as a capacitor and assigned an incremental capacitance per unit area which is obtained by differentiating Eq. 2.7.9:

$$C_d = -\frac{\partial Q_s}{\partial \psi_s} \tag{2.7.22}$$

$$= \pm\frac{\varepsilon_s}{\sqrt{2}L_D}\frac{1 - e^{-\psi_s/U_T} + \frac{n_{po}}{p_{po}}\left(e^{\psi_s/U_T} - 1\right)}{\sqrt{e^{-\psi_s/U_T} + \frac{\psi_s}{U_T} - 1 + \frac{n_{po}}{p_{po}}\left(e^{\psi_s/U_T} - \frac{\psi_s}{U_T} - 1\right)}}\,. \tag{2.7.23}$$

In the case of depletion and weak inversion this equation reduces to

$$C_d(0 < \psi_s < 2\psi_B) \approx \sqrt{\frac{qN_A\varepsilon_s}{2\psi_s}} = \frac{\varepsilon_s}{d} \tag{2.7.24}$$

and in the flat-band case to

$$C_d(\psi_s = 0) \approx \frac{\varepsilon_s}{L_D}\,. \tag{2.7.25}$$

The capacitances C_i and C_d are connected in series, as shown in Fig. 2.18, so that the MIS diode has a total incremental capacitance of

$$C = \frac{C_i C_d}{C_i + C_d}\,. \tag{2.7.26}$$

Since C_i is constant, it can be used to normalize C in terms of the *capacitive divider ratio*

$$\frac{C}{C_i} = \frac{C_d}{C_i + C_d}\,. \tag{2.7.27}$$

The dependence of the incremental capacitive divider ratio on the surface potential, as given by Eq. 2.7.22 is shown in Fig. 2.19 (curve (a)). The flat-band capacitance is obtained from Eqs. 2.7.15, 2.7.25, and 2.7.26 as

$$C(\psi_s = 0) = \frac{\varepsilon_i}{d_i + \frac{\varepsilon_i}{\varepsilon_s} L_D}\,. \tag{2.7.28}$$

In accumulation and in strong inversion Q_s depends exponentially on ψ_s and $C_d \gg C_i$, while in depletion and weak inversion C_d is comparable to C_i. Eq. 2.7.22 and curve (a) of Fig. 2.19 describe the system in thermal equilibrium. If the applied voltage is modulated in time, however, the characteristic of the capacitance C_d depends on the ability of the charge distribution to follow the signal. In the case of inversion, this ability is already impaired for relatively slow signals, because the inversion charge at the semiconductor surface is electrically almost insulated and its concentration can only be changed via the normal generation-recombination process. At large enough frequencies, the inversion charge thus remains practically constant and only the majority carriers in the substrate follow the signal. The incremental depletion capacitance C_d then stays at a value comparable to C_i and continues to determine the capacitance-voltage characteristic into the range beyond the threshold voltage. For typical silicon MIS structures, the change between the two characteristics occurs at frequencies between 5 Hz and 100 Hz. Around the threshold of inversion, the depletion region width reaches a maximum value and becomes

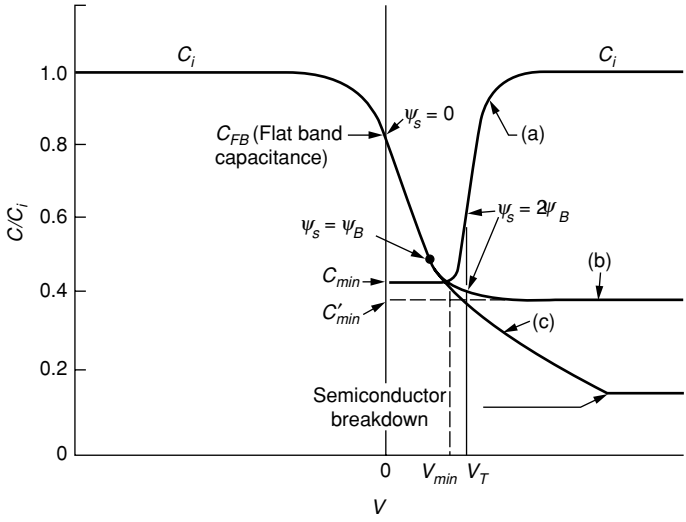


Figure 2.19
Capacitance-voltage characteristics of an ideal MIS diode. (a) Low frequency. (b) High frequency. (c) Deep depletion. Figure adapted from A. S. Grove et al. (1965), Investigation of thermally oxidised silicon surfaces using metal-oxide-semiconductor structures, *Solid-State Electronics*, **8**, 145-163. ©1965, with permission from Elsevier Science.

independent of the applied voltage for larger voltages. The depletion capacitance C_d is therefore also independent of the voltage (curve (b) in Fig. 2.19). If the voltage is switched rapidly beyond threshold from the accumulation domain, such that there is no time for inversion charge to build up at the surface, the depletion region width grows beyond the limits normally set by the inversion threshold. The depletion region width as a function of the applied voltage then resembles that of a reverse-biased *p-n* junction diode. This condition is called *deep depletion* and is shown by curve (c) in Fig. 2.19. The deep depletion domain is important for the operation of charge-coupled devices, which will be presented in Section 10.5.

The relationship between the applied potential and the surface potential is given by the coupling factor

$$\boxed{\kappa = \frac{\partial \psi_s}{\partial V}\,.} \tag{2.7.29}$$

With the help of Eqs. 2.7.14, 2.7.16, 2.7.22, and 2.7.26 κ can be expressed as

$$\boxed{\kappa = \frac{C_i}{C_i + C_d} = \frac{C}{C_d} = 1 - \frac{C}{C_i}\,,} \tag{2.7.30}$$

which is the incremental capacitive divider ratio as seen from the conductor side. The coupling factor κ appears as a parameter in the basic equations describing the operation of MOSFETs, where it will be called the *subthreshold slope factor*, and will be used extensively in this book. In the small-signal analysis of MOSFET circuits κ is assumed to be constant, but it must be kept in mind that κ varies quite strongly with applied voltage, as can be seen from Fig. 2.19.

Parasitic Charges

We conclude this chapter by noting that in MOS processing technologies the thin silicon dioxide layer is obtained by thermal oxidation of the silicon surface. A consequence of this fabrication process is the creation of traps and charges that must be taken into account in models of device behavior. There are several types of such traps and charges that can be distinguished according to their location either at the interface or within the oxide and according to whether they are fixed or mobile. The most important effect of these charges is a shift in the capacitance-voltage characteristics.

Typically, fixed positive charges are found close to the interface between the silicon substrate and the oxide. They are not changed by variation of ψ_s. Electrically neutral defects in the silicon dioxide, called *oxide traps*, can be charged or discharged by electrons and holes introduced into the oxide. In the presence of such charges an external voltage must be applied to the MOS diode to establish the flat-band condition. This voltage is called the *flat-band voltage*, V_{fb}. The relationship between applied voltage and surface potential is then modified from Eq. 2.7.16 to become

$$V = V_i + \psi_s + V_{fb} \tag{2.7.31}$$

and the expression for the threshold voltage (Eq. 2.7.21) is changed to

$$V_T \approx -\frac{Q_d}{C_i} + 2\psi_B + V_{fb} = \frac{\sqrt{4\varepsilon_s q N_A \psi_B}}{C_i} + 2\psi_B + V_{fb}\,. \tag{2.7.32}$$

This page intentionally left blank

3 MOSFET Characteristics

The two most common devices used in today's integrated circuit technology are the *Metal-Oxide-Silicon Field Effect Transistor* (MOSFET)[1] and the bipolar junction transistor (BJT)[2]. The currents in these devices comprise either positively-charged holes, negatively-charged electrons, or both holes and electrons. The BJT is called a *bipolar* device because the current in the transistor consists of both types of carriers, electrons and holes. The MOSFET[3] is called a *unipolar* device because the current has only one type of carrier, either holes or electrons.

In this book, we concentrate on MOSFETs and their current-voltage characteristics in the *subthreshold* (also known as *weak inversion*) domain. We use the transistor in this domain because the current here is exponentially dependent on the control voltages of the MOSFET just as the ionic conductances of a neuron are exponentially dependent on its membrane potential. Although the current in a BJT is also exponentially dependent on its control voltages, we only use BJTs when there is a requirement for higher current drive, and for lower offsets between transistors. Furthermore when MOSFETs are operated in the *subthreshold* domain, they draw small currents so power consumption is reduced.

We also describe to some extent the current-voltage characteristics of the MOSFET in the *above threshold* domain. There are numerous texts and papers that cover the transistor's characteristics in this domain (Weste and Eshraghian, 1994; Ismail and Fiez, 1994; Tsividis, 1996; Johns and Martin, 1997; Tsividis, 1998; Gray et al., 2001). There are a few sources in which the operation of the transistor in the subthreshold domain is described in detail (Mead, 1989; Maher, 1989; Andreou et al., 1991; Andreou and Boahen, 1994; Enz et al., 1995; Enz and Vittoz, 1997; Tsividis, 1998). We start by describing the MOSFET structure and the biasing necessary for the different modes of

1 The field-effect transistor structure was first described in a series of patents by J. Lilienfeld that were granted in the early 1930s. The MOSFET is the field-effect transistor type that is almost exclusively used today. Historically, other field-effect transistor types were invented including the junction field-effect transistor (JFET), and the metal-semiconductor field-effect transistor (MESFET).

2 The *pn* junction field-effect and the bipolar transistor were invented by Bardeen, Brattain, and Shockley and their colleagues at Bell Telephone Laboratories during 1947–1952. Even though the FET was conceived earlier than the BJT, the latter was the first to be mass produced.

3 This device is also called a MOST (MOS transistor) or an IGFET (insulated gate field-effect transistor).

operation. We then derive the current-voltage characteristics of the MOSFET in these different regimes using the equations derived for the pn junction in Chapter 2. We also look at a small-signal MOSFET model and some second-order effects that affect the operation of the transistor. Other issues such as noise and transistor mismatch are discussed in Section 3.6, and in great detail in Chapter 11 and Chapter 13 respectively.

3.1 MOSFET Structure

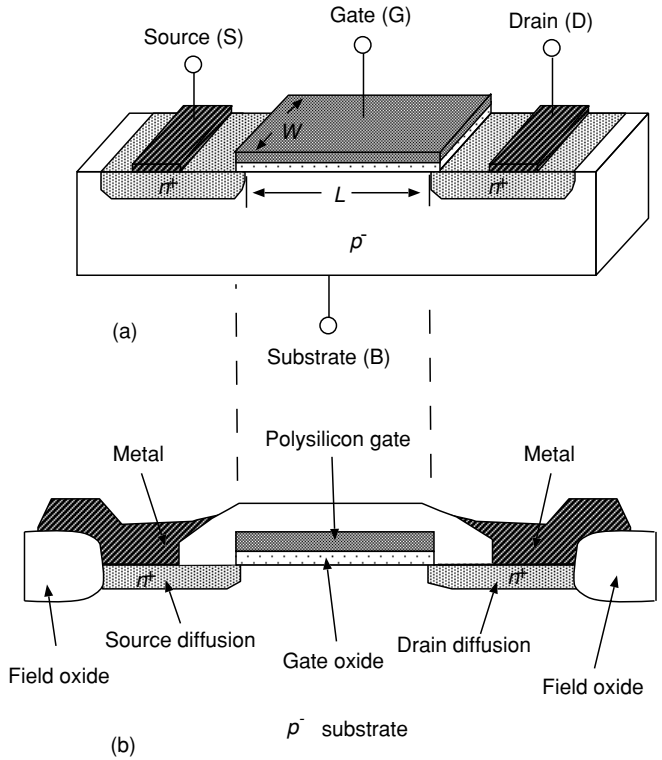


Figure 3.1
Structure of an n-type MOSFET in a p^- body. The MOSFET has four terminals; the drain (D), the source (S), the gate (G), and the bulk (B). (a) Pictorial view of the MOSFET. (b) A more realistic picture of a cross-section of a fabricated MOSFET. Note that the gate oxide is much thinner than the field oxide.

The MOSFET is composed of a MOS structure and two pn junction diodes that were both discussed in Chapter 2. The structure of an n-channel MOSFET in a p^- silicon *body* or *substrate* (lightly doped p-type) is shown in Fig. 3.1. Because the body is p-type (that is, doped with acceptors), the majority carriers in this region are holes. The n^+ regions (heavily doped n-type) are the *source* and *drain* regions of the transistor. These regions are symmetric to each other and as we will see in the next subsection, they are defined only by the voltages applied to these regions. Because the regions are n-type (doped with donors), the majority carriers are electrons. These regions have a low resistivity because of the heavy doping of donors.

The region underneath the gate and between the source and drain regions is called the *channel*. The channel has a width W, and a length L. The minimum transistor length in modern-day MOSFET processes is approaching 0.1 μm. (The issues in process scaling are discussed further in Chapters 13 and 14.) The channel is insulated from the gate above by a layer of silicon dioxide (or *gate oxide*). This oxide is thinner than the other oxide (*field oxide*), which covers the rest of the substrate. The gate is made of heavily doped polycrystalline silicon

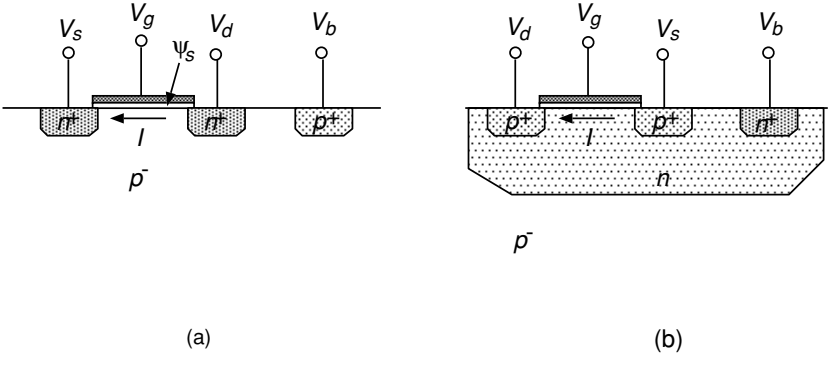


Figure 3.2
Physical structure of (a) an nFET and (b) a pFET in a common p^- substrate. The pFET rests in a n-well within the substrate.

(also called *polysilicon* or poly), which has low resistivity. We can view the transistor as having four terminals; the gate (G), the source (S), the drain (D), and the bulk (B). Because the n^+ source and drain regions can supply a lot of

electrons to the channel, this device is called an n-channel MOSFET (*nFET*)[4]. The complementary type of MOSFET is the p-channel MOSFET (*pFET*)[5], in which the charge in the channel is carried by holes supplied from the source and drain regions.

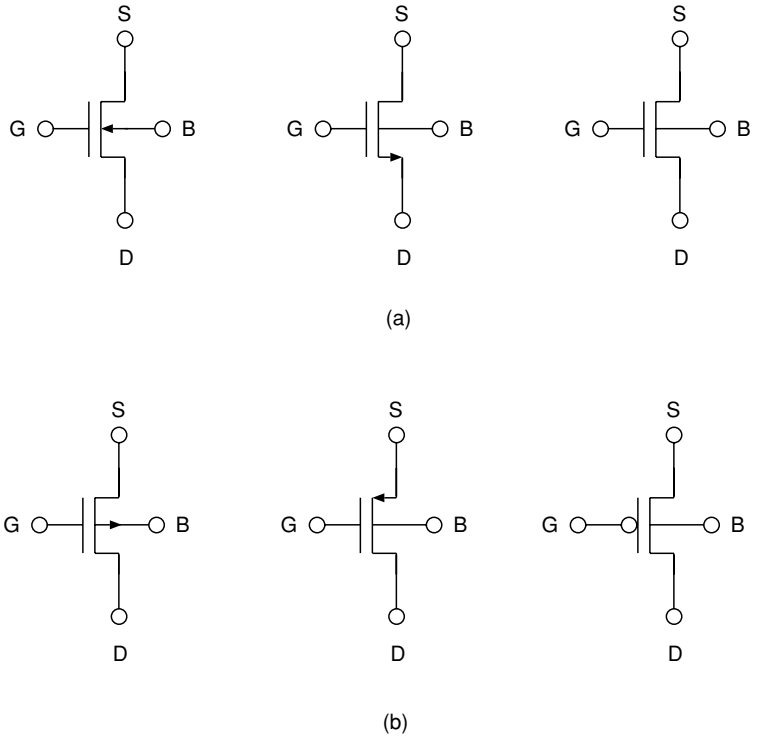


Figure 3.3
MOSFET symbols for (a) an nFET and (b) a pFET.

The fabrication steps needed to fabricate a MOSFET are described in Chapter 13. Almost all modern MOS processes are CMOS (Complementary Metal Oxide Semiconductor) In a CMOS process, both nFETs and pFETs are fabricated on the same substrate. In this chapter, we will assume a single well CMOS process in which the nFET rests in the common p^- substrate, and

4 It is also called an n-type MOSFET or NMOS.
5 It is also called p-type MOSFET or PMOS.

the pFET rests in an n-well within the substrate as shown in Fig. 3.2. In an alternative CMOS process, the common substrate could be n-type, with the nFET resting in a p-well within the substrate; while the pFET rests directly within the substrate. However, most CMOS processes now use a p-type starting substrate. There are also processes in which both types of transistors rest in individual bulks, within a common substrate. This arrangement is called a *twin-tub* CMOS process. In circuit diagrams, the two types of MOSFETs are indicated by the symbols shown in Fig. 3.3. There are three different sets of symbols in common use. The bulk terminal (B) is usually not drawn when the bulks of the transistors are connected to the appropriate power supply. In this book, we use the symbols (without the bulk terminal, unless necessary) shown in the last column.

Biasing the MOSFETs

First, we look at how an nFET should be biased. The drain voltage, V_d, and the source voltage, V_s, of an nFET (see Fig. 3.2(a)) should be greater than or equal to the bulk voltage, V_b, so that the pn junctions between the highly doped n^+ regions and the substrate will be reverse biased. That is, $V_{sb} = V_s - V_b \geq 0$ and $V_{db} = V_d - V_b \geq 0$. These bias conditions guarantee that there will only be a small reverse leakage current at these junctions and that most of the transistor's current will flow in the channel. In an nFET, the n^+ region biased at the higher voltage is called the *drain*, and the other n^+ region is called the source[6]. Because electrons are negatively charged, the direction of positive current flow, I, is from drain to source, as shown in Fig. 3.2(a) and Fig. 3.4(a), even though the carriers flow from source to drain. The currents measured at the source and at the drain are approximately the same, that is, there is very little loss of carriers along the channel.

For a pFET (shown in Fig. 3.2(b)), the p^+ regions should be biased negative relative to its bulk, that is, $V_{sb} \leq 0$ and $V_{db} \leq 0$ so that the pn junctions are again reverse-biased. The n-type bulk (or n-well) of the pFET should be biased higher than the p^- substrate. For a pFET, the p^+ region which is biased at the higher voltage is called the *source* and the other p^+ region is called the drain[7]. Because holes are positively charged, positive channel current, I, flows from the source to the drain as shown in Fig. 3.2(b) and Fig. 3.4(b). The bulk of the pFET is usually connected to the highest voltage (V_{dd}) supplied to the

6 Electrons are supplied to the channel by the source, and removed by the drain.

7 Holes are supplied to the channel by the source, and removed by the drain.

chip while the bulk of the nFET is tied to the lowest voltage (V_{ss}). In a p^- substrate, where the pFET rests in an n-well, the substrate is connected to V_{ss}, and the well to V_{dd}.

3.2 Current–Voltage Characteristics of an nFET

We start by deriving the current-voltage characteristics of the nFET. The characteristics of the MOSFET depend on the relative voltages between the four terminals. The terminal voltages are usually referenced to the source or to the bulk voltage.

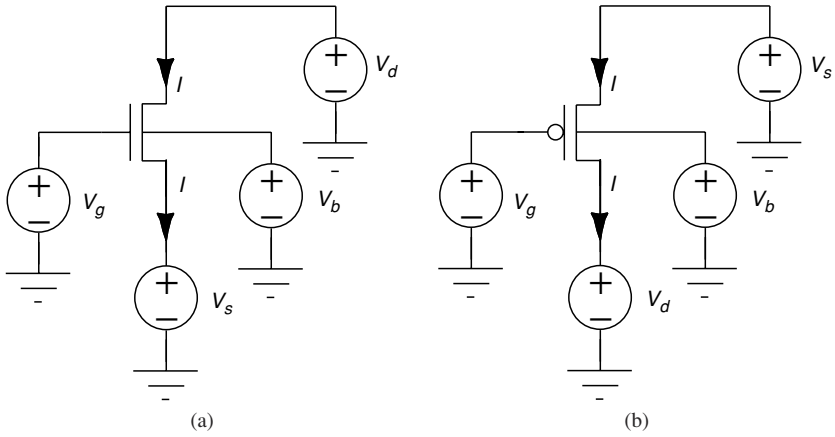


Figure 3.4
Biased MOSFETs showing the direction of conventional current flow. (a) For proper nFET operation, we should ensure that $V_g \geq V_b$, $V_s \geq V_b$, and $V_d \geq V_b$. If $V_d \geq V_s$, the channel current I is positive, flowing from drain to source as shown. If $V_d \leq V_s$, then I is negative and flows in the opposite direction. (b) For proper pFET operation, we should ensure that $V_g \leq V_b$, $V_s \leq V_b$, and $V_d \leq V_b$. If $V_d \leq V_s$, the channel current I is positive, flowing from source to drain as shown. If $V_d \geq V_s$, then I is negative and flows in the opposite direction.

We first consider the current flow in an nFET as a function of the gate-to-source voltage, V_{gs}. Depending on the relative values of the four terminals of the transistor, the nFET operates in one of the following regimes; *accumulation*, *depletion*, *weak inversion*, *moderate inversion*, or *strong inversion*. These regimes are equivalent to those of the MIS structure described in Chapter 2. We say the MOSFET is in *cutoff*, when it is in either the accumulation or de-

pletion regime; in *subthreshold*, when it is in weak inversion ; and to be *above threshold*, when the MOSFET operates in the strong inversion regime[8]. In the cutoff region, the current in the MOSFET is close to zero. For both the *subthreshold* and *above-threshold* regimes we can further divide the operation of the MOSFET into the *triode* and *saturation* modes depending on the value of V_{ds}.

Subthreshold Region

Increasing the gate voltage increases the positive charge on the gate. This charge repels the holes in the substrate and leaves behind negatively-charged ions, that balance out the gate charge. The MOSFET operates in the subthreshold regime when the positive charge on the gate is almost balanced by the negatively-charged depletion region underneath the gate (see Fig. 3.5(a)). There is also a very thin layer of electrons beneath the gate (the *inversion layer*). In subthreshold, we ignore the charge from the inversion layer because it is almost negligible compared with the depletion charge. The energy band diagrams in Figs. 3.5(b) and (c) are similar to the band diagrams for the pn junction in Chapter 2[9]. In these diagrams, the axis for the electron energy is directed upwards, while the positive voltage axis is downwards. That is, the quasi-Fermi level of the region that has a higher applied voltage (like the drain of the transistor) is lower than the quasi-Fermi level of the source region.

We now compute the electron concentrations at the source and drain ends of the channel. Assume for the moment that the potential in the channel (or the surface potential), ψ_s, is constant. This constant potential corresponds to zero drift current. The electron concentrations at the two ends of the channel depend on the energy barrier that the electrons encounter. This barrier is determined by the voltage difference between the surface potential ψ_s and the applied voltages V_s and V_d at the source and drain respectively. The barrier heights at the source end of the channel, θ_s, and at the drain end of the channel, θ_d, can be expressed as

$$\theta_s = \theta_0 - q(\psi_s - V_s) \tag{3.2.1}$$

$$\theta_d = \theta_0 - q(\psi_s - V_d) \tag{3.2.2}$$

8 The below-threshold regime includes the subthreshold and cutoff regimes.

9 To assist non-physicists in making qualitative band diagrams, we assume that the quasi-Fermi level, E_F, of each region is connected to the external bias voltage that is applied to the region (except for the ψ_s region). This assumption is not true in reality. Remember that the positive voltage axis points downwards.

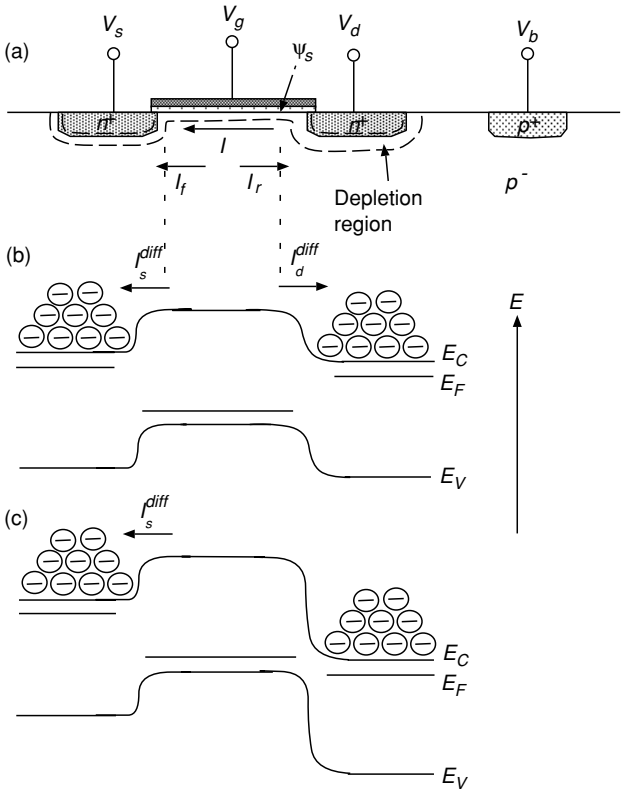


Figure 3.5
An nFET in subthreshold. (a) Cross-section. (b) Energy band diagram in the linear regime. (c) Energy band diagram in the saturation regime.

where $\theta_0 = qV_{bi}$ is the built-in energy barrier between the n^+ drain/source and the p^- substrate; V_{bi} is the built-in potential at this pn junction; and q is the positive elementary charge.

The electron density (the number of electrons per unit volume) at the source end of the channel, N_s, is given by

$$N_s = N_o e^{-\theta_s/kT} = N_o e^{-(\theta_0 - q(\psi_s - V_s))/kT} \tag{3.2.3}$$

where N_o is the effective number of states per unit area in the channel. A corresponding relationship holds for the electron density at the drain end of

the channel:

$$N_d = N_o e^{-\theta_d/kT} = N_o e^{-(\theta_0 - q(\psi_s - V_d))/kT}. \tag{3.2.4}$$

These equations are obtained by solving the boundary conditions of the energy states at the boundaries where the channel meets the source and drain depletion regions[10] [11]. Because the barrier at the drain is higher than that at the source, the electron concentration at the drain end of the channel is lower than at the source end. The concentration gradient leads to the diffusion of electrons from the source to the drain[12]. The electrons that diffuse to the drain end of the channel are swept into the drain by the electric field in the depletion region around the drain. The electrons in the channel form an inversion layer.

We can compute the current in the transistor using the electron diffusion current density equation, Eq. 2.5.7 in Chapter 2:

$$\begin{aligned} I &= J_{n,diff}\, W\, t \\ &= -qWtD_n \frac{dN}{dz} \end{aligned} \tag{3.2.5}$$

where $J_{n,diff}$ is the electron diffusion current density, W is the width of the channel, t is the depth of the channel, D_n is the diffusion coefficient of the electrons, and $\frac{dN}{dz}$ is the concentration gradient across the channel. In computing $\frac{dN}{dz}$, we assume that the conduction is lossless. Therefore, the current is constant as a function of position z in the channel; where z=0 at the source end of the channel. The concentration gradient can be expressed as

$$\frac{dN}{dz} = \frac{N_d - N_s}{L} = \frac{N_1}{L} e^{\psi_s/U_T} (e^{-V_d/U_T} - e^{-V_s/U_T}) \tag{3.2.6}$$

where L is the channel length, $U_T = kT/q$ is the thermal voltage[13], and

10 The carrier density at the boundary is given by the integral with respect to the energy E, of the product of the two-dimensional density of states, $N(E)$, in the channel and the Fermi-Dirac distribution of energy states in the source or drain region (see Eq. 2.3.3). This integral is as follows:

$$\int_{E_c}^{\infty} \left(N(E) \frac{1}{e^{(E-E_f)/kT} + 1} \right) dE$$

where E_c is the energy at the conduction band and E_f is the quasi-Fermi level.

11 This derivation, or a similar derivation can be found in Maher (1989); Tsividis (1998).

12 Similarly, the current flow in a neuron is due to the diffusion of ions between the intracellular space and extracellular space of the neuron.

13 $U_T \approx 25.8$ mV at room temperature.

$N_1 = N_o\, e^{-\theta_0/kT}$. Substituting Eq. 3.2.6 in Eq. 3.2.5, we get

$$I = -q\frac{W}{L}tD_nN_1e^{\psi_s/U_T}(e^{-V_d/U_T} - e^{-V_s/U_T}) \tag{3.2.7}$$

$$= I_{t0}\, e^{\psi_s/U_T}(e^{-V_s/U_T} - e^{-V_d/U_T}) \tag{3.2.8}$$

where $I_{t0} = q\frac{W}{L}tD_nN_1$ is the accumulated pre-exponential.

The surface potential ψ_s is modulated by the gate voltage V_g. In subthreshold, the gate charge per unit area Q_g is balanced primarily by the depletion charge per unit area Q_d underneath the gate as described in Section 2.7. As the gate voltage increases, the depletion region also increases so that the negative charge in this region balances the increased positive charge on the gate. The surface potential relative to the bulk also increases because of the wider depletion region (see Eq. 2.7.19). By assuming a small change around an operating point, ψ_0, we can express ψ_s as

$$\psi_s = \psi_0 + \kappa V_g \tag{3.2.9}$$

where κ is the *capacitive coupling ratio from gate to channel* and is defined by Eq. 2.7.29 in Chapter 2. Substituting Eq. 3.2.9 into Eq. 3.2.8, we can write the current–voltage (I–V) characteristics for an nFET in terms of the gate, source, and drain voltages which are referenced to the bulk voltage as

$$\boxed{I = I_0\, e^{\kappa V_g/U_T}(e^{-V_s/U_T} - e^{-V_d/U_T}).} \tag{3.2.10}$$

Equation 3.2.10 can be separated into the forward current, I_f; and the reverse current, I_r; so that it can rewritten as

$$\boxed{I = I_f - I_r.} \tag{3.2.11}$$

The directions of the currents are shown in Fig. 3.5. These component currents are

$$I_f = I_0\, e^{(\kappa V_g - V_s)/U_T} \tag{3.2.12}$$

$$I_r = I_0\, e^{(\kappa V_g - V_d)/U_T}. \tag{3.2.13}$$

The forward component I_f depends only on V_g and V_s, while the reverse component I_r depends only on V_g and V_d. This form of the subthreshold I–V equation will be useful when we analyze circuits in later chapters.

Measuring κ How can we determine κ? The derivation for κ was given by Eq. 2.7.30 in Chapter 2:

$$\kappa = \frac{C_{ox}}{C_{ox} + C_d} \tag{3.2.14}$$

where C_{ox} is the capacitance of the gate oxide per unit area, and C_d is the incremental capacitance of the depletion layer per unit area. The depletion capacitance is a non-linear function of the gate-to-bulk voltage, V_{gb} (see Eq. 2.7.24). As V_{gb} increases, the depletion width increases slowly; therefore C_d decreases, which in turn leads to an increase in κ.

One way of computing κ is to measure the slope of the log-linear plot of I versus V_g when the nFET operates in the subthreshold regime. An example of this measurement is shown in Fig. 3.6. The data was measured from an nFET fabricated in a 0.8 μm CMOS process. The dependence of κ on V_g can be measured from the local slopes of this curve. Because κ varies slowly with V_g, we can usually assume that κ is constant in an approximate analysis of a subthreshold circuit. Since κ determines the slope of the MOSFET in subthreshold, we also refer to κ as the *subthreshold slope factor*. The value of κ ranges from 0.5 to 0.9 depending on the process in which the transistors are fabricated.

Another way of computing κ is to source a constant current through the MOSFET, and then to measure the source voltage of the transistor as a function of the gate voltage. The slope of this curve also gives an approximate value for κ. A curve of κ plotted against V_g computed in this way is shown in Fig. 5.5 in Chapter 5.

Drain-to-Source Voltage Dependence in the Subthreshold Region

The subthreshold equation (Eq. 3.2.10) of the nFET encompasses two regions of operation; the *triode* region and the *saturation* region. Which of these regions the transistor operates in depends on the drain-to-source voltage (see Fig. 3.7). In the triode region, the current depends on V_{ds} while in the saturation region, the current is almost independent of V_{ds}.

Triode Region

The triode region describes the operation of the transistor for small V_{ds}. It is also called the *linear* region[14]. Eq. 3.2.10 describes the current-voltage

14 It is also called the *non-saturation* region, the *conduction* region, or the *ohmic* region. The term *linear* region comes from the above-threshold characteristics of the MOSFET where the current is

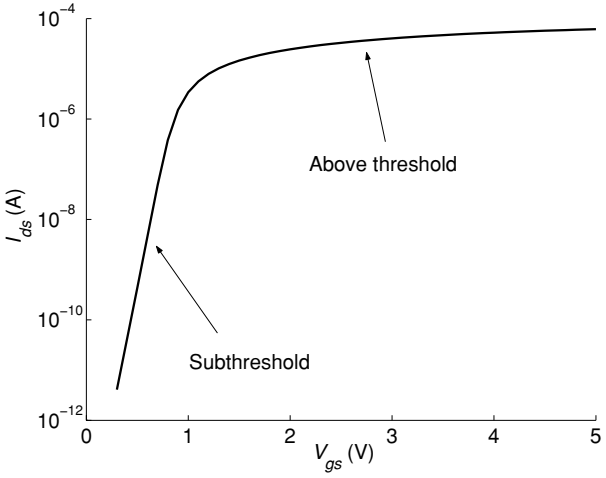


Figure 3.6
Current as a function of the gate-to-source voltage V_{gs}, as measured from an nFET (W/L=12.8/1.6) in a 0.8 μm CMOS process for a fixed V_{ds} and V_{sb}. The log-linear plot shows the different regimes of operation. In the subthreshold regime, the current I_{ds} is exponentially dependent on V_{gs}. For this device, the current changes by e-fold for every 37 mV with a measured κ of 0.576. In the above-threshold regime, the current has a quadratic dependence on V_{gs}. The moderate inversion regime lies in between the subthreshold and above-threshold regimes.

characteristic of the nFET operating in the linear region. This equation can be rewritten as

$$\boxed{I = I_0\, e^{(\kappa V_g - V_s)/U_T}\,(1 - e^{-V_{ds}/U_T}).} \tag{3.2.15}$$

By taking a Taylor's series expansion of Eq. 3.2.15, we can show that I is approximately linear with V_{ds} for $V_{ds} \leq U_T$ (see Fig. 3.7).

Saturation Region

As V_{ds} increases beyond $4\,U_T$, the concentration of electrons at the drain end of the channel can be neglected with respect to the concentration at the source end because of the larger barrier height as shown in Fig. 3.5(c). Any electrons in the channel that diffuse close to the drain are immediately swept into the drain by the electric field in this region. Because the diffusion current is no longer dependent on the electron concentration at the drain, the current in the transistor depends only on V_s and is approximately equal to I_f. The I–

linear with V_{ds}.

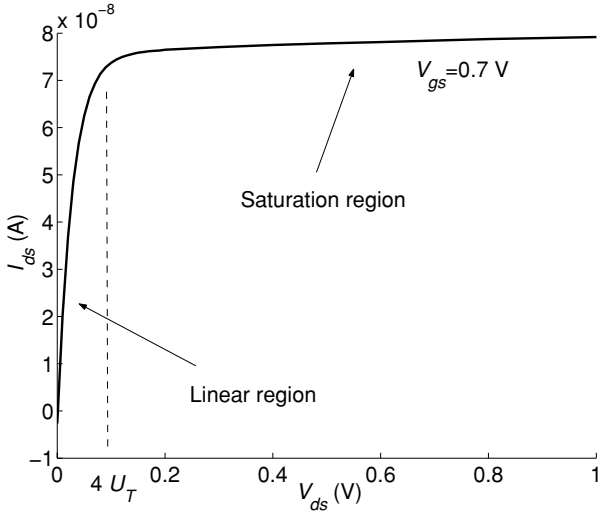


Figure 3.7
The current, I_{ds} as a function of V_{ds} for an nFET in the subthreshold region. The current was measured from the same device in Fig. 3.6. The current is approximately linear in V_{ds} for very small values of V_{ds} and is approximately constant for $V_{ds} > 4U_T$.

V relationship in this region is described by

$$\boxed{I = I_f = I_0\, e^{(\kappa V_g - V_s)/U_T}.} \tag{3.2.16}$$

This region of operation is called the *saturation* region. As seen in Fig. 3.7, the current is approximately constant in this region[15]. Figure 3.8 shows a family of curves measured from the same nFET in the subthreshold region with gate voltages ranging from 0.3 V to 0.7 V. In these curves, the transition point from the linear region to the saturation region occurs around V_{ds} = 100 mV and is independent of the gate voltage.

Above Threshold Region

In the subthreshold region, the transistor current flows primarily by diffusion. As the gate voltage increases, the surface potential ψ_s also increases, causing more electrons to enter the channel from the source and drain regions. An increasingly larger portion of the positive gate charge is balanced by the inver-

15 We will see in Section 3.5 that the current in this region is not actually constant because the drain voltage modulates the effective channel length of the transistor, causing I to depend on V_d.

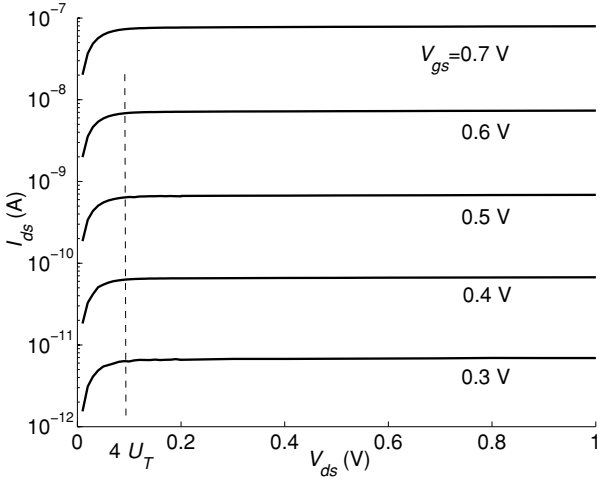


Figure 3.8
A family of curves showing I versus V_{ds}, as measured from a subthreshold nFET for V_{gs} between 0.3 V and 0.7 V in increments of 0.1 V. All curves start saturating around $V_{ds} \approx 4U_T$.

sion charge formed by the electrons in the channel. In this intermediate region (or *moderate inversion* region), the current consists of both drift and diffusion currents. As the gate voltage increases further, the transistor begins to operate in the *strong inversion* region (or above threshold region). Here, the current consists predominantly of drift current and the charge on the gate is balanced primarily by the inversion charge. Further increase in the gate charge is balanced by an increase in the inversion charge. Because of the larger electron concentrations at the source and drain ends of the channel (as compared to the concentrations in the subthreshold region), the now finite horizontal electric field, pointing from the drain to the source, creates a potential drop across the channel. In contrast to the subthreshold region, the surface potential along the channel is no longer constant but varies from $\psi_s(0)$ at the source end to $\psi_s(l)$ at the drain end, where $\psi_s(l) > \psi_s(0)$.

In above-threshold, the surface potential of the transistor no longer depends on V_g through κ. The electron density at the source end of the channel is exponentially dependent on ψ_s. Therefore, only a small change in ψ_s will supply sufficient electrons to offset the additional positive charge on the gate caused by the increase in V_g. The surface potential can be treated as if it is clamped in the above threshold regime. Figure 3.9 shows the dependence of

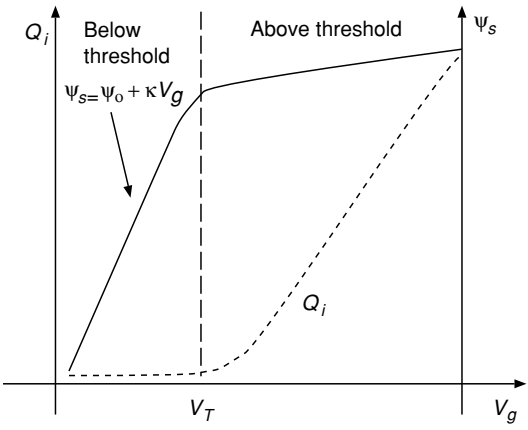


Figure 3.9
Qualitative plots showing the dependence of the inversion charge Q_i (dashed line) and the surface potential ψ_s (solid line) on the gate voltage in subthreshold, and above-threshold regimes of an nFET. The threshold voltage V_T delineates the two operating regions.

the inversion charge per unit area, Q_i, and the surface potential ψ_s, on the gate voltage in both the subthreshold and above threshold regions.

Threshold Voltage As shown in Fig. 3.9, the *threshold voltage*, V_T, delineates the two operating regimes of the transistor. This voltage is the minimum gate voltage required to obtain strong inversion in the MOS structure. The threshold voltage of the nFET at $V_s = 0$ (from Eq. 2.7.32 in Chapter 2) is

$$V_{T0} = V_{fb} + 2\psi_B - \frac{Q_d}{C_{ox}} \tag{3.2.17}$$

$$= V_{fb} + 2\psi_B + \gamma\sqrt{2\psi_B} \tag{3.2.18}$$

where V_{fb} is the flat-band voltage, Q_d is the incremental capacitance of the depletion region per unit area, $2\psi_B$ is the surface potential of the channel when the concentration of electrons in this region is equal to the concentration of acceptor ions, and $\gamma = \frac{\sqrt{2\epsilon_s q N_A}}{C_{ox}}$ is the *body factor coefficient*. The body factor coefficient γ is described in Section 3.2.

What happens when the source voltage increases? An increase in V_s leads to a decrease in the current of the transistor. To return to the original current, ψ_s has to increase by the same amount as the increase in V_s. Since $\delta\psi_s = \kappa\,\delta V_g$, the gate voltage has to increase by a factor of ($\frac{1}{\kappa} > 1$) more than V_s. The larger

increase required for V_g can be treated as an increase in the effective threshold voltage V_T of the transistor:

$$V_T = V_{T0} + \frac{V_s}{\kappa}. \tag{3.2.19}$$

As mentioned previously, when the transistor is abvoe threshold, increases in the gate charge are balanced primarily by increases in the inversion charge so we can write the relationship between the inversion charge and the threshold voltage as

$$Q_i = -C_{ox}(V_g - V_T) \tag{3.2.20}$$

$$= -C_{ox}(V_g - V_{T0} - \frac{V_s}{\kappa}). \tag{3.2.21}$$

This equation will be used in the derivation of the I–V characteristics of the nFET in above-threshold.

Drain-to-Source Voltage Dependence in the Above Threshold Region

As in subthreshold, above-threshold also encompasses two regions of operation: the triode regime, and the saturation regime. These regimes can be seen in Fig. 3.10, which shows a family of I versus V_{ds} curves taken from an nFET in the above-threshold region. Notice that the current do not longer saturates around $V_{ds} \approx 100\,mV$ (compare with Fig. 3.8 for the subthreshold curve). The drain-to-source voltage at which the current saturates depends on the gate voltage.

Triode Region

We first compute the I–V characteristic in in the deep *triode* or *ohmic* regime, where V_{ds} is very small. Here, the drift current depends linearly on V_{ds}. The energy band diagram of the nFET is shown in Fig. 3.11(b). To compute the drift current, we begin from the electron drift current density equation (Eq. 2.5.11). The drift current density, $J_{n,drift}$, (the drift charge flowing through a given cross-section during a given time interval) is given by

$$J_{n,drift} = q\mu_n n\mathcal{E} \tag{3.2.22}$$

where n is the carrier concentration, μ_n is the mobility of the electrons, and $\mathcal{E}$ is the horizontal field. The charge concentration can be re-expressed in terms

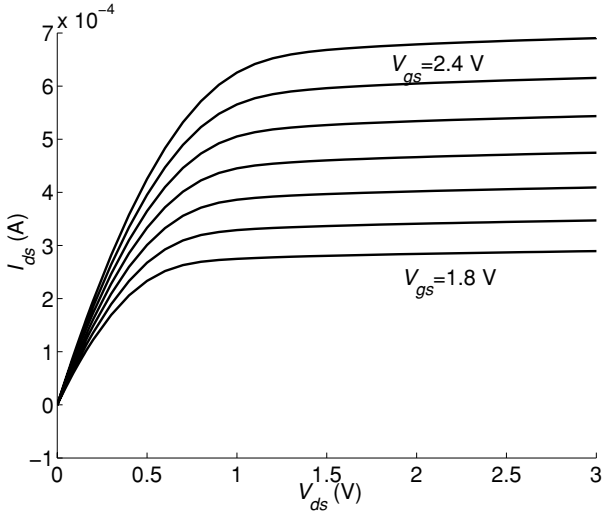


Figure 3.10
A family of curves showing the dependence of I on V_{ds} as measured from an nFET in above-threshold. The curves were taken for V_{gs} between 1.8 V and 2.4 V in increments of 0.1 V.

of the inversion charge per unit area, Q_i:

$$qn = -\frac{Q_i W L}{W L t} = -\frac{Q_i}{t} \tag{3.2.23}$$

where W, L, and t are the width, length, and depth of the channel respectively. We can then compute the drift current as

$$I = J_{n,drift}\, W t = \mu_n\, Q_i W \mathcal{E} \tag{3.2.24}$$

We now substitute Eq. 3.2.20 into Eq. 3.2.24 and, assuming a constant field across the channel for small V_{ds}, we get[16]

$$\begin{aligned} I &= \mu_n\, C_{ox} (V_g - V_T) W \frac{V_d - V_s}{L} \\ &= \mu_n\, C_{ox} \frac{W}{L} (V_g - V_T)\, (V_d - V_s). \end{aligned} \tag{3.2.25}$$

We can rewrite Eq. 3.2.25 as

$$\boxed{I = \beta (V_g - V_T)\, (V_d - V_s)} \tag{3.2.26}$$

16 We also assume that the electron mobility is constant because of the small lateral field.

where $\beta = \mu_n C_{ox} \frac{W}{L}$. From Eq. 3.2.26, we see immediately that the current through the MOSFET is linearly proportional to V_{ds}, and hence the transistor acts like a linear resistor in this region. The conductance of the resistor can be set by the gate voltage.

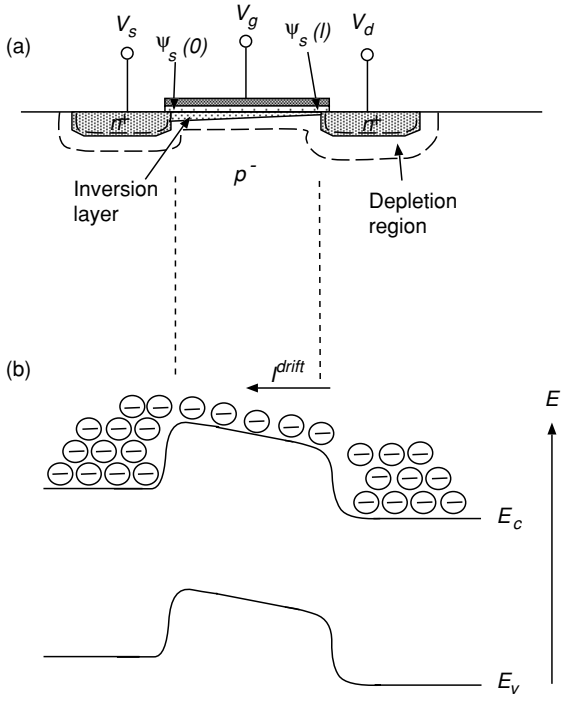


Figure 3.11
An nFET in the linear region in the above threshold domain. (a) Cross-section. (b) Energy band diagram.

In the derivation of Eq. 3.2.26, we assumed that Q_i is uniformly distributed through the channel. In the triode and saturation regimes, (excluding the deep triode regime), the distribution of Q_i is non-uniform because the voltage drop across the gate oxide, $V_i(z)$=$V_g - \psi_s(z)$, varies with position along the channel length z[17]. Because the voltage drop $V_i(z)$ decreases towards the drain end of the channel, Q_i also decreases towards the drain end as shown in Fig. 3.11.

17 Remember that the gate charge per unit length is determined by $C_{ox}\ V_i(z)$.

However, the current is constant along the channel if we assume no loss of carriers along the channel.

Again, we use Eq. 3.2.24 to compute the drift current in the triode regime. The inversion charge Q_i can be computed indirectly from the incremental channel capacitance per unit area, C, which consists of the oxide capacitance, C_{ox}, and the depletion capacitance, C_d:

$$C = \frac{\partial Q_i}{\partial \psi_s} = C_{ox} + C_d. \tag{3.2.27}$$

We no longer assume that the field along the channel is constant. Instead, it depends on the gradient of the surface potential ψ_s along the length of the channel:

$$\mathcal{E} = -\frac{d\psi_s}{dz}. \tag{3.2.28}$$

If we substitute Eq. 3.2.27 into Eq. 3.2.28, we can rewrite $\mathcal{E}$ as

$$\mathcal{E} = -\frac{1}{C}\frac{dQ_i}{dz}. \tag{3.2.29}$$

Substituting Eq. 3.2.29 back into the Eq. 3.2.24, we solve

$$I = -\mu_n Q_i W \mathcal{E} = \frac{\mu_n}{C} W Q_i \frac{dQ_i}{dz} = \frac{1}{2}\frac{\mu_n}{C} W \frac{d}{dz} Q_i^2. \tag{3.2.30}$$

Integrating both sides of Eq. 3.2.30, we rewrite I as

$$I\,L = W\frac{\mu}{2C}(Q_d^2 - Q_s^2). \tag{3.2.31}$$

where Q_s and Q_d are the inversion charges at the source end and the drain end of the channel respectively. In this equation, I is negative because the z-axis is directed from the source towards the drain. Our definition of positive current flow is from the drain towards the source. So we rewrite I to accommodate this convention:

$$I = \frac{W}{L}\frac{\mu}{2C}(Q_s^2 - Q_d^2). \tag{3.2.32}$$

We use Eq. 3.2.21 to compute Q_s and Q_d. The threshold at the source end of the channel is given by $V_T(s) = V_{T0} + \frac{V_s}{\kappa}$. Therefore, the inversion charge at the source end of the channel is

$$Q_s = C_{ox}(V_g - V_{T0} - \frac{V_s}{\kappa}). \tag{3.2.33}$$

The threshold at the drain end of the channel is given by $V_T(d) = V_{T0} + \frac{V_d}{\kappa}$.

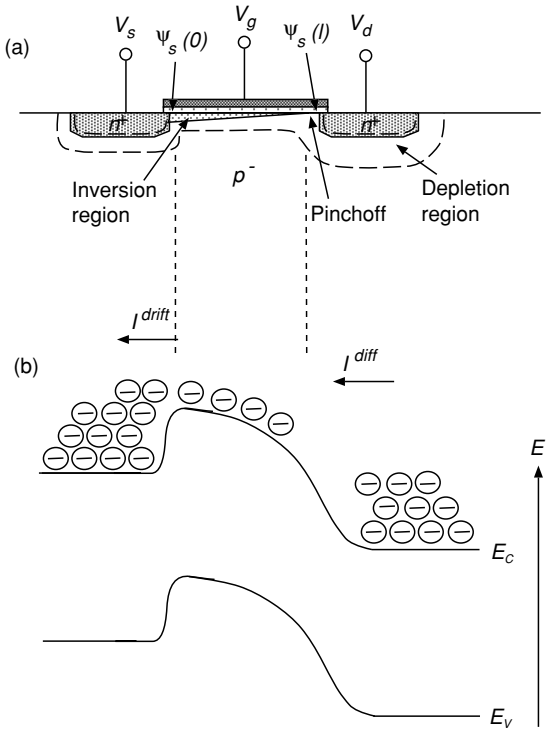


Figure 3.12
An nFET in the above-threshold saturation region. (a) Cross-section. The pinchoff region lies between the pinchoff point and the drain, where the inversion charge is negligible. (b) Energy band diagram.

Therefore, the inversion charge at the drain end is

$$Q_d = C_{ox}(V_g - V_{T0} - \frac{V_d}{\kappa}). \tag{3.2.34}$$

Replacing the definitions for Q_s and Q_d into Eq. 3.2.32, and using the definition $\kappa = C_{ox}/C$, we get

$$\boxed{I = \frac{\beta}{2\kappa}\left[(\kappa(V_g - V_{T0}) - V_s)^2 - (\kappa(V_g - V_{T0}) - V_d)^2\right]} \tag{3.2.35}$$

where $\beta = \mu C_{ox}\frac{W}{L}$. As in the subthreshold case, this current is the sum of a

Table 3.1
Current-voltage relationship of nFET in subthreshold and above-threshold.

Mode	Subthreshold	Above threshold
Triode	$I_0 e^{\kappa V_g/U_T}(e^{-V_s/U_T} - e^{-V_d/U_T})$	$\frac{\beta}{2\kappa}[{Q_s}^2 - {Q_d}^2]$
Saturation	$I_0 e^{(\kappa V_g - V_s)/U_T}$	$\frac{\beta}{2\kappa}[(\kappa(V_g - V_{T0}) - V_s)^2]$
Cutoff	0	0

forward component I_f and a reverse component I_r:

$$I_f = \frac{\beta}{2\kappa}\big[(\kappa(V_g - V_{T0}) - V_s)^2\big] \tag{3.2.36}$$

$$I_r = \frac{\beta}{2\kappa}\big[(\kappa(V_g - V_{T0}) - V_d)^2\big]. \tag{3.2.37}$$

We can further reduce Eq. 3.2.35 to

$$I = \beta\big[(V_g - V_T)(V_d - V_s) - \frac{1}{2\kappa}((V_d - V_s)^2)\big]. \tag{3.2.38}$$

By assuming that V_{ds} is very small, we can neglect the second term in Eq. 3.2.38 and obtain the same equation as Eq. 3.2.26 for the deep triode region.

Saturation Region

As we increase V_d further, the threshold towards the drain end of the channel increases. Consequently, the inversion charge disappears at some point along the channel because the gate charge needs only to be balanced by the depletion charge. The point in the channel where the inversion charge first disappears is called the *pinchoff point*. The region between the pinchoff point and the drain is called the *pinchoff region* and is almost depleted of electrons (see Fig. 3.12). In fact, this pinchoff region is in subthreshold. The current in the pinchoff still flows by drift and the electrons are swept into the drain region by the electric field resulting from the potential difference between the channel and the drain. The current in the transistor only depends on the source voltage, and is independent of the drain voltage. Accordingly, this region of operation is called the *saturation region*.

We derive the I–V relationship of the transistor in the above-threshold saturation regime by setting $Q_d = 0$ in Eq. 3.2.35:

$$I = I_f = \frac{\beta}{2\kappa}\big[(\kappa(V_g - V_{T0}) - V_s)^2\big]$$

which reduces to

$$\boxed{I = \frac{\beta\kappa}{2}\left[(V_g - V_T)^2\right].} \tag{3.2.39}$$

The triode and saturation regimes of an nFET operating above threshold are shown in Fig. 3.10 for different values of V_{gs}. The value of V_d where the transition between the triode and saturation regimes occurs, depends on V_g. This dependence on V_{gs} is unlike the subthreshold case where the transition is independent of the gate voltage. The transition value is computed by setting $Q_d = 0$ in Eq. 3.2.35:

$$\boxed{V_d = \kappa(V_g - V_{T0}).} \tag{3.2.40}$$

The I–V characteristics of the nFET in the subthreshold and above-threshold regions are summarised in Table 3.1. *In above-threshold, the definition of κ is invalid because the incremental increase in charge underneath the gate due to an increase in the gate charge arises from the inversion charge not the depletion charge. We set $\kappa = 1$ in the above-threshold equations.*

Body Effect

In the I–V equations that we have derived so far, the terminal voltages of the transistor are referenced to the bulk. However, the bulk is also an input to the transistor. In the subthreshold region, we can describe the influence of V_b through the series of capacitors C_{ox} and C_d (as we saw also for the gate input). The effect on the surface potential can be written as

$$\partial\psi_s = (1 - \kappa)\,\partial V_b. \tag{3.2.41}$$

The influence of the bulk on the transistor[18] is called the *body effect* or *substrate effect*. The bulk is sometimes called the *back gate*; hence the body effect is also called the *backgate effect*.

In a later chapter, we show a circuit in which the bulk is used as the input to a transistor instead of the gate.

In the strong inversion, or above-threshold, regime the influence of the bulk voltage is usually treated as an increase in the threshold voltage of the transistor. If V_b decreases, then there is practically no change in the gate charge because the voltage across the gate oxide is essentially unchanged (the surface

18 Remember that since the surface potential is referenced to the bulk, the actual surface potential change will be $-\kappa\partial V_b$.

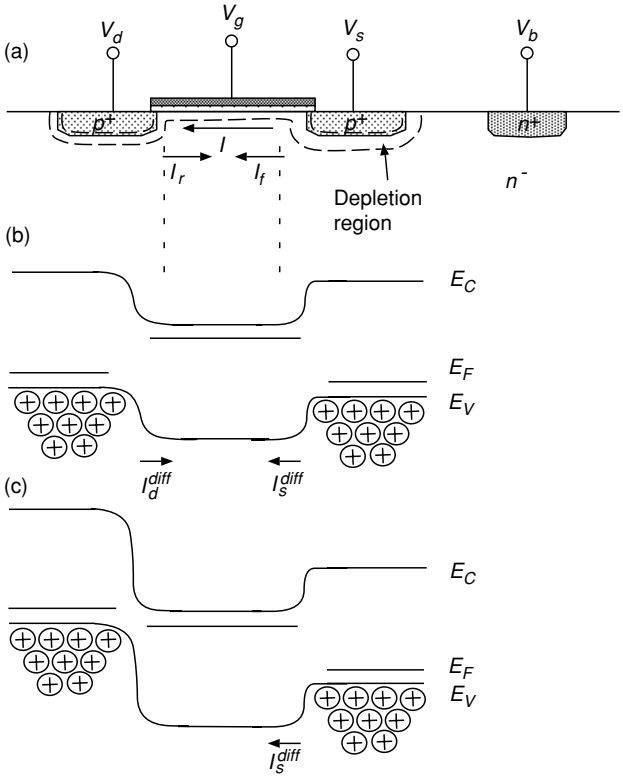


Figure 3.13
A pFET in subthreshold. (a) Cross-section. (b) Energy band diagram in the linear region. (c) Energy band diagram in the saturation region.

potential remains approximately the same). However, the depletion region underneath the gate increases[19]. Since the negative charge from the depletion region is now larger, less charge is required in the inversion region to balance the gate charge. The inversion region becomes smaller, so leading to a smaller I. To restore I to its original value, we increase the gate voltage V_{gs}. The effective threshold is now

$$\boxed{V_T = V_{fb} + 2\psi_B + \gamma\sqrt{V_{bs} + 2\psi_B}} \tag{3.2.42}$$

19 Remember from Chapter 2 that increasing the reverse bias across a pn junction causes the depletion width to increase.

where γ is the *body effect coefficient* described in Eq. 3.2.18, and V_{bs} is the bulk-to-source voltage[20].

Assuming that we do not forward the pn junctions between the drain/source regions and the bulk: What happens if the bulk voltage V_b is increased by ΔV? This scenario is the same as decreasing V_g, V_s, and V_d by the same ΔV. In the subthreshold region, the change in ψ_s will now be $-\kappa\,\Delta V$. The barrier height is decreased at both ends of the channel, and the current increases. Hence, the bulk acts like the gate, but it has a weaker influence on the transistor current.

3.3 Current–Voltage Characteristics of a pFET

The current in a pFET arises from the transport of holes across the channel from the source to the drain[21]. In subthreshold, the current is primarily due to diffusion. The structure of the pFET and its energy band diagrams in the linear and saturation regions of the subthreshold domain are shown in Fig. 3.13(a), (b), and (c) respectively. Since the diffusion process in a pFET obeys the same laws as in the nFET, we can derive the current–voltage characteristic in the same way as described in Section 3.2. The corresponding I–V equation for the pFET is

$$\boxed{I = I_0\, e^{\kappa(V_w - V_g)/U_T}\,(e^{-(V_w - V_s)/U_T} - e^{-(V_w - V_d)/U_T})} \tag{3.3.1}$$

where V_w is the bulk voltage of the MOSFET. The pre-exponentials I_0 and κs in the pFET and nFET equations are not equivalent. The κ value for the pFET is different than that for the nFET because of the different doping concentrations underneath the gate in the two types of transistors.

If the pFET rests in a n-well, then the well voltage, V_w, is usually connected to the highest potential, V_{dd}. But, as we have seen earlier, the bulk of the pFET can also be used as an input as long as the pn junctions at the source/drain regions and at the well are not forward-biased. If V_g, V_s, and V_d are referenced to V_w, Eq. 3.3.1 can be written as

$$I = I_0\, e^{-\kappa V_g/U_T}\,(e^{V_s/U_T} - e^{V_d/U_T}). \tag{3.3.2}$$

20 It can be shown Enz et al. (1995) that κ is given by

$$\kappa = 1 - \frac{\gamma/2}{\sqrt{V_{gb} - V_{fb} - \gamma^2/4}}.$$

21 Recall that in the pFET, the source is biased at a higher voltage than the drain.

The equation for the pFET in saturation is then

$$I = I_0\, e^{(-\kappa V_g + V_s)/U_T}. \tag{3.3.3}$$

The above-threshold equation for the pFET is

$$\boxed{I = \frac{\beta}{2\kappa}\left[(-\kappa(V_g - V_{T0}) + V_s)^2 - (-\kappa(V_g - V_{T0}) + V_d)^2\right].} \tag{3.3.4}$$

3.4 Small-Signal Model at Low Frequencies

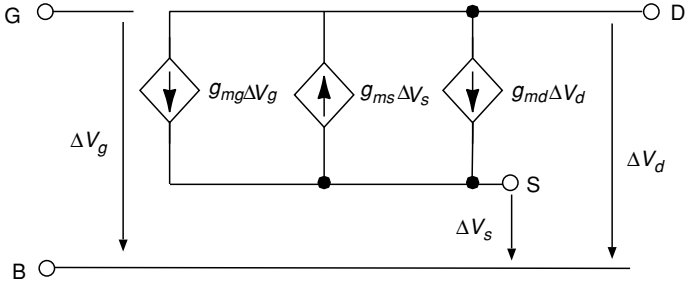


Figure 3.14
Small-signal model of an nFET at low frequencies. The current due to g_{ms} flows in the opposite direction to the currents that are due to the gate transconductance and the drain conductance. This convention takes into account the fact that the current decreases when V_s increases.

We now look at the case of small variations in the terminal voltages so that we can express the resulting currents using linear equations. The aim is to replace the nFET by a linear circuit that is the *small-signal equivalent model* of the transistor in a quasi-static condition (that is, at low frequencies)[22]. This model is shown in Fig. 3.14. Since the signal variations are slow, we can ignore the capacitive effects at the nodes.

Conductance Definitions

In the small-signal model, the change in the transistor current due to very small changes in each of the terminal voltages, can be expressed as conductances. Because three terminal voltages are are normally referenced to the voltage

22 We discuss the small-signal model of an nFET at moderate frequencies in Appendix 3.7.

Table 3.2
Conductances of Subthreshold nFET at Low Frequencies.

Conductance	Subthreshold
g_{mg}	$\frac{\kappa I}{U_T}$
g_{ms}	$\frac{I_0 e^{(\kappa V_g - V_s)/U_T}}{U_T}$
g_{md}	$\frac{I_0 e^{(\kappa V_g - V_d)/U_T}}{U_T} + \frac{I}{V_e}$

at the fourth terminal (either the source or the bulk), only three conductance parameters are required. In most textbooks, the analysis assumes that the three terminals are referenced to the source voltage. However, we first look at the case in which the voltages are referenced to the *bulk voltage*. The small-signal gate transconductance, g_{mg}, is defined as

$$g_{mg} = \frac{\partial I}{\partial V_g}. \tag{3.4.1}$$

This parameter describes how the current changes with a small change in the gate voltage. It is a "transconductance" because the the gate terminal only indirectly determines the transistor current. The source conductance, g_{ms}, is the parameter that describes how the current changes with a small change in the source voltage. The source and drain conductances are real "conductances" because the current actually flows between the drain and source terminals. The source conductance is given by

$$g_{ms} = -\frac{\partial I}{\partial V_s}, \tag{3.4.2}$$

and the small-signal drain conductance, g_{md}, is given by

$$g_{md} = \frac{\partial I}{\partial V_d}. \tag{3.4.3}$$

In Fig. 3.14, the current due to the source conductance flows in the opposite direction to those due to the drain conductance and the gate transconductance, because the current decreases when the source voltage increases.

The total change in the current i due to small variations in V_g, V_d, and V_s, (referenced to the bulk V_b) is

$$\begin{aligned} i &= \frac{\partial I}{\partial V_g}\Delta V_g + \frac{\partial I}{\partial V_s}\Delta V_s + \frac{\partial I}{\partial V_d}\Delta V_d \\ &= g_{mg}\Delta V_g - g_{ms}\Delta V_s + g_{md}\Delta V_d. \end{aligned} \tag{3.4.4}$$

This equation can be recast in terms of conductances that are referenced to the source:

$$i = g_m \Delta V_{gs} + g_{mb} \Delta V_{bs} + g_{ds} \Delta V_{ds}. \tag{3.4.5}$$

The relationships between these conductances and the bulk-referenced conductances are

$$\begin{aligned} g_m &= \frac{\partial I}{\partial V_{gs}} = g_{mg} \\ g_{mb} &= \frac{\partial I}{\partial V_{bs}} = g_{ms} - g_{mg} - g_{md} \\ g_{ds} &= \frac{\partial I}{\partial V_{ds}} = g_{md} \end{aligned} \tag{3.4.6}$$

where g_m is the gate transconductance, g_{mb} is the body or substrate transconductance, and g_{ds} is the drain-to-source conductance. These bulk-referenced conductances are summarised in Table 3.2 for the transistor operating in subthreshold and in Table 3.3 for the transistor operating in above threshold.

Bulk-Referenced Conductances

Gate Transconductance In *subthreshold*, the gate transconductance, g_{mg} (or g_m) can be derived by differentiating Eq. 3.2.10 with respect to V_g:

$$g_{mg} = \kappa \frac{I_0}{U_T} e^{(\kappa V_g - V_s)/U_T} = \frac{\kappa I}{U_T}. \tag{3.4.7}$$

In *above-threshold*, the gate transconductance in the linear region can be derived from Eq. 3.2.38:

$$g_{mg} = \beta (V_d - V_s), \tag{3.4.8}$$

and in the saturation region, from Eq. 3.2.39:

$$\begin{aligned} g_{mg} &= \beta \kappa (V_g - V_T) \\ &= \beta (\kappa (V_g - V_{T0}) - V_s). \end{aligned} \tag{3.4.9}$$

Table 3.3
Conductances of Above Threshold nFET at Low Frequencies.

Conductance	Above threshold	
	Triode	Saturation
g_{mg}	$\beta(V_d - V_s)$	$\beta(\kappa(V_g - V_{T0}) - V_s)$
g_{ms}	$\frac{\beta}{\kappa}(\kappa(V_g - V_{T0}) - V_s)$	$\frac{\beta}{\kappa}(\kappa(V_g - V_{T0}) - V_s)$
g_{md}	$\frac{\beta}{\kappa}(\kappa(V_g - V_{T0}) - V_d)$	$\frac{I}{V_e}$

Drain and Source Conductances In *subthreshold*, the drain and source conductances can be derived from Eq. 3.2.10:

$$\begin{aligned} g_{ms} &= -\frac{\partial I}{\partial V_s} = \frac{I_f}{U_T} \\ &= \frac{I_0 e^{(\kappa V_g - V_s)/U_T}}{U_T} \end{aligned} \tag{3.4.10}$$

$$\begin{aligned} g_{md} &= \frac{\partial I}{\partial V_d} = \frac{I_r}{U_T} + \frac{I}{V_e} \\ &= \frac{I_0 e^{(\kappa V_g - V_d)/U_T}}{U_T} + \frac{I}{V_e} \end{aligned} \tag{3.4.11}$$

where the second term in Eq. 3.4.11 expresses the channel-length modulation effect[23] in the saturation region. The drain conductance reduces to $g_{md} = \frac{I}{V_e}$.

In above-threshold linear regime, we use Eq. 3.2.35 to solve for these conductances:

$$g_{ms} = \frac{\beta}{\kappa}(\kappa(V_g - V_{T0}) - V_s) \tag{3.4.12}$$

$$g_{md} = \frac{\beta}{\kappa}(\kappa(V_g - V_{T0}) - V_d). \tag{3.4.13}$$

In the saturation regime, g_{ms} is also given also by Eq. 3.4.12, while g_{md} is given by the slope of the I–V_d curve:

$$g_{md} = \frac{I}{V_e}. \tag{3.4.14}$$

Source-Referenced Conductances

The gate transconductance g_m and the drain-to-source conductance g_{ds} are equal to the gate transconductance g_{mg} and the drain conductance g_d respectively. The latter two conductances were derived in Section 3.4. The bulk

23 This effect is described in detail in Section 3.5.

transconductance, g_{mb} can be derived easily as shown in Eq. 3.4.6. In *sub-threshold*, the bulk transconductance is

$$\begin{aligned} g_{mb} &= g_{ms} - g_{mg} - g_{md} \\ &= \frac{(1-\kappa)I}{U_T}. \end{aligned} \tag{3.4.15}$$

In the above-threshold linear regime,

$$g_{mb} = \beta(\kappa(V_g - V_{T0}) - V_s)(1 - \frac{1}{\kappa}) - \frac{I}{V_e}, \tag{3.4.16}$$

and in the saturation regime, $g_{mb} = 0$.

3.5 Second-Order Effects

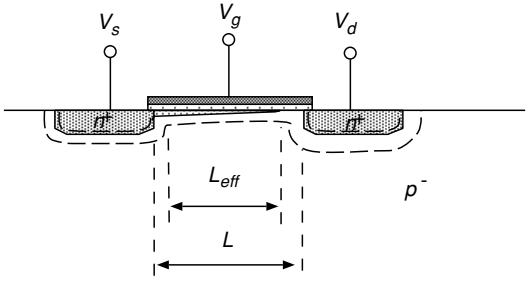


Figure 3.15
The effective channel length L_{eff} of a transistor operating in the above-threshold saturation region decreases with increasing V_d because the pinchoff point moves into the channel, away from the drain. The effective channel length can be described by the transistor length minus the length of the pinchoff region in the channel.

In the derivation of the current-voltage characteristics of the MOSFET, we assumed that certain properties (for example, the electron and hole mobilities) are constant under all operating conditions. These ideal assumptions do not apply in a real device. Some examples of the non-idealities in a MOSFET are described below.

Early Effect

When deriving the I–V characteristics of the nFET we assumed that the current is constant in the saturation regime. This assumption is not sufficient, particularly for short-length MOSFETs[24]. The drain voltage can modulate the channel current even in saturation. In the above-threshold saturation regime, the effective length of the transistor, L_{eff}, decreases when V_d increases because the pinchoff region extends further along the channel away from the drain (see Fig 3.15). We describe L_{eff} as the difference between the drawn channel length, L, and the size of the pinchoff region. Hence, the transistor current increases with V_d.

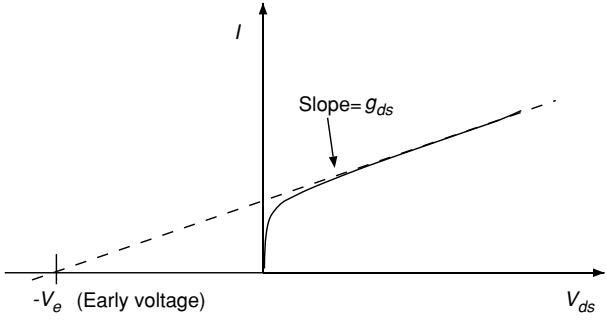


Figure 3.16
Plot of current versus drain-to-source voltage, showing the slope of the curve g_{ds} in the saturation regime. The intersection of the slope with the V_{ds} axis is called the Early voltage.

The same phenomenon occurs in subthreshold as is evident from the finite slope of the I–V_{ds} curve in Fig. 3.7. This slope is the *output conductance* of the transistor:

$$g_{ds} = \frac{\partial I}{\partial V_{ds}} = \frac{\partial I}{\partial L_{eff}} \frac{\partial L_{eff}}{\partial V_{ds}}. \tag{3.5.1}$$

By taking the derivative of Eq. 3.2.7 with respect to L_{eff}, we can make the replacement, $\frac{\partial I}{\partial L_{eff}} = -\frac{I}{L_{eff}}$ and rewrite Eq. 3.5.1 as

$$g_{ds} = -\frac{I}{L_{eff}} \frac{\partial L_{eff}}{\partial V_{ds}} = \frac{I}{V_e} \tag{3.5.2}$$

24 As we will see later, when computing the gain of a two-transistor amplifier, it is important to use a long transistor for high gain.

where $V_e = -L_{eff} \frac{\partial V_{ds}}{\partial L_{eff}}$ is the absolute value of the voltage at which this slope intersects the voltage axis of the curve in Fig. 3.16. This voltage is called the *Early voltage*[25]. In the subthreshold domain, the saturation current, I_{sat}, which is given by Eq. 3.2.16 can be re-expressed as

$$\boxed{I = I_{sat} + g_{ds} V_{ds} = I_{sat}(1 + \frac{V_{ds}}{V_e})} \tag{3.5.3}$$

to account for the Early effect.

Figure 3.17 shows the measured Early voltage plotted against transistor length, for transistors operating in subthreshold. The Early voltage varies from 20 V to 750 V and shorter transistors have a smaller Early voltage. This dependence of the channel current on V_d is sometimes called the *channel-length modulation effect*:. The modulation also occurs in transistors operating above threshold. This effect is sometimes modeled as a reduction in the effective threshold of the transistor, or as a reduction in the energy barrier at the source. The latter is similar to the *drain-induced barrier lowering (DIBL) effect*[26].

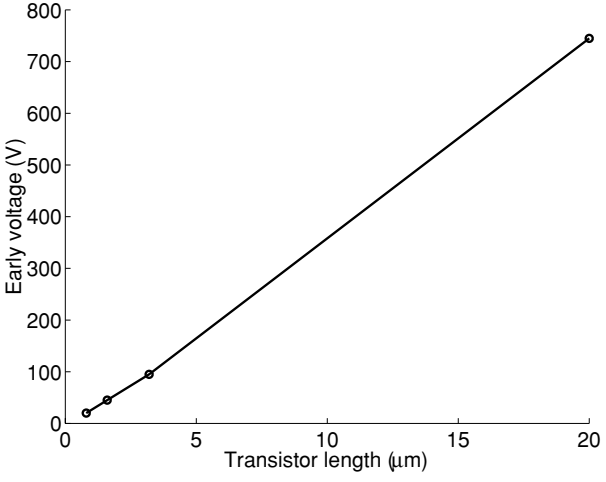


Figure 3.17
Early voltage versus the transistor length of an nFET fabricated in a 0.8 μm CMOS process.

25 It is named after Jim Early who first analyzed this effect in BJTs.

26 See Chapter 14 for a description of the DIBL effect.

Mobility Dependence on the Vertical Field

In Section 3.2, we assumed that the electron mobility is constant. In reality, the effective electron mobility, μ_{eff}, is a function of the electric field in the channel. The horizontal component of the field drives the electrons from the source to the drain while the vertical component (which depends on V_g) attracts the electrons to the oxide-silicon interface. When the vertical field increases, the electrons collide more frequently with the interface, leading to a lower effective mobility. The effective mobility can be derived from the current versus V_{gs} relation of an nFET operating in the deep triode above-threshold regime. The equation governing this relation is $I = \mu_{eff} C_{ox} \frac{W}{L} (V_g - V_T) V_{ds}$.

Velocity Saturation

Previously we assumed that the drift velocity v_d of the electrons is linearly proportional to the longtitudinal (or lateral) field component, $\mathcal{E}$. This assumption is only valid if the field is small. For fields above a critical value $\mathcal{E}_c$ the velocity saturates at a constant v_s, which has the same magnitude as the thermal velocity. This velocity saturation effect is also described in Section 2.5. The critical value at which the velocity saturates depends on the doping concentration. As a first-order approximation, the carrier-velocity saturation equation is

$$v_d = \frac{\mu \mathcal{E}}{1 + \mathcal{E}/\mathcal{E}_c}. \tag{3.5.4}$$

Because of velocity saturation, the current saturates at a smaller V_{ds} for a given V_{gs} than predicted by the current–voltage equations. This effect becomes more pronounced for small-channel length transistors.

Narrow- and Short-Channel Effects

In our current–voltage derivations, we assumed that the channel of the transistor is sufficiently long and wide that we could neglect any "edge" effects. As the dimensions of transistors continue to shrink, this assumption is no longer true. The field distribution in the channel becomes more complicated: The field components along the channel, and along the width axis of the transistor become significant. These changes result in short-channel and narrow-channel

effects[27]. In sub-micron processes, fabrication steps are engineered to minimize these effects (See Chapter 13.).

Narrow-Channel Effects

The narrow-channel effect (due to narrow-width transistors) can be modeled by an increase in the threshold voltage of the transistor. This effect is due to the extension of the depletion region underneath the gate toward the sides, under the *bird's beak*. The bird's beak is the part of the field oxide that encroaches into the transistor's channel. A cross-section of the transistor showing this encroachment can be seen in Fig. 13.1 in Chapter 13. The effective threshold voltage appears higher because some of the field lines from the gate end in the depletion region underneath the thick oxide (also called the *fringing field*) rather than in the depletion region underneath the gate. As a result, it takes a higher gate voltage to create the inversion layer beneath the gate.

Short-Channel Effects

We have previously discussed an example of a short-channel effect in Section 3.5. Because the Early effect was studied before the recognition of short-channel effects, it is often not recognised as such. Nonetheless, it is a short-channel effect, because the drain voltage changes the current by modulating the effective length of the transistor.

As the effective channel length decreases, the voltage drop across the pinchoff region increases leading to an increasing field around the drain. Eventually, velocity saturation of the carriers occurs as described in Section 3.5. This saturation leads to a decrease in the transistor current. As the field around the drain increases further, *hot-carrier effects* occur: The carriers in the channel become "hot" because they gain kinetic energy as they travel along the channel. Hot-carrier effects include impact ionization, avalanche multiplication, hot-electron injection, and punch-through. These effects are usually undesirable but the hot-electron injection mechanism can be used for interesting purposes, such as programming floating-gate memories.

Impact Ionization and Avalanche Multiplication In the presence of a high longtitudinal or lateral field, the carriers collide with the lattice, breaking the Si

27 A channel is short if the sum of the depletion region widths around the source and drain become comparable to the length. A channel is narrow when the width of the transistor becomes comparable to the depth of the depletion region under the gate.

bonds, and creating additional electron-hole pairs. This phenomenon is called *impact ionization* in an nFET. The generated electrons are swept into the drain and the generated holes are swept into the substrate creating a substrate current. The drain current is now higher than the source current ($I_d \neq I_s$). This effect prevents current sources from achieving a high output impedance. With even higher electric fields, the *impact ionization* process turns into an *avalanche* phenomenon: More and more electron-hole pairs are created even when the field is kept constant.

Hot-Electron Injection The electrons in the channel under the high longtitudinal field can gain sufficient kinetic energy to surmount the barrier at the interface of the silicon and the gate oxide and some carriers are injected into the oxide. Most of these electrons will cross into the gate and the remainder will be trapped in the oxide. The trapped charges alter the "threshold" of the MOSFET. This phenomenon is usually undesirable because it leads to a effective threshold voltage change. But, as we will see in Chapter 4, this mechanism can be used to our advantage: In the construction of a non-volatile memory.

Punch-through Punch-through is a form of transistor breakdown. This phenomenon occurs when the depletion regions around the drain and source merge. The gate becomes ineffective in controlling the current. The channel length becomes effectively zero so a huge current flows even at low V_{gs}.

3.6 Noise and Transistor Matching

To fully characterise the operation of a circuit, the different noise sources that contribute to the current in a transistor should be considered. This topic is discussed in detail in Chapter 11. Transistor mismatch is also a major concern, especially in analog circuit design. The mismatch can be reduced by using a large layout area for the transistors and also by operating the transistors above threshold. This means that good matching requires sacrifices in both chip area and power consumption.

If small device geometries are used, the transistors are susceptible to spatial variations from process-dependent parameters (see Section 12.2).

3.7 Appendices

A. Small-Signal Model at Moderate Frequencies

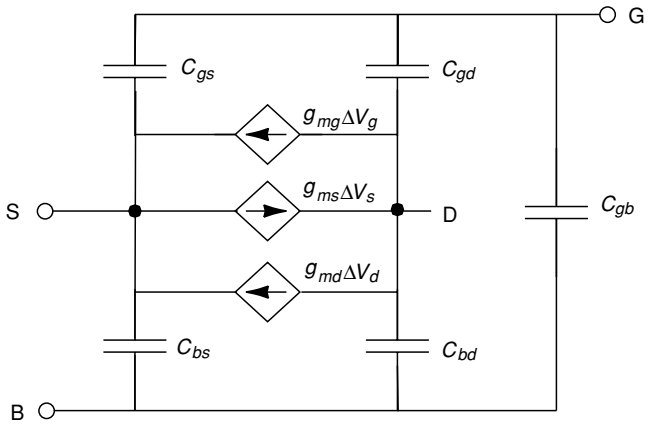


Figure 3.18
Small-signal equivalent model of a MOSFET operating at moderate frequencies.

The operation of the transistor is also determined by the charges on its different parts. In addition to the currents from the conductance parameters, finite capacitive currents arise due to voltage changes ΔV at any of its four terminals. In computing these capacitive currents, we assume that the transistor operates in the *quasi-static* condition: The change in the charge is proportional to the change in the voltage. These capacitive currents (which become important when the transistor operates at moderate frequencies) are incorporated into the small-signal model in Fig. 3.14 by the addition of capacitors (as shown in Fig. 3.18).

The MOSFET can divided into two parts; the *intrinsic* part, and the *extrinsic* part. The *intrinsic* part of a MOSFET is defined as the region between the source and drain (the inversion layer and the depletion region), and the gate oxide and the gate (see Fig. 3.19). The intrinsic part effects transistor action. The undesirable parasitic elements constitute the *extrinsic* part of a MOSFET: These elements include the drain and source resistances, the junction capacitances between the drain and source regions and the bulk, and the overlap capacitances between the gate and the source/drain.

The capacitive currents at the different nodes of a MOSFET become significant when the MOSFET is operated at moderate frequencies. We will consider the contribution of only the intrinsic capacitances to these currents. The voltages at the four terminals are referenced to a separate "ground". The voltages at three terminals are kept constant while the voltage at the fourth terminal (for example, V_s) is varied by ΔV_s. We consider how this voltage variation affects the charge on the gate (Q_g), the charge in the bulk (Q_b), and the inversion charge (Q_i). *Here, the symbol, Q is the total charge rather than the charge per unit area. Hence Q_g is the total charge on the gate.* The five trans-capacitors that describe the change in the charges as a function of the terminal voltages are shown in Fig. 3.18[28]. *The capacitance symbol C defines the total capacitances rather than the capacitance per unit area.*

The gate-to-source capacitance, C_{gs}, represents the change in the gate charge, ΔQ_g, as a function of a change in V_s for a fixed V_g, V_b, and V_d:

$$C_{gs} = -\frac{\partial Q_g}{\partial V_s}|_{V_g,V_b,V_d}. \tag{3.7.1}$$

The charge ΔQ_g represents the change in the gate charge when V_s increases. A positive increase in V_s means that the potential across the gate oxide decreases: So ΔQ_g is negative, and C_{gs} is then positive (Eq. 3.7.1). Similar capacitances can be associated with the gate for a given change in V_d and V_b respectively:

$$C_{gd} = -\frac{\partial Q_g}{\partial V_d}|_{V_g,V_b,V_s} \tag{3.7.2}$$

$$C_{gb} = -\frac{\partial Q_g}{\partial V_b}|_{V_g,V_d,V_s}. \tag{3.7.3}$$

A positive ΔV_s also affects the depletion charge in the bulk. Because ΔV_s is positive, the depletion region around the source increases, making the total charge more negative (more acceptor ions are exposed in a p^- substrate). This charge is balanced by an equivalent positive charge Q_b leaving the substrate. The capacitance C_{bs} represents the negative change in Q_b resulting from the increase in V_s:

$$C_{bs} = -\frac{\partial Q_b}{\partial V_s}|_{V_g,V_b,V_d}. \tag{3.7.4}$$

28 These capacitances are not associated with a physical parallel-plate structure in the channel.

Similarly the bulk-drain capacitance is:

$$C_{bd} = -\frac{\partial Q_b}{\partial V_d}|_{V_g, V_b, V_s}. \tag{3.7.5}$$

A good discussion of these capacitances in both the subthreshold and above threshold regimes of MOSFET operation can be found in Enz et al. (1995); Tsividis (1998).

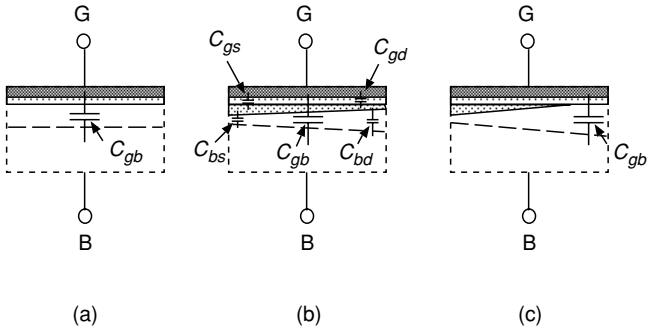


Figure 3.19
Intrinsic capacitances of a nFET in weak inversion and strong inversion. (a) In weak inversion, there is one dominant capacitance C_{gb}. When the FET operates above threshold, additional capacitances appear both in the linear region (b) and in the saturation region (c).

Intrinsic Capacitances in Subthreshold

In subthreshold, the inversion charge Q_i underneath the gate is negligible. Any variation in the gate charge caused by a change in the terminal voltages affects primarily the bulk charge. An explicit expression for four of the capacitances in the subthreshold region can be found in Enz et al. (1995) and Tsividis (1998). However, to a first-order approximation, we assume that these capacitances are approximately zero. The gate-to-bulk capacitance, C_{gb}, is the combination of two C_{ox} and C_d in series:

$$\begin{aligned} C_{gb} = C_{ox} \,||\, C_d &= C_{ox}\left(\frac{C_d}{C_{ox} + C_d}\right) \\ &= (1 - \kappa) C_{ox}. \end{aligned} \tag{3.7.6}$$

Intrinsic Capacitances Above Threshold

Above threshold, the intrinsic capacitances depend on whether the MOSFET is operating in the triode or the saturation regime.

Triode Regime Here, the inversion layer is approximately constant throughout the channel. Any change in V_s or V_d affects the inversion charge, and hence the gate charge. The total capacitance between the channel and the gate is given by the oxide capacitance, C_{ox}. By symmetry

$$C_{gs} = C_{gd} = \frac{C_{ox}}{2}, \tag{3.7.7}$$

and similarly,

$$C_{bs} = C_{bd} = \frac{C_b}{2} \tag{3.7.8}$$

where C_b is the total capacitance of the reversed-bias depletion region formed by the inversion layer and the bulk. C_{gb} is equal to zero because the inversion layer shields the bulk from the influence of the gate.

Saturation Regime In this case, the inversion charge disappears at the drain end of the channel. Changing V_d has no effect on the intrinsic charges, so $C_{bd} = C_{gd} = 0$. However the source voltage affects the inversion layer, so C_{bs} and C_{gs} are non-zero. The inversion layer decreases towards the drain end of the channel, and the channel is depleted near the drain. Consequently, the region of the channel near the drain is not affected by changes in V_s. From detailed analysis in Tsividis (1998), $C_{gs} = \frac{2}{3}C_{ox}$, and $C_{bs} \approx \frac{2}{3}C_b$. Because of the pinchoff (depletion) region near the drain, C_{gb} is equal to the series combination of two capacitances:

$$C_{gb} = 1/3(C_{ox} \,||\, C_d) = \frac{1}{3}C_{ox}(1-\kappa). \tag{3.7.9}$$

Unity-Gain Cutoff Frequency

The gate current of a MOSFET is close to zero at low frequencies, but at higher frequencies, it becomes significant due to the current flow through the capacitances (C_{gs}, C_{gb}, and C_{gd}) associated with the gate. The small-signal equivalent model for the transistor in Fig. 3.20(a) is shown in Fig. 3.20(b). The *unity current gain cutoff frequency*, f_T, is the frequency at which the input current (gate current) and output current (drain current) are equal. This cutoff

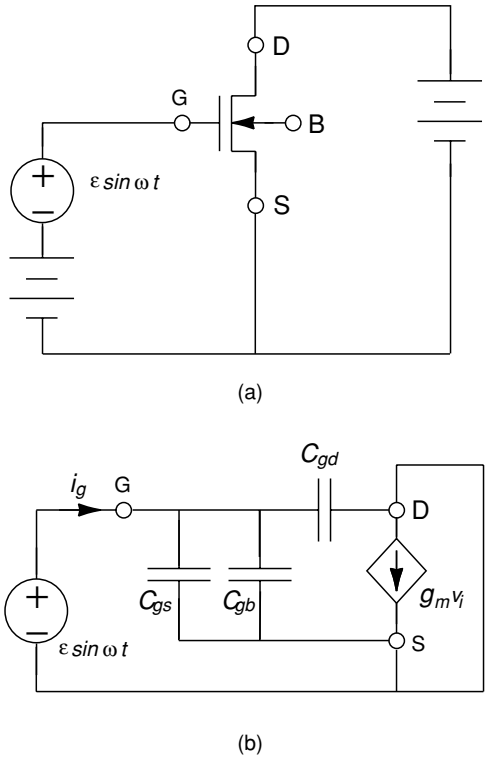


Figure 3.20
Small-signal model of an nFET for unity-gain frequency calculation. (a) nFET operating in saturation with a small-signal input voltage. (b) Small-signal equivalent circuit of (a).

frequency specifies the maximum operating frequency of the transistor.

f_T is computed by supplying a small-signal input $v_i = \epsilon \sin \omega t$ to the gate[29]. The small-signal drain current, i_d, is given by $g_m v_i$ and the gate current is

$$i_g = j2\pi f(C_{gs} + C_{gb} + C_{gd})v_i \approx j2\pi f C_{gs} v_i. \tag{3.7.10}$$

We can ignore C_{gb} and C_{gd} because $C_{gb}, C_{gd} < C_{gs}$. Solving for f_T with

29 The radian frequency, ω is given in radians per second and is equal to $2\pi f$ where f is the frequency in Hertz.

$|i_d| = |i_g|$, we get

$$\boxed{f_T = \frac{g_m}{2\pi C_{gs}}.} \tag{3.7.11}$$

For an above threshold transistor in saturation, the unity-gain frequency is

$$\boxed{f_T \approx \frac{3\mu_{eff}(V_{gs} - V_T)}{4\pi L^2}.} \tag{3.7.12}$$

We see that f_T is inversely proportional to the square of the channel length. To obtain a high f_T, we need a high carrier mobility and a short-channel transistor[30].

B. The Enz-Krummenacher-Vittoz Transistor Model

There are a small number of models that describe the transistor's operation continuously from subthreshold to above threshold. These models include those of Maher and Mead (Mead, 1989; Maher, 1989) and others (Tsividis, 1998; Enz et al., 1995; Enz and Vittoz, 1997; Montoro et al., 1999). In Maher and Mead's model, the current flow in a MOSFET is computed from the mobile charge distribution in the channel as the terminal voltages are varied. This model is used in the SPICE (Simulation Program with Integrated Circuit Emphasis) simulation package from *Tanner Tools*. This circuit simulation program can be used to perform many different types of analysis such as steady-state, small-signal, time-domain, frequency, temperature, and noise analysis.

The Enz-Krummenacher-Vittoz (EKV) model (Enz et al., 1995) can also be used with SPICE. This model provides a simple, closed-form expression for the channel current of a MOS transistor in terms of the terminal voltages, each of which is referenced to the transistor's bulk voltage. It is a continuous model that is valid in all normal regimes of MOS transistor operation, that is, the drain-bulk and the source-bulk junctions are reverse-biased. The model is also symmetric with respect to the interchange of source and drain terminals. The channel current is

$$I = I_f - I_r \tag{3.7.13}$$

where I_f and I_r are the *forward* and *reverse* components of the channel current respectively. The forward component is a function only of the gate-to-bulk and source-to-bulk voltages, whereas the reverse component is a function only of

30 We compute an approximate value for f_T. When $L = 0.8\mu$m ,V_{gs}=3V, V_T=1V, $\mu_{eff} = 400 cm^2/(V.s)$, $f_T \approx$ 30 GHz.

the gate-to-bulk and drain-to-bulk voltages. Under this formulation a MOS transistor operates in saturation when $I_f \gg I_r$, which implies that $I \approx I_f$. In this case, the channel current (neglecting the Early effect) is not a function of the drain voltage. For a MOS transistor operating in the ohmic regime, $I_f \approx I_r$.

In the EKV model with the nFET biased as shown in Fig. 3.4a, the forward component of the channel current is

$$\begin{aligned} I_f &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa(V_{gb} - V_{T0}) - V_{sb})/2U_T}\right) \\ &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa V_g + (1-\kappa) V_b - \kappa V_{T0} - V_s)/2U_T}\right) \end{aligned}$$

where V_{T0} is the zero-bias threshold voltage (referenced to the bulk). The reverse component, I_r, is given by a similar expression

$$\begin{aligned} I_r &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa(V_{gb} - V_{T0}) - V_{db})/2U_T}\right) \\ &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa V_g + (1-\kappa) V_b - \kappa V_{T0} - V_d)/2U_T}\right). \end{aligned}$$

The function $\log^2\left(1 + e^{x/2}\right)$[31] interpolates smoothly between an exponential (subthreshold behavior) when $x < 0$ and a quadratic (above-threshold behavior) when $x > 0$. Thus, when $V_{gb} < V_{T0} + V_{sb}/\kappa$, the forward component of the channel current becomes

$$I_f \approx \frac{W}{L} I_0 e^{(\kappa V_g + (1-\kappa) V_b - V_s)/U_T} \tag{3.7.14}$$

where I_0 is the subthreshold pre-exponential current factor (Mead, 1989), given by

$$I_0 = 2U_T^2 \mu C_{ox} e^{-\kappa V_{T0}/U_T}/\kappa.$$

On the other hand, when $V_{gb} > V_{T0} + V_{sb}/\kappa$, the forward component of the channel current becomes

$$I_f \approx \frac{W}{L} \frac{\mu C_{ox}}{2\kappa} \left(\kappa V_g - (1-\kappa) V_b - \kappa V_{T0} - V_s\right)^2. \tag{3.7.15}$$

Likewise, the reverse component of the channel current becomes

$$I_r \approx \frac{W}{L} I_0 e^{(\kappa V_g + (1-\kappa) V_b - V_d)/U_T} \tag{3.7.16}$$

31 This function was chosen for correct mathematical and qualitative properties and not for any physical reason.

when $V_{db} > \kappa\left(V_{gb} - V_{T0}\right)$, and it becomes

$$I_r \approx \frac{W}{L}\frac{\mu C_{ox}}{2\kappa}\left(\kappa V_g - (1-\kappa)V_b - \kappa V_{T0} - V_d\right)^2 \tag{3.7.17}$$

when $V_{db} < \kappa\left(V_{gb} - V_{T0}\right)$.

When $V_{gb} < V_{T0} + V_{sb}/\kappa$ and $V_{db} > \kappa\left(V_{gb} - V_{T0}\right)$, we can recover the usual subthreshold MOS transistor model (Mead, 1989) by substituting Eqs. 3.7.14 and 3.7.16 into Eq. 3.7.13 and factoring out the common part of each term. This way, we obtain

$$I = \frac{W}{L} I_0 e^{(\kappa V_g + (1-\kappa)V_b)/U_T}\left(e^{-V_s/U_T} - e^{-V_d/U_T}\right) \tag{3.7.18}$$

which covers both the the ohmic regime (when $V_{ds} < 4U_T$) and the saturation regime (when $V_{ds} > 4U_T$) in subthreshold. To see that it does so, we write

$$\begin{aligned} I &= \frac{W}{L} I_0 e^{(\kappa V_g + (1-\kappa)V_b - V_s)/U_T}\left(1 - e^{-V_{ds}/U_T}\right) \\ &\approx \frac{W}{L} I_0 e^{(\kappa V_g + (1-\kappa)V_b - V_s)/U_T} \end{aligned} \tag{3.7.19}$$

because if $V_{ds} > 4U_T$, then $e^{-V_{ds}/U_T} < e^{-4} \ll 1$ and so $1 - e^{-V_{ds}/U_T} \approx 1$, thus removing the dependence of the channel current on the drain voltage.

When $V_{gb} > V_{T0} + V_{sb}/\kappa$ and $V_{db} < \kappa\left(V_{gb} - V_{T0}\right)$, the transistor is operating in the above-threshold ohmic regime and by substituting Eqs. 3.7.15 and 3.7.17 into Eq. 3.7.13 and factoring out the common part of each term, we obtain an expression for the channel current:

$$\begin{aligned} I = \frac{W}{L}\frac{\mu C_{ox}}{2\kappa}\big(&\left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_s\right)^2 - \\ &\left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_d\right)^2\big). \end{aligned} \tag{3.7.20}$$

If $V_{db} > \kappa\left(V_{gb} - V_{T0}\right)$, then the drain end of the channel will be only weakly inverted and $I_f \gg I_r$. In this case, the transistor is operating in the above-threshold saturation regime and the channel current is given by the first term of Eq. 3.7.20:

$$I = \frac{W}{L}\frac{\mu C_{ox}}{2\kappa}\left(\kappa V_g + (1-\kappa)V_b - \kappa V_{T0} - V_s\right)^2. \tag{3.7.21}$$

The (bulk-referenced) saturation voltage is thus

$$V_{sat} = \kappa\left(V_{gb} - V_{T0}\right). \tag{3.7.22}$$

Taken together, Eqs. 3.7.20 and 3.7.21 constitute an above-threshold model

that differs from those found in most elementary circuit texts. This model has the advantage that it exposes the source-drain symmetry of the MOS transistor. It also makes clear how the above-threshold ohmic and saturation equations are related to one another. Moreover, this model accounts directly for the body effect (to first order) via the κ parameter without any auxiliary equations and without adding too much complexity to the model. For a pFET, biased as shown in Fig. 3.4b, the forward component of the channel current is

$$\begin{aligned} I_f &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa(V_{bg} - |V_{T0}|) - V_{bs})/2U_T}\right) \\ &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(V_s - \kappa V_g - (1-\kappa)V_b - \kappa|V_{T0}|)/2U_T}\right) \end{aligned}$$

where all parameters are defined as for the nFET, except that μ is the mobility of holes. Likewise, the reverse component is

$$\begin{aligned} I_r &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(\kappa(V_{bg} - |V_{T0}|) - V_{bd})/2U_T}\right) \\ &= \frac{W}{L} 2U_T^2 \frac{\mu C_{ox}}{2\kappa} \log^2 \left(1 + e^{(V_d - \kappa V_g + (1-\kappa)V_b - \kappa|V_{T0}|)/2U_T}\right). \end{aligned}$$

C. Bipolar Junction Transistor

Traditionally, microcircuits were implemented with bipolar junction transistors (BJTs); but in the past 30 years MOSFETs have replaced BJTs in microcircuits except for high-speed applications or in analog applications that require low 1/f noise or precise transistor matching. Bipolar CMOS (BiCMOS) processes are available where both bipolar and MOS transistors are fabricated on the same substrate. These technologies are attractive for the implementation of mixed analog-digital designs. There are two types of BJTs: The **npn**, and the **pnp** transistors, whose symbols are shown in Fig. 3.21. We give a cursory treatment of the basic operation of an **npn** bipolar transistor. As we will see in Chapter 10, the bipolar transistor can also be used as a photo-sensing element (phototransistor).

Basic Operation

The structure of an **npn** bipolar transistor is shown in Fig 3.22. It consists of two back to back pn junctions. The emitter and collector terminals are both

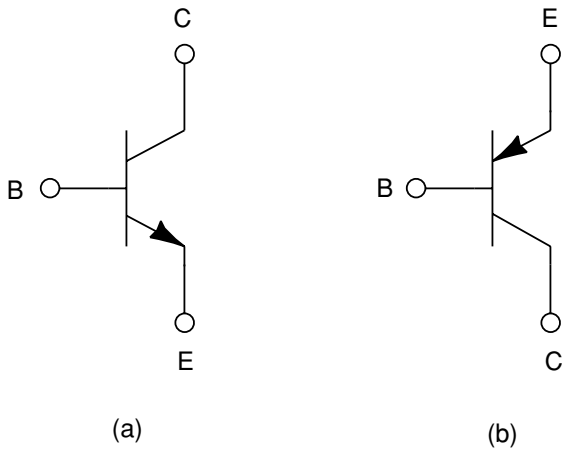


Figure 3.21
Bipolar transistor symbol. (a) An **npn** bipolar transistor. (b) An **pnp** bipolar transistor.

n-type regions but the emitter is doped higher than the collector. The base is doped p type. In the normal operating mode, the emitter-base junction is forward-biased and the collector-base junction is reverse-biased. If the base-emitter junction is reverse-biased (V_{be} $<0.5\,\text{V}$) the transistor is in *cutoff*. When the base-emitter voltage V_{be} is greater than approximately $0.7\,\text{V}$, current begins to flow from the base to the emitter. The current flow consists of holes moving from the p base to the emitter and electrons moving from the n^+ emitter to the base. Because the emitter is doped higher than the base, more electrons are injected from the emitter into the base. The collector-base junction is reversed biased, and so the holes from the base are not attracted to the collector. The base width is narrow, so that electrons injected from the emitter will diffuse easily into the depletion region at the collector-base junction. Once they are in this region, they are swept into the collector. Most of the electrons from the emitter reach the collector, consequently, the collector current I_c is approximately equal to the emitter current I_e. However some of the injected electrons recombine with the holes in the base. This loss of electrons is equivalent to the base current. For an almost ideal BJT, we require the base current I_b to be much smaller than I_e. The common-emitter current gain, β, of the device is defined as

$$\beta = \frac{I_c}{I_b}. \tag{3.7.23}$$

The value of β varies between 50 and 100. β depends on the base width and the lifetime of the carriers in the base. It also depends on the magnitude of the collector current

$$I_c = I_{cs}(e^{V_{be}/U_T} - 1) \tag{3.7.24}$$

where I_{cs} is the pre-exponential constant.

An in-depth discussion of the operation of BJTs can be found in many texts (Grove, 1967; Gray et al., 2001). The bipolar transistor has different characteristics to the MOSFET. Because of its finite base current and its higher output current drive, it has a lower input impedance than the MOSFET. However, the BJT and the subthreshold MOSFET share similar exponential current-voltage characteristics.

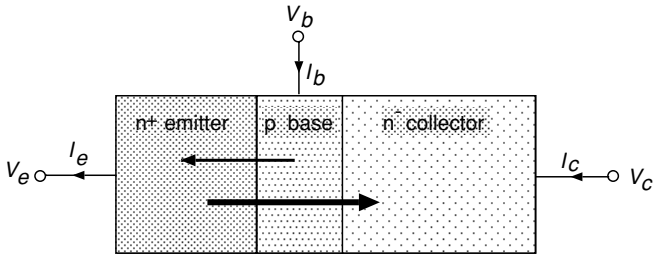


Figure 3.22
Current components in an **npn** bipolar junction transistor.

This page intentionally left blank

4 Floating-Gate MOSFETs

Our goal in this chapter is to examine the physics of floating-gate devices, and develop the technology of analog memory transistors.

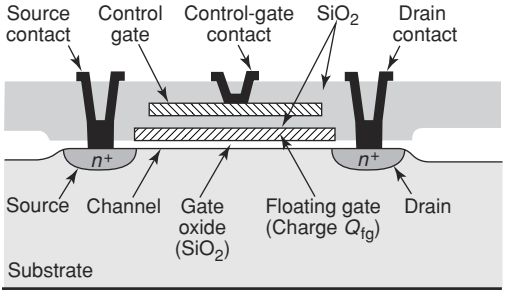
4.1 Floating-Gate MOSFETs

The semiconductor industry has developed a variety of nonvolatile solid-state memory devices, including the EPROM (electrically programmable read only memory), EEPROM (electrically erasable, programmable, read-only memory), and flash EEPROM. A common feature of these memories is that they use a *floating-gate MOSFET* (FGMOSFET) whose gate is isolated because it is surrounded on all sides by an insulator (typically SiO_2). The basic organization of such a floating gate is shown in Fig. 4.1(a).

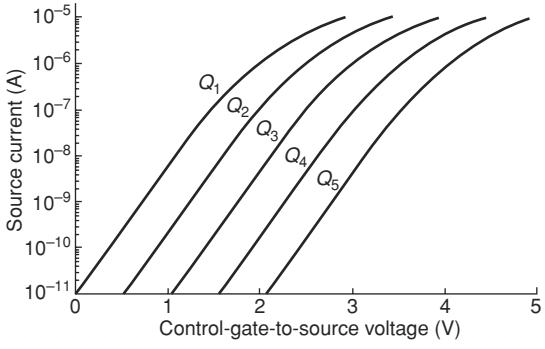
Electrons on the floating gate are prevented from escaping by the surrounding insulator. Voltage inputs to a secondary *control gate* couple capacitively to the floating gate, thereby modulating the transistor's channel current. From the perspective of this control gate, the presence (or absence) of excess charge on the floating gate causes a voltage offset and so an apparent shift in the MOSFET's transfer function (Fig. 4.1(b)).

The earliest experiments exploring charge storage on the insulated gate of a MOSFET were performed in the 1960s (Kahng and Sze, 1967; Kahng, 1967; Frohman-Bentchkowsky, 1971). At that time, the goal was to find a replacement for magnetic (core) memories and so work focused on the use of floating gates for binary storage. Digital EEPROM chips achieved commercial success roughly a decade later. In the 1980s, Alspector and Allen suggested that the memory function in a neural network could be realized by using analog charge on a floating gate (Alspector and Allen, 1987; Yang et al., 1992; Fujita and Amemiya, 1993). Unfortunately, changing the charge on a floating gate involves the difficult task of transporting electrons through the SiO_2 insulator. It has taken yet another decade to achieve a measure of success in solving this problem. Even now, analog floating-gate devices are confined primarily to the research laboratory.

Neuromorphic engineers have adapted the floating-gate technology used in digital EEPROMs both to allow nonvolatile analog storage, and to perform a local learning function. In particular, a family of single-transistor floating-gate devices called *synapse transistors* have been developed (Diorio, 1997;



(a)



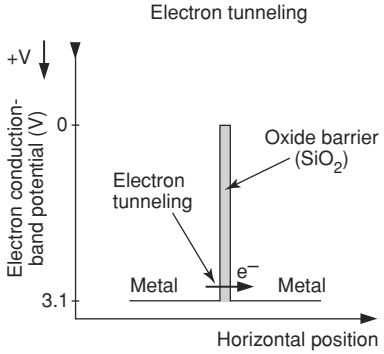
(b)

Figure 4.1
An *n*-channel floating-gate MOSFET (a) and its associated *I–V* transfer function (b). Voltage inputs to the poly2 control gate are coupled capacitively to the poly1 floating gate. From the control gate's perspective, changing the floating-gate charge $Q_{f\,g}$ shifts the transistor's threshold voltage (bidirectionally).

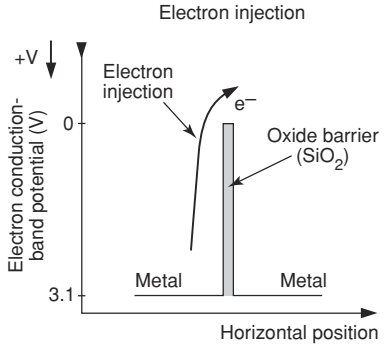
Hasler, 1997; Diorio et al., 1996, 1997c, 1998b,a, 1997a). These devices, like neural synapses, implement long-term nonvolatile analog memory; allow bidirectional memory updates; and learn from an input signal without interrupting the ongoing computation.

Most electrically programmable solid-state memory technologies use two basic mechanisms to modify the floating-gate charge[1] (Kerns et al., 1991).

1 Other floating gate devices, such as programmable read-only memories (PROMs), uses UV light



(a)



(b)

Figure 4.2
Overcoming an SiO_2 barrier. An electron can either (a) tunnel through the barrier, provided the barrier is thin enough; or it can (b) acquire enough energy to inject over the barrier.

These two mechanisms are illustrated in the energy-band diagrams of Fig. 4.2. Stated simply, moving electrons through SiO_2 requires overcoming the difference in electron affinities between a metal[2] and the SiO_2. We can either

to read and/or write the floating-gate memory (Kerns et al., 1991). In this case, UV photons excite electrons to energy levels high enough to overcome the difference in electron affinities between the metal and the SiO_2.

2 The "metal" can be an actual metal, a polysilicon gate, or a degenerately doped silicon implant.

push the electrons *through* the potential barrier; or force them *over* the top of the barrier. These two processes are called electron tunneling and hot-electron injection, respectively (Takeda et al., 1995). Electron tunneling is a quantum-mechanical process; we restrict our study to a particular form, called Fowler-Nordheim (FN) tunneling (Lenzlinger and Snow, 1969), in which an electric field is applied across the SiO_2 to facilitate tunneling. Hot-electron injection comes in many flavors that differ in the mechanism by which the electrons are made "hot". Synapse transistors use both tunneling and injection to alter their floating-gate charge.

Electron Tunneling

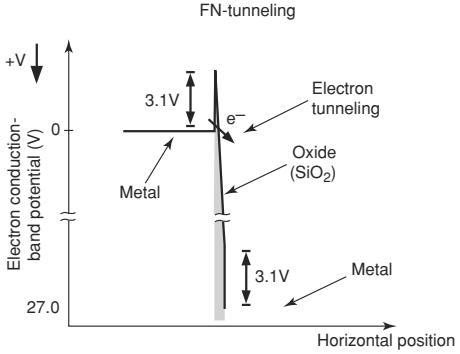
The FN-tunneling process is illustrated in Fig. 4.3(a). In practice, the barrier is SiO_2 gate oxide; the "metal" on one side of the barrier is the polysilicon floating gate; and the "metal" on the other side is a heavily doped region of the silicon surface called a *tunneling implant*. We tunnel electrons from the floating gate to the tunneling implant. A potential difference between the floating gate and the implant reduces the effective thickness of the SiO_2 barrier, facilitating electron tunneling from the floating gate, through the barrier, into the oxide conduction band. These electrons are then swept over to the implant by the field across the barrier.

In general, we can tunnel electrons both onto and off the floating gate; and we can tunnel them through any oxide (gate oxide, inter-poly oxide, etc.). In practice we avoid tunneling electrons onto the floating gate, because to do so we must either apply a large negative voltage to the tunneling implant, or pull the floating gate to a high positive voltage. The former is unattractive because most commercial processes are single-tub *n*-well. The latter is also unattractive because of the large change in channel current during tunneling. We also avoid tunneling electrons through the inter-poly oxide, because the inferior oxide quality (compared with gate oxide) causes poor matching and unreliable behavior.

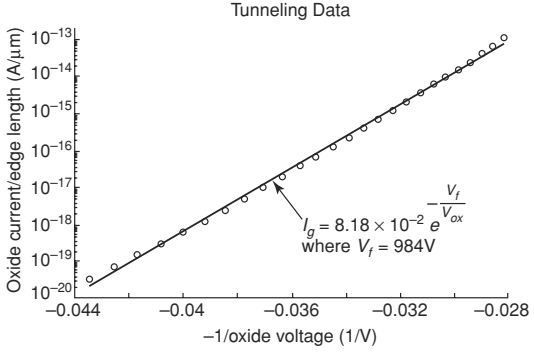
Figure 4.3(b) shows the measured FN-tunneling current versus the reciprocal of the voltage across the oxide, for a 400 Å gate oxide fabricated in a 2 μm CMOS process. We can approximate the data by

$$\boxed{I_g = I_{to}\, e^{-\frac{V_f}{V_{ox}}}} \tag{4.1.1}$$

where I_g is the gate current,V_{ox} is the oxide voltage, V_f is a constant that depends on the oxide thickness, and I_{to} is a pre-exponential current. Equa-



(a)



(b)

Figure 4.3
Fowler-Nordheim (FN) tunneling. (a) By applying a voltage across the oxide, the band diagram is altered, thereby facilitating electron tunneling through the "thinned" barrier into the oxide conduction band. (b) A plot of tunneling current versus reciprocal oxide voltage, measured from the tunneling junction shown in Fig. 4.4 (a). V_{ox} is the voltage across the gate oxide; Given an oxide thickness of 400 Å, V_f =984 V is consistent with a survey (Mead, 1994) of Fowler-Nordheim tunneling in SiO_2. We plot the data as oxide current divided by the gate-to-n^+ edge length (in lineal microns) of the tunneling implant because the floating gate induces a depletion region in the lightly doped n^- well, reducing the effective oxide voltage and with it the tunneling current. The gate cannot appreciably deplete the n^+ well contact, so the oxide field is higher where the self-aligned floating gate overlaps the n^+. Because tunneling increases exponentially with oxide voltage, tunneling in analog memory transistors is primarily an edge phenomenon.

tion 4.1.1 is a sufficient (for our purposes) approximation to a well-known FN-tunneling equation (see, for example, (Sze, 1981)).

Electron Injection

We defer illustrating the hot-electron injection process (Sanchez and DeMassa, 1991) until we have described the *n*FET and *p*FET synapses, because the injection process differs slightly between these two devices. We merely note at this time that we accelerate electrons to high energies in the drain-to-channel electric field of a transistor; that a fraction of these electrons scatter upward into the gate oxide; and that a fraction of these scattered electrons inject into the oxide's conduction band. The injected electrons can then be swept over to the floating gate by an electric field across the gate oxide.

4.2 Synapse Transistors

The programming characteristics of floating-gate devices are quite variable, even for nominally identical transistors on the same chip (Sin et al., 1992). For digital memories, these variations can be compensated by applying excess charge during the write or erase processes. For analog memories, these device variations require feedback to ensure accurate memory writes. Various feedback mechanisms have been tried: Most employ either multi-step, iterative writes (Sin et al., 1992); or single-step, open-loop writes with frequent oxide calibration to compensate mismatch and degradation (Holler et al., 1989). Neither approach permits the locally computed, parallel weight updates that are needed for silicon learning arrays.

Synapse transistors are new types of floating-gate devices that permit simultaneous reading and writing of memory. The channel *and* oxide currents are simultaneous, analog, and continuous. Consequently, we can use continuous analog feedback to write the charge on the floating gate, and to implement either a weight-update rule or a weight constraint in a silicon learning system.

Four-terminal *n*FET and *p*FET synapse transistors are shown in Fig. 4.4. Both comprise a single MOSFET (with poly1 floating gate and poly2 control gate) and an associated *n*-well tunneling implant. Both use hot-electron injection to add electrons to their floating gates, and FN tunneling to remove the electrons. The *n*FET synapse differs from a conventional *n*-type MOSFET by its use of a moderately-doped channel implant, which facilitates hot-electron injection (Diorio et al., 1997a). The *p*FET synapse achieves a hot-

electron gate-current (Chung et al., 1997) using a conventional p-type MOSFET: No special channel implant is required. Various researchers have fabricated nFET synapses in 2 μm and 1.2 μm n-well double-poly processes, and pFET synapses in several double-poly processes ranging from 2 μm to 0.35 μm.

Both the nFET and pFET synapses store a weight as charge on their floating gate. We can choose source current, drain current, or channel conductance as the output. Typically we choose source current, although this choice is arbitrary because these quantities are related. Signal inputs are applied to the poly2 control gate, which in turn couples capacitively to the poly1 floating gate. Typically the synapses are operated in their subthreshold regime (Mead, 1989), for three reasons. Firstly, subthreshold channel currents ensure low power consumption; typically less than 100 nW per device. Secondly, because the source current in a subthreshold MOSFET is an exponential function of the gate voltage, small changes to the floating-gate charge shift the transistor's operating point measurably. Thirdly, the synapse output is the product of a stored weight and the applied input:

$$I_s = I_o e^{\frac{\kappa V_{fg}}{U_t}} = I_o e^{\frac{\kappa (Q_{fg} + C_{in} V_{in})}{C_T U_t}} = I_o e^{\frac{Q_{fg}}{Q_T}} e^{\frac{\kappa' V_{in}}{U_t}} \tag{4.2.1}$$

$$= W I_o e^{\frac{\kappa' V_{in}}{U_t}} \tag{4.2.2}$$

where I_s is the source current, I_o is the pre-exponential current, κ is the coupling coefficient from the floating gate to the channel, Q_{fg} is the floating-gate charge, C_T is the total capacitance seen by the floating gate, U_t is the thermal voltage kT/q, C_{in} is the input (poly1 to poly2) coupling capacitance, V_{in} is the control-gate voltage; $Q_T \equiv C_T U_t/\kappa$, $\kappa' \equiv \kappa C_{in}/C_T$, $W \equiv \exp(Q_{fg}/Q_T)$; and, for simplicity, we assume the source potential to be ground (V_s=0). Equation 4.2.1 and Eq. 4.2.2 imply an n-channel MOSFET; if we change the sign of all the variables, the equations describe a p-channel MOSFET.

The weight W is a learned quantity. Its value derives from the floating-gate charge, which can change with synapse use. The floating-gate transistor output is the product of W and the source current of an idealized MOSFET, which has a control-gate input V_{in}, and a coupling coefficient κ' from the control gate to the channel.

Both the nFET and pFET synapses achieve bidirectional weight updates by using FN-tunneling to remove electrons from the floating gate, and by using hot-electron injection to add electrons to the floating gate. The magnitudes

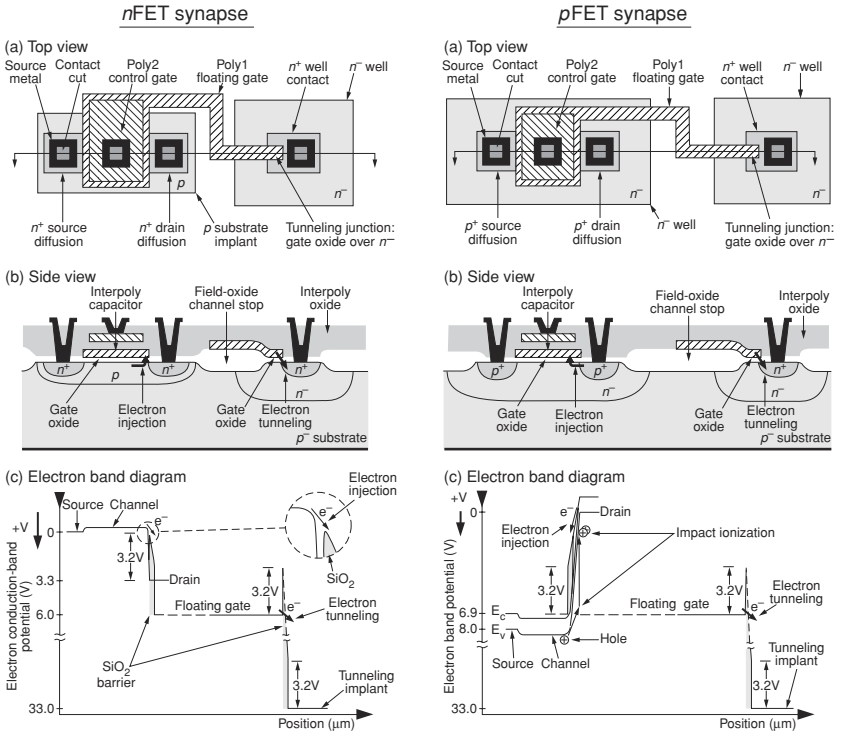


Figure 4.4
*n*FET and *p*FET synapse transistors, showing the electron tunneling and injection locations. The diagrams are aligned vertically; (a) and (c) are drawn to scale; the vertical scale in (b) is exaggerated and all voltages in the conduction-band diagram are referenced to the source potential. The transistors operate in subthreshold ($I_s < 100$ nA). The oxide band diagrams, and the trajectory of (scattered) injection electrons, both project vertically. To better illustrate the injection process, we overlook the scattering and rotate the oxide band diagrams by 90^o, drawing them in the channel direction (horizontally). The tunneling process in the *p*FET synapse is identical to that in the *n*FET. The injection process is different in the two devices, as we describe in the text.

of the weight updates depend on the transistor's terminal voltages and source current. Consequently, a synapse's future weight W varies with the terminal voltages, which are imposed on the device; and with the source current, which is the output. As a result, synapse transistors learn: Their future weight value depends on both the applied input and the present weight value.

Synapse transistors retain all the attributes of conventional transistors, and in addition, they have long-term nonvolatile analog memory; bidirectional memory updates; they compute the product of their stored memory and the

applied input; and they learn from an input signal without interrupting the ongoing computation. Because synapse transistors permit both local computation and local weight updates, they can be used to build autonomous learning arrays in which both the system outputs, and the memory updates, are computed locally and in parallel.

The nFET Synapse

The top and side views of the *n*FET synapse are shown in Fig. 4.4. Tunneling increases the synapse's weight W. A high voltage applied to the n^+ well contact causes electrons to tunnel off the floating gate. The n^+ is surrounded by a lightly doped n^- well to prevent reverse-bias *pn*-junction breakdown (from n^+ to substrate) during tunneling. The breakdown voltage of n^+ to substrate is typically about one-quarter that of the breakdown of n^- to substrate. For example, in a typical 0.35 μm process, n^+ breaks down at about 6V, whereas n^- breaks down at about 25V.

The weight W is decreased by injecting channel electrons onto the floating gate. Hot-electron injection is well known in conventional MOSFETs (Sanchez and DeMassa, 1991). It occurs in short-channel devices with continuous channel currents, when a high gate voltage is combined with a large potential drop across the short channel. It also occurs in switching transistors, when both the drain and gate voltages are transiently high. In neither case is the injection suitable for use in an analog learning system: The short-channel injection requires large channel currents, consuming too much power; and the switching-induced injection is poorly controlled, and transient. Instead, nFET synapse transistors use the drain-to-channel electric field in a subthreshold MOSFET to accelerate channel electrons to high energies, and a fraction of these electrons are injected into the oxide conduction band. The process is shown in the energy-band diagram of Fig. 4.4.

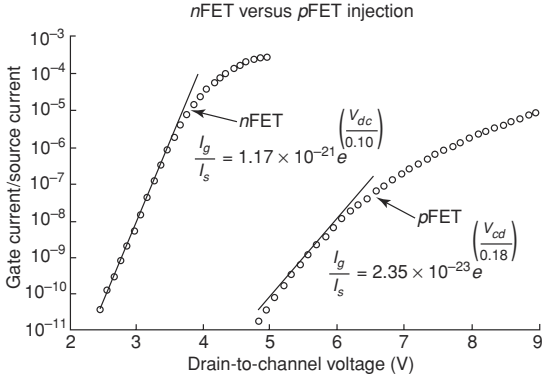
Channel electrons, accelerated in the *n*FET's drain-to-channel depletion region, lose energy by colliding with the semiconductor lattice. A fraction of these electrons scatter upwards toward the gate oxide; and a fraction of these possess sufficient energy to overcome the 3.1 eV difference in electron affinities between the Si and SiO_2 conduction bands and enter the SiO_2. These electrons are then swept over to the floating gate by the oxide electric field. For electrons to be collected at the floating gate, the following two conditions must be satisfied: The electrons must possess the 3.1 eV required to overcome the difference in electron affinities; and the oxide electric field must be oriented in the direction required to transport the injected electrons to the floating gate.

In a conventional *n*-type MOSFET we can easily satisfy the first requirement. We merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about 3 V. Because the subthreshold channel-conduction band is flat, the drain-to-channel transition is steep, meaning that the drain-to-channel electric field is large. Channel electrons accelerate rapidly in this field, and a fraction of them acquire 3.1 eV of energy. A fraction of these 3.1 eV electrons scatter upward into the gate oxide. It is largely the oxide field orientation requirement that prevents a gate current in a conventional subthreshold *n*FET. Subthreshold operation usually implies gate-to-source voltages less than 0.8 V (assuming a 2 μm process). With the drain at 3 V, and the gate at 0.8 V, the drain-to-gate electric field opposes transport of the injected electrons to the floating gate and so the electrons return to the drain instead.

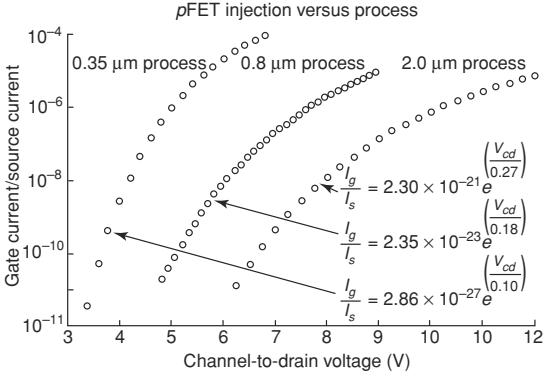
In the *n*FET synapse transistor, the transport of injected electrons to the floating gate is promoted by adding a bulk *p*-type implant ($\sim 10^{17}/\text{cm}^3$) to the channel region. Here, an NPN bipolar-transistor base implant is used. This implant serves two purposes: Firstly, it increases the drain-to-channel electric field, thereby increasing the hot-electron population in the drain-to-channel depletion region; and secondly, it raises the transistor's threshold voltage V_T from 0.8 V to 6 V, thereby allowing the MOSFET to operate with both high floating-gate voltages and subthreshold source currents. Consequently, for typical floating-gate and drain voltages of about 5.5 V and 3 V respectively, the channel current is still subthreshold, but the drain-to-gate electric field transports injected electrons over to the floating gate.

From the perspective of the control gate, raising the MOSFET's threshold voltage is inconsequential, because the control and floating gates are isolated capacitively. Therefore, we can use control-gate inputs in the conventional 0 V to 5 V or 0 V to 3.3 V ranges, regardless of the requirement that the floating gate be near 6 V.

Figure 4.5(a) shows injection efficiency (gate current I_g divided by source current I_s) for a 2 μm *n*FET synapse. The data are plotted as efficiency, because the gate current increases linearly with source current over the entire subthreshold range (see Fig. 4.6(a)). The drain is referenced to the channel potential because the hot-electron population derives from the drain-to-channel electric field. When V_{dc} is less than 2 V, the hot-electron gate current is exceedingly small, and the weight *W* remains nonvolatile. When V_{dc} exceeds 2.5 V, the hot-electron gate current causes measurable changes in the synapse weight *W*. Neuromorphic systems typically use slow adaptation (small oxide currents): V_{dc} for the *n*FET synapse is typically less than 3 V, and is always



(a)



(b)

Figure 4.5
(a) Hot-electron injection efficiency (gate current divided by source current) versus drain-to-channel voltage, for both *n*FET (2 μm process) and *p*FET (0.8 μm process) synapses. The drain voltage is referenced to the channel potential because the hot-electron injection probability varies with the drain-to-channel electric field. The drain voltage can be re-referenced to the source voltage using the relationship between source and channel potential in a subthreshold MOSFET (Enz et al., 1995; Andreou and Boahen, 1994). For the purposes of deriving a synapse weight-update rule, we fit the injection data empirically, using the simple exponential as shown. (b) Hot-electron injection efficiency for *p*FET synapses fabricated in 2.0 μm, 0.8 μm, and 0.35 μm processes. The injection probability increases with decreasing process linewidth, due to higher drain-to-channel electric fields (due to increased implant-impurity concentrations) and thinner gate oxides.

less than 3.5 V. Consequently, we can approximate the data of Fig. 4.5 using a simple exponential:

$$\boxed{I_g = \beta I_s e^{\frac{V_{dc}}{V_{inj}}}} \tag{4.2.3}$$

where I_g is the gate current, I_s is the source current, V_{dc} is the drain-to-channel potential; and β, V_{inj} are fit constants. Because of the *n*FET synapse's 6 V threshold, the floating-gate voltage almost always exceeds 5 V. If V_{dc} <3.5 V, then the drain-to-gate oxide electric field strongly favors the transport of injected electrons to the floating gate. Consequently, we can safely omit gate-voltage dependencies from Eq. 4.2.3.

The injection process that we have described is known in the literature as channel hot-electron injection (CHEI). For more details of the injection physics, see Hasler et al. (1998).

The pFET Synapse

The top and side views of a *p*FET synapse are shown in Fig. 4.4. The tunneling implant is identical to that of the *n*FET synapse, and as in the *n*FET, the electrons are removed from the floating gate when high voltages are applied to the n^+ well contact. However, because the *p*FET and *n*FET synapses are complementary, tunneling has the opposite effect on a *p*FET synapse: It decreases, rather than increases *W*.

The charge carriers in a *p*FET are holes, which cannot be injected onto a floating gate. Furthermore electrons, not holes, must be added to the floating gate to increase the *p*FET synapse's weight. So, to increase W, electrons are first generated and then injected. Electrons are generated by a process called *impact ionization*: Channel holes, accelerated in the transistor's channel-to-drain depletion region, lose energy by colliding with the semiconductor lattice. If the channel-to-drain electric field is large, as is the case for a subthreshold *p*FET with large source-to-drain voltage, then a fraction of these holes collide with sufficient energy to generate free electron-hole pairs. The silicon physics then naturally provides electron injection: The liberated electrons, promoted to their conduction band by the collision, are expelled rapidly from the drain region by the same channel-to-drain electric field. These electrons can, if scattered upward into the gate oxide, inject onto the floating gate. This *impact-ionized hot-electron injection* (IHEI) process is illustrated in the energy band diagram in Fig. 4.4.

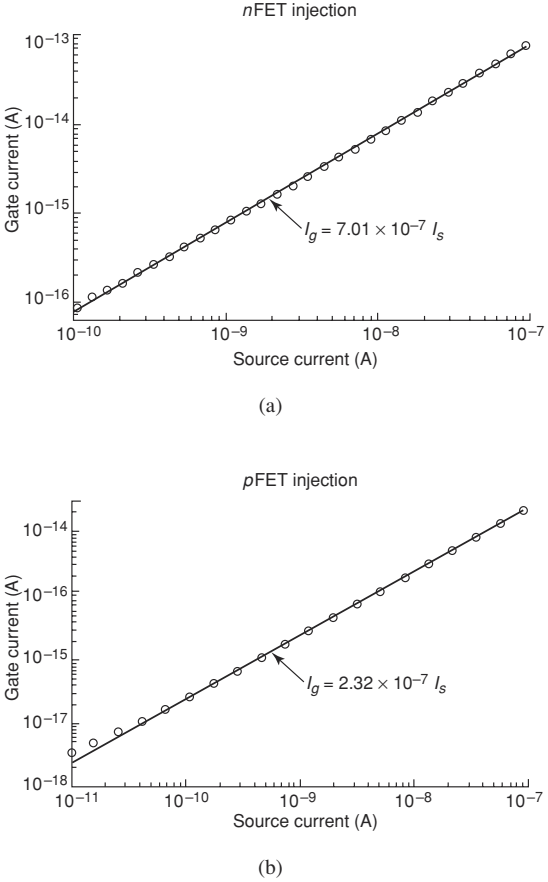


Figure 4.6
(a) Four-terminal *n*FET-synapse gate current I_g versus source current I_s. (b) Four-terminal *p*FET-synapse gate current I_g versus source current I_s. For both devices, the gate current is linearly proportional to the source current over the entire subthreshold range.

To collect electrons at the floating gate of a *p*FET synapse, the same two conditions must be satisfied as for the *n*FET synapse: The electrons must possess the 3.1 eV needed to overcome the difference in electron affinities between the Si and SiO_2; and the oxide electric field must be oriented in the direction required to transport the injected electrons to the floating gate.

The first requirement is satisfied by using a sufficiently large source-to-drain voltage in a subthreshold MOSFET. The silicon physics naturally satisfies the second requirement: In a subthreshold *p*FET, the source-to-gate voltage is typically less than 1 V, whereas the source-to-drain voltage is at least several volts. Consequently, the floating gate is always a few volts higher than the drain, and so the oxide electric field naturally transports the injected electrons to the floating gate. Unlike conventional *n*FET transistors, conventional *p*FET transistors do not need special channel implants to facilitate injection.

As shown in Fig. 4.5, a *p*FET synapse in a 2 μm process achieves an IHEI efficiency of about 10^{-10} when operated with a drain-to-channel voltage of about 6.5 V. An *n*FET synapse in a 2 μm process achieves the same efficiency at a drain-to-channel voltage of about 2.5 V. The *p*FET synapse's higher drain-voltage requirement arises for three reasons: Firstly, for any given drain-to-channel voltage, the *n*FET experiences a higher drain-to-channel electric field than does the *p*FET. This higher field is a consequence of the bulk implant that we add to the *n*FET synapse's channel region. Secondly, phonon loss mechanisms in silicon are naturally more efficient for holes (the *p*FET charge carriers) than they are for electrons (the *n*FET charge carriers). Consequently, a *p*FET requires a larger voltage across its depletion region to achieve the same hot-carrier population. And thirdly, a *p*FET synapse's two-step injection process requires more energy, because of the loss associated with the impact ionization.

A *p*FET synapse's gate current increases linearly with its source current (see Fig. 4.6) because the gate current derives from the impact-ionized electron population, and this population increases linearly with the source current (assuming subthreshold operation). Because the *p*FET synapse's floating-gate voltage is high, the dependency of the gate current on the gate-to-channel potential is small. Consequently, we ignore gate-to-channel dependencies in our (approximate) fit. We use the same fit for the *p*FET synapse as we did for the *n*FET:

$$\boxed{I_g = \beta I_s e^{\frac{V_{dc}}{V_{inj}}}.} \tag{4.2.4}$$

The IHEI efficiency is plotted in Fig. 4.5(b) for *p*-channel synapse transistors fabricated in 2 μm, 0.8 μm, and 0.35 μm CMOS processes. These data show clearly that the results for the 2 μm process scale directly to more modern processes[3].

3 Three-terminal silicon synapses have also been fabricated (Diorio, 2000). The *n*FET 3-terminal

The Gate-Current Equation

Because the tunneling and hot-electron gate currents flow in opposite directions, the final gate-current equation for both the *n*FET and *p*FET synapses is obtained by subtracting Eq. 4.2.3 from Eq. 4.1.1:

$$I_g = I_{to} e^{-\frac{V_f}{V_{ox}}} - \beta I_s e^{\frac{V_{dc}}{V_{inj}}}. \tag{4.2.5}$$

This equation describes the gate current for both types of synapse, over the entire drain-voltage and subthreshold channel-current ranges.

The gate current I_g changes a synapse's W (bidirectionally) by modifying the floating-gate charge Q_{fg}. We use Eq. 4.2.5 for both the *n*FET and *p*FET synapses, although the sign of the weight updates is different in the two cases. In the *n*FET, tunneling increases the weight W, whereas injection decreases it. In the *p*FET, tunneling decreases the weight W, whereas injection increases it.

4.3 Silicon Learning Arrays

A synaptic array can form the basis for a silicon learning system, such as the generic learning array shown in Fig. 4.7. Diorio et al. (1997b) have implemented a 4 by 4 silicon array with this general structure, using a single *n*FET synapse for each "synapse" in the array. Unlike traditional neural-network simulations that use continuous valued inputs, we use pulse inputs. The array computes the inner product of the pulsed input vector and the stored analog weight matrix. The synaptic weights are nonvolatile. Column input pulses that are coincident with row learn-enable pulses cause weight increases at synapses that they address. Unbounded weight values are bounded by a constraint: The time-averaged sum of the synaptic weights in each row of the array, is held constant. This constraint forces row synapses to compete for floating-gate charge and so stabilizes the learning.

Before considering the learning array, we derive a weight-update rule for an *n*FET synapse. The weight-update rule for a *p*FET synapse is identical in form to that for the *n*FET, except for a sign inversion due to the opposite effects that tunneling and injection have on a *p*FET synapse.

---

device incorporates the tunneling function into the transistor's drain. The *p*FET 3-terminal device incorporates the tunneling function into the well contact. Because 3-terminal synapses are not fundamentally different from the 4-terminal synapses described here, and because they are more difficult to use in practice, we will not consider them further.

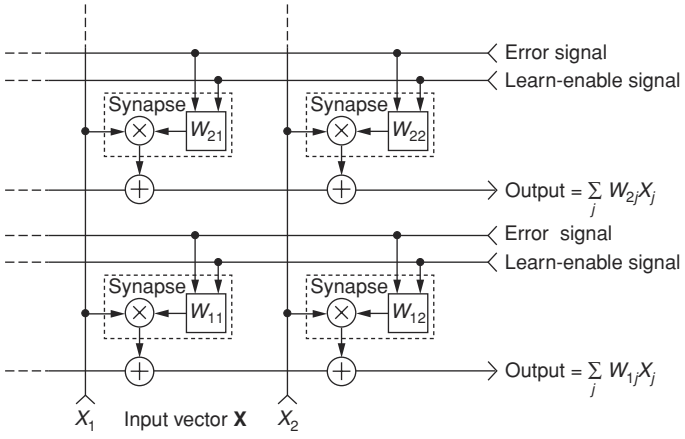


Figure 4.7
A learning-array block diagram. Each synapse multiplies its column input with its nonvolatile analog weight, and outputs a current to the row-output wire, which sums the synapse-output currents along that row. Column inputs that are coincident with the row learn-enable signals cause weight increases at selected synapses. The error signal constrains the time-averaged sum of the row-synapse weights to be a constant, bounding the row weights by forcing the synapses to compete for weight value.

Synaptic Weight Updates

Weight updates depend on the tunneling and injection oxide currents that alter the floating-gate charge. Figure 4.8 shows the temporal derivative of the source current versus the source current, for an *n*FET synapse with a set of fixed tunneling voltages (Fig. 4.8(a)), and a set of fixed drain voltages (Fig. 4.8(b)). In these experiments, the control-gate input V_{in} was held fixed. Consequently, these data show the synaptic weight updates $\partial W/\partial t$, as can be seen by differentiating Eq. 4.2.2.

In Appendix 4.4 we show that the tunneling-induced weight increments follow a power law:

$$\boxed{\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)}} \tag{4.3.1}$$

where σ and τ_{tun} are defined in Eq. 4.4.5 and Eq. 4.4.6, respectively. In Appendix 4.4, we show that the CHEI-induced weight decrements also follow

a power law:

$$\boxed{\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)}} \tag{4.3.2}$$

where ε and τ_{inj} are defined in Eq. 4.4.15 and Eq. 4.4.16, respectively.

SiO_2 trapping is a well-known issue in floating-gate transistor reliability (Aritome et al., 1993). In the synapse, oxide trapping decreases the weight-update rates. Fortunately, our synapses require only small quantities of charge for their weight updates and we can usually ignore oxide trapping.

The Learning Array

Figure 4.9 shows one row of the learning array which comprises a synapse transistor at each node, and a normalization circuit at the row boundary. The column inputs X_i and the row learn-enable signals Y_j are digital pulses. Each synapse multiplies its binary-valued input X_i with its stored weight W_{ij}, and outputs a source current I_{sij} whose magnitude is given by Eq. 4.2.2. The total row current I_{out} is the sum of the source currents from all the synapses in the row. Synapses are ordinarily on; low-true gate inputs X_i turn off selected synapses, decreasing the current I_{out} transiently. This decrease in I_{out} in response to an input vector **X**, is the row computation.

The row learn-enable inputs Y_j are high-voltage inputs that can increase the synapse weights when $V_{ox} \equiv V_{tun} - V_{fg}$ (where $V_{tun} = Y_j$ +25 and V_{fg} is the floating-gate voltage) is large enough to induce tunneling. Synapse-weight increases occur only when both the row and column inputs, Y_j and X_i, are true. To see why, first consider the case when the row learn-enable signal Y_j is false (V_{tun} is low). Because $V_{ox} = V_{tun} - V_{fg}$, V_{ox} is small for every synapse in the row when V_{tun} is low. In this case, the tunneling currents are small, and there is no weight increase at any row synapse.

Now consider the case when Y_j is true (V_{tun} is high). V_{ox} increases as V_{fg} decreases, and V_{fg} follows X_i. Therefore, if a low-true column input X_i is true, then V_{fg} is low; V_{ox} is large; and electron tunneling causes a weight increase at the selected synapse. If, on the other hand, the low-true column input X_i is false, then V_{fg} is high; V_{ox} is too small to cause appreciable tunneling; and there is little change in the synaptic weight.

Tunneling increases the weight value of a row-column selected synapse. Because this weight update is single quadrant (tunneling only increases a synapse's weight), tunneling allows unbounded weight values. To constrain the array weights, we renormalize each row of the array. The array allows

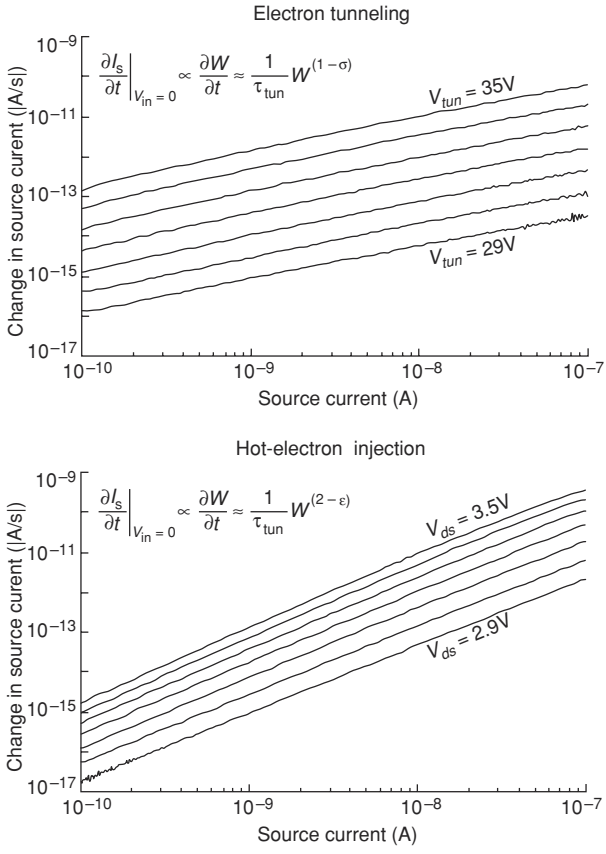


Figure 4.8
*n*FET synapse transistor (a) tunneling and (b) CHEI weight updates. We measured the synapse's source current I_s versus time, and plotted — $\partial I_s/\partial t$ — versus I_s. We fixed the synapse's terminal voltages; consequently, the change in I_s is a result of changes in the synapse's weight W. In (a), we applied V_{in}=5 V, V_s=0 V, V_{ds}=2 V, and stepped V_{tun} from 29 V to 35 V in 1 V increments; in (b), we applied V_{in}=5 V, V_s=0 V, V_{tun}=20 V, and stepped V_{ds} from 2.9 V to 3.5 V in 0.1 V increments. We turned off the tunneling and CHEI at regular intervals, to measure I_s. Because, for a fixed V_{in}, the synapse's weight updates $\partial W/\partial t$ are proportional to $\partial I_s/\partial t$ (see Eq. 4.2.2), these data show that the weight updates follow a power law. The mean values of σ and ε are 0.17 and 0.24, respectively.

unsupervised learning (Hertz et al., 1991) under the constraint that the sum of the row-synapse weights, averaged over time, is constant. CHEI feedback along each row enforces the constraint.

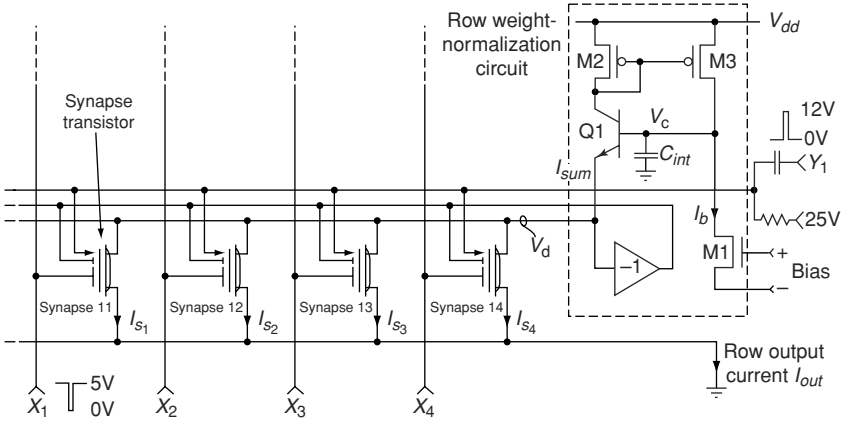


Figure 4.9
One row of the learning array. The column input vector **X** comprises low-true, 5 V, 10 μs digital pulses. The row input vector **Y** comprises high-true, 12 V, 10 μs digital pulses. Because the 2 μm CMOS process that we use has 400 Å gate oxides, the tunneling voltages are high: To cause measurable tunneling, we superimpose the row inputs onto a 25 V DC bias. The voltage coupling between a synapse's control and floating gates is about 0.8. Consequently, a 5 V (low-true) input on column wire X_1 causes a 4 V decrease in synapse11's floating-gate voltage, which in turn causes a 4 V increase in synapse11's tunneling-oxide voltage. A column input X_1 that is coincident with a row learn-enable pulse Y_1 causes a 16 V increase in the tunneling-oxide voltage at synapse11, but only a 12 V increase at the other synapses. Because electron tunneling increases exponentially with tunneling-oxide voltage (see Fig. 4.3), synapse11's floating gate receives about 100 times more charge than do the other synapses' floating gates. Because W increases exponentially with floating-gate charge (see Eq. 4.2.1), synapse11's weight increases much more than do the other synapses' weights. The weight increase causes I_{sum} to rise, which in turn causes the normalization circuit to raise V_d. Because the CHEI efficiency increases with V_{ds} (see Fig. 4.5), a higher V_d causes CHEI in all the synapses, decreasing all the weights. The array eventually settles back to equilibrium, with I_{sum} equal to I_b, but synapse11 now takes a larger share of the total row current, and the other synapses each take a smaller share. The inverting amplifier in the weight-normalization circuit enhances loop stability, for reasons that we discuss in Section 4.3. Each row of the array has its own normalization circuit.

Weight Normalization

The weight-normalization circuit (see Fig. 4.9) compares I_{sum}, the sum of the synapse drain currents in a row, with I_b, the bias current in transistor M_1. If $I_{sum} > I_b$, then the circuit uses CHEI to renormalize the weights. To understand the renormalization, we begin by defining equilibrium: A row is in equilibrium when $I_{sum} = I_b$. In equilibrium, the drain voltage V_d causes little or no CHEI in the row synapses.

The normalization circuit constrains I_{sum} as follows: Assume that the row is initially in equilibrium, and that tunneling then raises the weight values

of selected synapses, increasing I_{sum}. The excess drain current (I_{sum}-I_b) is mirrored by M_2 and M_3 into capacitor C_{int}, causing V_c to rise; Q1 forces V_d to follow V_c. When V_d rises, all the row synapses undergo CHEI, decreasing all the weights and so causing I_{sum} to fall. As I_{sum} falls, V_d also falls, and the row returns to equilibrium. The drain-current constraint requires that, over time, $I_{sum} = I_b$. The normalization circuit creates a negative resistance at the synapses' common drain node, causing V_d to rise when I_{sum} increases.

We now show how the drain-current constraint renormalizes the synapse weights. We begin with the constraint

$$\sum_i I_{s_i} \approx \sum_i I_{d_i} \equiv I_{sum} = I_b. \tag{4.3.3}$$

The renormalization time constant τ_a typically exceeds 10s; this value is 10^6 times longer than the 10 μs input pulses X_i (where $V_{in} = X_i$). Consequently, for renormalization, we replace V_{in} in Eq. 4.2.2 with its temporal average $\overline{V_{in}}$, and we assume that $\overline{V_{in}}$ is time invariant, and has the same value for all the row synapses. Substituting Eq. 4.2.2 into Eq. 4.3.3, we obtain

$$\sum_i W_i \, I_o e^{\frac{\kappa' \overline{V_{in}}}{U_t}} = I_o e^{\frac{\kappa' \overline{V_{in}}}{U_T}} \sum_i W_i \; = I_b \tag{4.3.4}$$

which implies that

$$\sum_i W_i \; = \; \frac{I_b}{I_o} \, e^{\frac{-\kappa' \overline{V_{in}}}{U_T}} \equiv W_{sum} = constant. \tag{4.3.5}$$

The drain-current and weight-value constraints are equivalent. Consequently, row feedback renormalizes the synapse weights.

Renormalization forces the row synapses to compete for floating-gate charge: When one synapse's weight value increases, the sum of the weight values of its row neighbors must decrease by the same amount. However, when a selected synapse tunnels, thereby increasing its weight, renormalization forces *all* the row synapses to undergo CHEI, decreasing *all* the row-synapse weights. The selected synapse undergoes both tunneling and CHEI. Because the exponent in the CHEI weight-update rule is larger than that in the tunneling rule (see Eq. 4.3.1 and Eq. 4.3.2), renormalization constrains a synapse's weight-update rate, in addition to its weight value.

Tunneling and CHEI effectively redistribute a fixed quantity of floating-gate charge among the row synapse transistors. In Appendix 4.4 we derive the

array learning rule, for coincident (*x*,*y*) pulse inputs to synapse *j*:

$$W_{i,\ i\neq j}\,(n+1) = W_{i,\ i\neq j}\,(n) - f_{learn}\,W_i\,(n)^{(2-\varepsilon)} \tag{4.3.6}$$

$$W_j\,(n+1) = W_j\,(n) + f_{learn}\sum_{i,\ i\neq j} W_i\,(n)^{(2-\varepsilon)} \tag{4.3.7}$$

where ε and f_{learn} are defined in Eq. 4.4.15 and Eq. 4.4.25 respectively. Figure 4.10 shows unsupervised learning in one row of the 4 by 4 array. These data highlight both the synapse weight and the update-rate constraints. The data are fit by applying Eq. 4.3.6 and Eq. 4.3.7 recursively; the only inputs to the fit equations are the synapse weights at *n*=0 and the fit constants τ_{tun}, t_{pw}, σ, and ε.

The parasitic coupling between a synapse's tunneling junction and its floating gate is about 5 fF. With C_T=1 pF, a 12 V row learn-enable pulse Yj increases the floating-gate voltage of every row synapse by about 60 mV. This coupling does not affect the row computation significantly, for two reasons: Firstly, 5 V low-true column inputs X_i always turn off selected synapses, regardless of Y_j. Secondly, because row learn-enable pulses Y_j increase the floating-gate voltage of every de-selected synapse by a fixed 60 mV, we can calculate the corresponding source-current increase using Eq. 4.2.1, and adjust I_{out} accordingly.

Normalization-Circuit Stability

The normalization circuit creates a negative resistance at the synapses' common drain node, V_d. The loop output is V_d and the loop feedback comprises CHEI oxide currents. When I_{sum} increases, V_d rises and CHEI decreases the synapse weights, causing I_{sum} to fall. Because the CHEI oxide currents increase exponentially with V_d, the loop dynamics are highly nonlinear. This nonlinearity is inherent to synapse transistors, and is likely be present in any circuit that employs oxide currents in the feedback path. A quantitative stability analysis of this circuit is beyond the scope of this discussion. However, rather than omit any discussion of stability, we describe qualitative loop-stability criteria.

The normalization circuit employs positive feedback. To ensure stability, we must make the loop gain less than unity for all frequencies. This requirement implies that the small-signal impedance z_d looking into the synapse drain terminals, must be greater than the total impedance z_c at capacitor C_{int}. To see why, we assume instead that $z_c > z_d$. A rising V_d induces a small-signal current

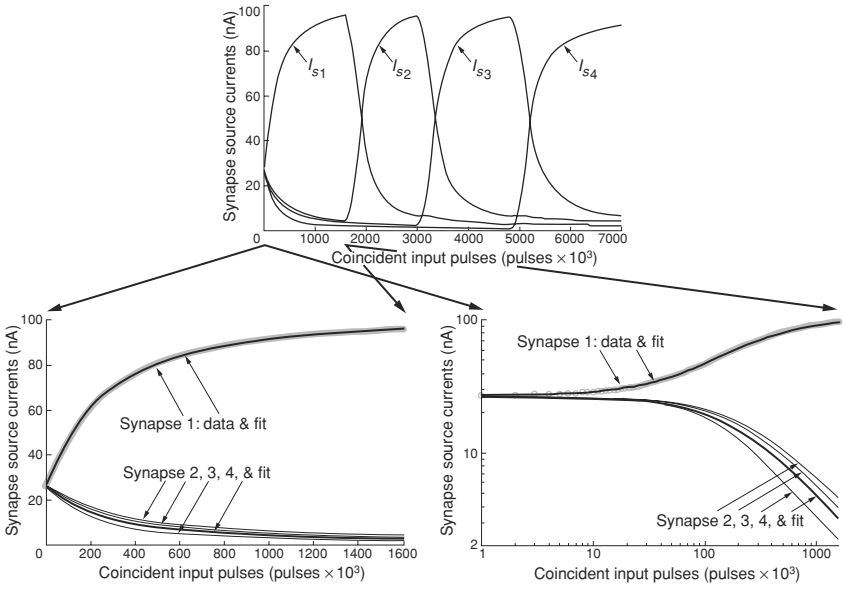


Figure 4.10
Array learning behavior, with fits. We first initialized all synapses to the same source-current value. We then applied a train of coincident (*x*,*y*) 10 μs pulses to synapse11, causing its weight value and source current to increase. Renormalization caused the weight values and source currents of the other synapses to decrease. Once synapse11 had acquired 90% of the total row current, the pulse-train stimulus was removed and applied instead to synapse12, and then in turn to synapses 13 and 14. The synapse source currents were measured after every 10^3 input pulses. In the lower half of the figure, the first 1600 data points are fit by applying Eq. 4.3.6 and Eq. 4.3.7 recursively. The linear plot shows the fit accuracy over the entire sweep, whereas the logarithmic plot shows that the weight values of deselected synapses do not saturate, but instead follow a power-law decay as predicted by Eq. 4.3.2 and Eq. 4.3.6. The inputs to the fit equations are the initial synapse source-current values (at n=0); the pulsewidth t_{pw}=10 μs; and the empirical constants τ_{tun}=10 ms, σ=0.14, and ε=0.21. These data show that individual synapses can be addressed with good selectivity, and that wide separation in the weight values of selected versus deselected synapses can be achieved.

$i_{sum} = v_d/z_d$; i_{sum} is mirrored by M_2 and M_3 into C_{int}, causing V_c to rise by an amount v_c=$i_{sum}z_c$. Because V_d follows V_c, if $z_c > z_d$, then $v_c > v_d$; i_{sum} will increase rapidly, causing V_c to rise toward V_{dd}.

The impedance z_d is limited by interconnect capacitances, and by synapse-transistor channel-length modulation, floating-gate-to-drain overlap capacitance, and drain-current impact ionization. We consider each of these limitations in turn.

Interconnect Capacitance Interconnect capacitance at the synapses' common drain node causes z_d to decrease with frequency. We choose C_{int} to be much larger than this parasitic capacitance, so the reactive impedance ratio, z_c/z_d, favors loop stability for all frequencies.

Channel-Length Modulation Channel-length modulation reduces a synapse's drain impedance, limiting z_d. Fortunately, the synapse transistor's Early voltage exceeds 100 V, as a result of both the 10 μm channel length and the *p*-type channel implant; consequently, the channel-length modulation is small.

Floating-Gate-to-Drain Overlap Capacitance V_d couples to a synapse transistor's floating gate, by means of the floating-gate-to-drain overlap capacitance C_{dg}. The coupling coefficient is C_{dg}/C_T, where C_T is the total floating-gate capacitance. Because I_s increases exponentially with V_{fg}, C_{dg} causes I_{sum} to increase exponentially with V_d, limiting z_d. To minimize the effect, we use a large interpoly capacitor (C_T=1 pF); we also apply inverting feedback from V_d to the floating gate, increasing z_d (see Fig. 4.9). We used an off-chip amplifier to generate this inverting feedback; an on-chip adaptive floating-gate amplifier (Hasler, 1997) can be used instead.

Drain-Current Impact Ionization Channel electrons that possess sufficient energy for CHEI also possess sufficient energy for impact ionization (Shockley, 1961; Tam et al., 1984). In the synapse transistor, a drain-to-channel electric field that causes CHEI also creates additional electron-hole pairs, causing I_d to increase exponentially with V_{dc}. As a result, I_{sum} increases exponentially with V_d, limiting z_d. If V_d becomes greater than about 4 V, the rate of drain-current increase causes loop instability, and V_d rises rapidly. As V_d rises, CHEI decreases all the synapse-transistor weights. As V_d saturates near V_{dd}, CHEI causes I_{sum} to fall below I_b; V_d falls and the loop returns to a stable operating regime. Loop instability causes V_d to undergo a single brief ($\sim$10 μs) voltage spike, and reduces all the synapse weights substantially. Fortunately, because the synapse CHEI efficiency is high, weight renormalization rarely causes V_d to exceed 3.5V; consequently, the loop is stable.

Local Learning in Silicon

The synaptic array that we have analyzed, although simplistic from a computational point of view, demonstrates the potential for performing large-scale

computation and unsupervised learning on a silicon chip. The computation and synapse-weight modification occur locally and in parallel: The array performs fast, single-transistor analog computation and slow, locally computed weight adaptation. In addition, we can describe the computation and learning behavior using rules derived directly from the silicon-MOS and silicon-oxide physics.

Although the present array affords unsupervised learning, it uses a feedback error signal to constrain the weight values. Feedback error signals typically are used in supervised neural networks, to adjust the array weights according to a network learning rule. In future floating-gate arrays, rather than using unsupervised learning, we can use CHEI to adjust the synapse weights in a supervised fashion, using either pulsed, or continuously valued analog (Hasler, 1997), inputs and row-error signals. Further examples of learning circuits can be found in Cauwenberghs and Bayoumi (1999).

4.4 Appendices

We derive the tunneling weight-increment rule for an *n*FET synapse in Appendix 4.4; the CHEI weight-decrement rule for an *n*FET synapse in Appendix 4.4; and the *n*FET array learning rule in Appendix 4.4.

The Tunneling Weight-Increment Rule

We begin by taking the temporal derivative of the synapse weight W, where $W \equiv \exp(Q_{fg}/Q_T)$:

$$\frac{\partial W}{\partial t} = \frac{W}{Q_T}\,\frac{\partial Q_{fg}}{\partial t} = \frac{W}{Q_T} I_g \tag{4.4.1}$$

and substitute Eq. 4.1.1 for the tunneling gate current I_g:

$$\frac{\partial W}{\partial t} = \frac{I_{to}}{Q_T}\, W e^{-\frac{V_f}{V_{ox}}}. \tag{4.4.2}$$

We substitute $V_{ox} = V_{tun} - V_{fg}$ (where V_{tun} and V_{fg} are the tunneling-implant and floating-gate voltages, respectively), assume that $V_{tun} \gg V_{fg}$, expand the exponent using $(1\text{-}x)^{-1} \approx 1+x$, and solve:

$$\frac{\partial W}{\partial t} \approx \frac{I_{to}}{Q_T}\, W e^{-\frac{V_f}{V_{tun}} - \frac{V_f V_{fg}}{V_{tun}^2}} \tag{4.4.3}$$

Substituting $V_{fg} = U_T\, Q_{fg}/\kappa Q_T$, and solving for the tunneling weight-

increment rule yields

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} \tag{4.4.4}$$

where

$$\sigma \equiv \frac{V_f U_T}{\kappa V_{tun}^2} \tag{4.4.5}$$

and

$$\tau_{tun} \equiv \frac{Q_T}{I_{to}} e^{\frac{V_f}{V_{tun}}}. \tag{4.4.6}$$

The parameters σ and τ_{tun} vary with the tunneling voltage V_{tun}.

The CHEI Weight-Decrement Rule

We begin by defining a synapse transistor's drain-to-channel potential, V_{dc}, in terms of V_{ds} and I_s. In a subthreshold floating-gate MOSFET, the source current is related to the floating-gate and source voltages (Mead, 1989) by

$$I_s = I_o e^{\frac{\kappa V_{fg} - V_s}{U_T}} \tag{4.4.7}$$

and the channel-surface potential Ψ is related to the floating-gate voltage, V_{fg} (Andreou and Boahen, 1994; Enz et al., 1995) by

$$\Psi \approx \kappa V_{fg} + \Psi_o \tag{4.4.8}$$

where κ is the coupling coefficient from the floating gate to the channel, and Ψ_o depends on the MOS process parameters.

Using Eq. 4.4.7 and Eq. 4.4.8, we solve for the surface potential Ψ in terms of I_s and V_s:

$$\Psi = V_s + \Psi_o + U_T \, ln \left(\frac{I_s}{I_o} \right). \tag{4.4.9}$$

We now solve for V_{dc}:

$$V_{dc} = V_d - \Psi = V_{ds} - \Psi_o - U_T \, ln \left(\frac{I_s}{I_o} \right). \tag{4.4.10}$$

The CHEI gate current I_g is given by Eq. 4.2.3. We add a minus sign to I_g, because CHEI decreases the floating-gate charge; and substitute for V_{dc} using

Eq. 4.4.10:

$$I_g = -\beta I_s e^{\frac{V_{ds}-\Psi_o - U_T \ln(I_s/I_o)}{V_{inj}}} = -\beta I_o^{\frac{U_T}{V_{inj}}} e^{\frac{V_{ds}-\Psi_o}{V_{inj}}} I_s^{\left(1-\frac{U_T}{V_{inj}}\right)}. \tag{4.4.11}$$

Substitute for I_s using Eq. 4.2.2, and solve:

$$I_g = -\beta I_o e^{\frac{\kappa' V_{in}}{U_T} + \frac{V_{ds}-\kappa' V_{in}-\Psi_o}{V_{inj}}} W^{\left(1-\frac{U_T}{V_{inj}}\right)}. \tag{4.4.12}$$

Substitute Eq. 4.4.12 into $\partial W/\partial t$ (Eq. 4.4.1):

$$\frac{\partial W}{\partial t} = -\frac{\beta I_o}{Q_T} e^{\frac{\kappa' V_{in}}{U_T} + \frac{V_{ds}-\kappa' V_{in}-\Psi_o}{V_{inj}}} W^{\left(2-\frac{U_T}{V_{inj}}\right)} \tag{4.4.13}$$

to obtain the final weight-decrement rule:

$$\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \tag{4.4.14}$$

where

$$\varepsilon \equiv \frac{U_T}{V_{inj}} \tag{4.4.15}$$

and

$$\tau_{inj} \equiv \frac{Q_T}{\beta I_o} e^{-\frac{\kappa' V_{in}}{U_T} - \frac{V_{ds}-\kappa' V_{in}-\Psi_o}{V_{inj}}}. \tag{4.4.16}$$

The parameter ε is a constant, whereas τ_{inj} varies with V_{in}.

The Array Learning Rule

We consider the row-synapse weights at discrete time intervals $t \equiv n\mathrm{T}$, where *n* is the step number and T is the timestep and derive the row-learning rule for a single coincident (*x,y*) input to a single row synapse. We begin with the equilibrium condition for the row-weight normalization:

$$\sum_i W_i(n) = W_{sum} \tag{4.4.17}$$

We assume that the normalization time constant τ_a is fixed, for the following reason: Coincident (*x,y*) input pulses cause a weight increase at a synapse; the normalization circuit responds by establishing a drain voltage V_d for which the total weight decay, summed over all the row synapses, balances the weight increase at the single synapse. If we assume that the mean density of the coincident input pulses is time-invariant, then V_d's mean value, $\overline{V_d}$, is constant,

and therefore the low-frequency loop time constant, τ_a, also is constant. We assume that $\tau_a \ll \mathrm{T}$.

The synapse weight values can violate Eq. 4.4.17 for times $t \ll \tau_a \ll \mathrm{T}$, but we require that they satisfy Eq. 4.4.17 at our measurement time intervals t=nT. We permit array inputs at times (t+δt)≡(n+δ)T, immediately after the synapse weight values at t=nT are measured. The array inputs comprise a pulsed column vector $\mathbf{X}(n+\delta)$, where $X_i \in [0,1] \equiv$ [5V,0V], and a pulsed row vector $\mathbf{Y}(n+\delta)$, where $Y_j \in [0,1] \equiv$ [0V,12V]. Without loss of generality, we assume that at time t=nT, the circuit is in equilibrium, and that at (t+δt+t_{pw})≡(n+δ_{pw})T, coincident row and column inputs of duration t_{pw} have caused synapse j's weight to increase:

$$W_j\left(n+\delta_{pw}\right) \approx W_j\left(n\right) + \frac{\partial W_j\left(n\right)}{\partial t}\, t_{pw} \tag{4.4.18}$$

$$\approx W_j\left(n\right) + \frac{t_{pw}}{\tau_{tun}} W_j\left(n\right)^{(1-\sigma)} \tag{4.4.19}$$

where in Eq. 4.4.18 we have made the first-order approximation that $\partial W/\partial t$ is constant over t_{pw}, and in Eq. 4.4.19 we have substituted for $\partial W/\partial t$ using Eq. 4.3.1. Because $t_{pw} \ll \tau_a$, at time (n+δ_{pw}) the circuit no longer is in equilibrium,

$$\sum_i W_i\left(n+\delta_{pw}\right) > W_{sum} \tag{4.4.20}$$

and the synapse weights inject down to reestablish equilibrium.

We wish to find the synapse weights at (n+1), when the row again satisfies Eq. 4.4.17. Using Eq. 4.4.14 and Eq. 4.4.19, we write weight-decrement expressions for the row synapses

$$\Delta W_{i,i\neq j}\left(n+1\right) = -\frac{T}{\tau_{inj}} W_{i,i\neq j}\left(n\right)^{(2-\varepsilon)} \tag{4.4.21}$$

$$\Delta W_j\left(n+1\right) \approx -\frac{T}{\tau_{inj}} \left(W_j\left(n\right) + \frac{t_{pw}}{\tau_{tun}} W_j\left(n\right)^{(1-\sigma)}\right)^{(2-\varepsilon)} \tag{4.4.22}$$

where, because the row drain voltage V_d settles during renormalization, τ_{inj} may vary over T (recall that $\mathrm{T} \gg \tau_a \gg t_{pw}$). For reasonable values of V_{tun} and t_{pw}, the weight increment from a single coincident (x,y) input is small;

consequently, Eq. 4.4.22 can be simplified using $(1+x)^n \approx 1+nx$,

$$\Delta W_j(n+1) \approx -\frac{T}{\tau_{inj}} W_j(n)^{(2-\varepsilon)} \left(1 + (2-\varepsilon)\frac{t_{pw}}{\tau_{tun}} W_j(n)^{-\sigma}\right). \tag{4.4.23}$$

Because τ_{inj} varies over T, T/τ_{inj} can re-expressed in terms of quantities that we know at *n*. We equate the weight increment at synapse *j* (see Eq. 4.4.19) to the sum of the weight decrements at synapses *i*,*i*≠*j* (Eq. 4.4.21) and *j* (Eq. 4.4.23)

$$\begin{aligned}\frac{t_{pw}}{\tau_{tun}} W_j(n)^{(1-\sigma)} &= \frac{T}{\tau_{inj}} \sum_{i,\, i\neq j} W_i(n)^{(2-\varepsilon)} \\ &+ \frac{T}{\tau_{inj}} W_j(n)^{(2-\varepsilon)} \left(1 + (2-\varepsilon)\frac{t_{pw}}{\tau_{tun}} W_j(n)^{-\sigma}\right)\end{aligned} \tag{4.4.24}$$

and solve for T/τ_{inj}:

$$\frac{T}{\tau_{inj}} = \frac{\frac{t_{pw}}{\tau_{tun}} W_j(n)^{(1-\sigma)}}{(2-\varepsilon)\frac{t_{pw}}{\tau_{tun}} W_j(n)^{(2-\varepsilon-\sigma)} + \sum_i W_i(n)^{(2-\varepsilon)}}. \tag{4.4.25}$$

We define $f_{learn} \equiv$T/τ_{inj}, substitute f_{learn} into Eq. 4.4.21, and use Eq. 4.4.17 to solve for the row-learning rule:

$$W_{i,\, i\neq j}(n+1) = W_{i,\, i\neq j}(n) - f_{learn} W_i(n)^{(2-\varepsilon)} \tag{4.4.26}$$

$$W_j(n+1) = W_j(n) + f_{learn} \sum_{i,\, i\neq j} W_i(n)^{(2-\varepsilon)}. \tag{4.4.27}$$

Eq. 4.3.6 and Eq. 4.3.7 describe the row weight-update rule for a single coincident (*x*,*y*) pulse input to synapse *j*.

II STATICS

This page intentionally left blank

5 Basic Static Circuits

In this chapter we present some basic analog VLSI circuits that are widely used as building blocks of more complex circuits and that are more extensively treated in various textbooks (Gregorian and Temes, 1986; Horowitz and Hill, 1989; Mead, 1989; Johns and Martin, 1997; Maloberti, 2001; Gray et al., 2001; Razavi, 2001; Allen and Holberg, 2002). Other basic circuits that are less commonly used, but particular to the type of structures presented in this book, are introduced in the next chapter and some of the later chapters. Most of the circuits in this chapter are described only in one configuration. An almost equivalent circuit is obtained by exchanging the types of all MOSFETs and by reversing the signs of all voltage differences. We restrict ourselves to a steady-state analysis, which is valid if the largest signal frequency is much smaller than the bandwidth of the circuit. Furthermore, unless otherwise noted, we only consider the subthreshold domain and neglect second-order effects, such as the Early effect. The equations governing the behavior above threshold can be found in standard text books. We only point out qualitative differences between the two domains. Note that the larger currents above threshold increase the bandwidths of the circuits. In order to keep the equations simple, calculations are for MOSFETs with unity width-to-length ratios, but extension to other values is straightforward.

We saw in Chapter 3 that the natural reference potential for a MOSFET is its bulk potential. However, for the analysis of circuits whose MOSFETs do not all have the same bulk potential a common reference potential must be chosen. In traditional CMOS circuits, all nFETs have a common bulk potential, usually called V_{ss}, and all pFETs have a common bulk potential called V_{dd}. In order to avoid large currents from the sources and drains into the bulks, all source and drain diodes to the bulk are reverse-biased: V_{ss} is the lowest and V_{dd} is the highest potential in the circuit. The V_{ss} and V_{dd} lines are therefore called *power rails*. It is convenient to choose either V_{ss} or V_{dd} as the reference potential. In the following, we will reference all voltages to V_{ss} such that all the voltages in the circuit are positive. In our circuits, the bulks of the MOSFETs do not necessarily have to be connected to the power rails, but we use the convention that the bulk connections are omitted in the MOSFET symbols if they are connected to the power rails. Note, however, that in standard CMOS processes, all bulks of one MOSFET type are connected to the lightly-doped silicon substrate and are thus at a common potential, which is one of the power

rail potentials. MOSFETs of the other type rest in wells, whose potentials can be individually chosen.

In the circuit schematics, nodes at larger potentials are drawn above nodes at lower potentials for each branch of the circuit, such that the currents flow downward. Connections to V_{ss} are denoted by one of the common symbols used for ground connections, and connections to V_{dd} are denoted by a slanting line. All other connections to external circuitry are denoted by circles. Some of the input or output nodes of the circuits are not labeled with a voltage parameter. If such a node is the drain of a MOSFET, the voltage value at that node can vary freely. Unlabeled source nodes are assumed to be held at a constant potential.

The steady-state subthreshold characteristics of MOSFETs (neglecting the Early effect) are described by Eq. 3.2.10 and 3.3.1 for nFETs and pFETs respectively. With the above conventions and under the assumption that the bulks are connected to the respective power rails, we obtain for the drain current of the nFET

$$I = I_{n0} e^{\kappa_n V_g / U_T} \left(e^{-V_s / U_T} - e^{-V_d / U_T} \right) \tag{5.0.1}$$

where I_{n0} denotes the nFET current-scaling parameter, κ_n the nFET subthreshold slope factor, U_T the thermal voltage, V_g the gate voltage, V_s the source voltage, and V_d the drain voltage. The current is defined to be positive if it flows from the drain to the source. The corresponding equation for the pFET is

$$I = I_{p0} e^{\kappa_p (V_{dd} - V_g) / U_T} \left(e^{-(V_{dd} - V_s) / U_T} - e^{-(V_{dd} - V_d) / U_T} \right) \tag{5.0.2}$$

where the values of the pFET current-scaling parameter I_{p0} and the subthreshold slope factor κ_p are different from the corresponding nFET values.

5.1 Single-Transistor Circuits

The rich behavior of a MOSFET in its different regimes allow it to perform several functions.

Current Source

One of the simplest functions of a MOSFET is obtained when its source and gate potentials are held at constant values. As long as the difference between the drain and source voltages is larger than approximately $4U_T$, the MOSFET

is in saturation and Eq. 5.0.1 reduces to

$$\boxed{I = I_{n0} e^{(\kappa_n V_g - V_s)/U_T}} \tag{5.1.1}$$

(cf. Eq. 3.2.16). In this first-order approximation, the drain current is independent of the drain voltage. A device that supplies an output current that is independent of the voltage applied to the same terminal is called a *current source*. The transistor deviates systematically from an ideal current source because of the Early effect. This deviation can be reduced by making the length of the transistor large. Above threshold, the saturation current is given by Eq. 3.2.39, which we rewrite here as

$$I = \frac{\beta}{2\kappa_n} \left[(\kappa_n (V_g - V_{T0}) - V_s)^2 \right] . \tag{5.1.2}$$

Above threshold, the drain voltage required for saturation is larger than below threshold and depends on V_g. Furthermore, the Early voltage is smaller and hence, the MOSFET is a less ideal current source than when it is operated below threshold.

Linear Resistor

The drain current of the transistor in the triode regime strongly depends on the relative values of the source and drain voltages. We can rewrite Eq. 5.0.1 for subthreshold operation as

$$I = I_{n0} e^{\kappa_n V_g/U_T - (V_d+V_s)/2U_T} \left(e^{(V_d - V_s)/2U_T} - e^{-(V_d - V_s)/2U_T} \right) \tag{5.1.3}$$

$$= 2 I_{n0} e^{\kappa_n V_g/U_T - (V_d+V_s)/2U_T} \sinh \left(\frac{V_d - V_s}{2U_T} \right) . \tag{5.1.4}$$

If $V_d - V_s$ is small enough to neglect third-order and higher order terms, a Taylor series expansion yields

$$I \approx I_{n0} e^{\kappa_n V_g/U_T - (V_d+V_s)/2U_T} \frac{V_d - V_s}{U_T} . \tag{5.1.5}$$

For a given common mode voltage $(V_d + V_s)/2$ and a given gate voltage V_g, a MOSFET acts as a *linear resistor* with resistance

$$\boxed{R = \frac{U_T}{I_{n0}} e^{(V_d+V_s)/2U_T - \kappa_n V_g/U_T} .} \tag{5.1.6}$$

Above threshold, the current-voltage relationship in the triode regime is described by Eq. 3.2.38. Neglecting second-order terms in V_s and V_d, the resistor

is linear:

$$R = (\beta(V_g - V_T))^{-1}. \tag{5.1.7}$$

As in the subthreshold case, the resistance depends on V_g, and the common-mode voltage through V_T.

Hence, in the triode regime, the transistor is approximately linear both above and below threshold. The resistance is controlled by the gate voltage of the transistor. Below threshold, the resistance is an exponential function of the gate voltage. Above threshold, the resistance is a function of the inverse of the gate voltage. The triode regime of the transistor extends to larger drain-to-source voltages above threshold than below threshold, but the linear approximation holds only for first-order terms, whereas below threshold the approximation is valid up to second-order terms.

Transistors operated in the linear regime are suitable for use as *pseudo-conductances* in resistive networks (see Chapter 6).

Nonlinear Voltage-Current / Current-Voltage Converter

A MOSFET operating in saturation in the subthreshold region, as described by Eq. 5.1.1 for an nFET, has a drain current I that is an exponential function of V_s and of V_g. In this configuration, the MOSFET acts as an exponential voltage-to-current converter if one of the two voltages is fixed and the other node is the input. Above threshold, there is a quadratic voltage-to-current conversion as described by Eq. 3.2.39. The inverse function (a current-to-voltage conversion) is obtained by simply making I the input signal and V_g or V_s the output signal. The MOSFET acts as a logarithmic current-to-voltage converter below threshold, and a square-root current-to-voltage converter above threshold. The current-to-voltage conversion function in subthreshold can be derived by solving for the output voltage in Eq. 5.1.1:

$$V_s = \kappa_n V_g - U_T \log\left(\frac{I}{I_{n0}}\right) \tag{5.1.8}$$

if the source is the output terminal and

$$V_g = \kappa_n^{-1}\left(V_s + U_T \log\left(\frac{I}{I_{n0}}\right)\right) \tag{5.1.9}$$

if the gate is the output terminal. In the latter case, we should remember that V_g is determined by the gate charge, which cannot be directly influenced by the input current due to the infinite impedance between channel and gate. In

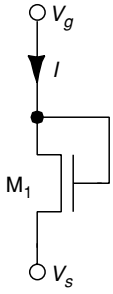


Figure 5.1
Diode-connected nFET.

order to make the circuit work, the input node and the gate have to be in a negative feedback loop that controls the gate charge and keeps the MOSFET in saturation. When the input current is fed into the drain terminal, the feedback from gate to drain is negative and thus the feedback from drain to gate must be positive. As shown in Fig. 5.1, the gate can simply be shorted to the drain to complete the negative feedback loop. The MOSFET is then reduced to a two-terminal device with similar characteristics to a diode, and is said to be *diode-connected*. Since the drain of the diode is always reverse-biased with respect to the channel, a diode-connected MOSFET is always in saturation as long as any appreciable current flows. If the source is chosen as the input node, the feedback from source to gate must be made positive, which requires additional transistors.

5.2 Two-Transistor Circuits

Now that we have seen how one transistor can perform multiple functions, we show how two-transistor circuits can perform additional functions; for example, replication, amplification, and reduction of either current inputs or voltage inputs.

Current Mirror

Figure 5.2 shows a diode-connected MOSFET that has a common gate node with another MOSFET of the same type. If both MOSFETs have fixed source voltages and are in saturation, they act as current sources. Moreover, if both MOSFETs are of the same size and have the same source voltage, they source

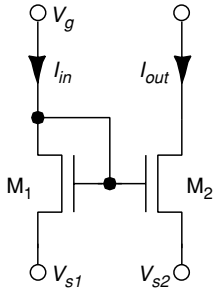


Figure 5.2
Current mirror with adjustable amplification.

the same current, which is why the device is called a *current mirror*. The input current I_{in} through the diode-connected transistor M_1 sets the common gate voltage V_g and hence the output current I_{out} of the second transistor M_2. The output current can be scaled by choosing different transistor sizes, or by choosing different source potentials V_{s1} and V_{s2} for the two MOSFETs. The dependence of the output on the difference in the source potentials is described by

$$I_{out} = e^{(V_{s1}-V_{s2})/U_T} I_{in}. \tag{5.2.1}$$

While this strategy scales the current by a factor that is exponential in the source-voltage difference, as seen from Eq. 5.2.1, the transistor size strategy scales the current by a fixed design factor that is linear in the ratio of the width-to-length ratios.

Above threshold, the matching of input and output currents depends more strongly on the drain voltage of M_2, due to the larger Early effect, and the dependence of the gain on the difference of the source voltages is weaker.

For bidirectional input currents I_{in}, a current mirror acts as a *half-wave rectifier*, because the input transistor has a very high impedance in the opposite direction. Half-wave rectification is useful in many applications, for example, in the implementation of linear-threshold neurons (Chapter 6).

Source Follower

The *source-follower* circuit linearly transforms a voltage at a high-impedance input terminal into a voltage at a lower-impedance output terminal, such that the output signal is able to drive larger loads than the input signal. It is

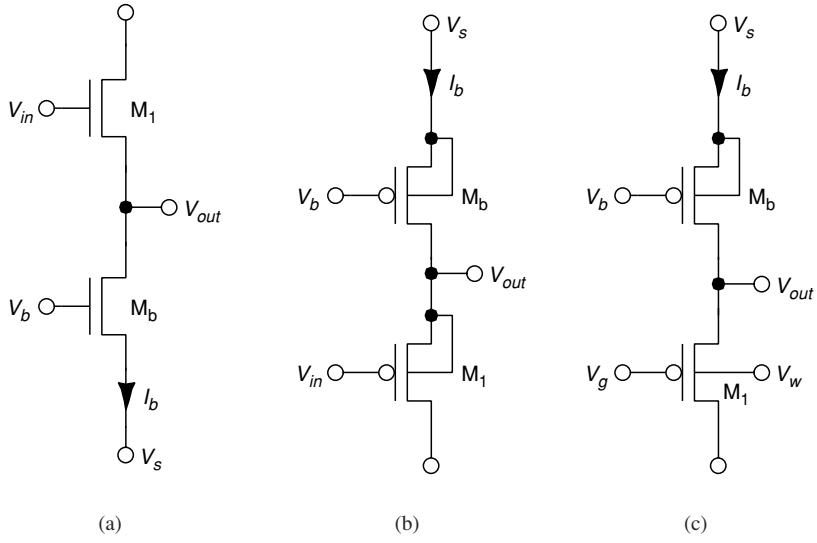


Figure 5.3
Source-follower circuits with (a) range reduction, (b) unity gain, and (c) range reduction and independent time-constant and offset control.

constructed by connecting a fixed current source to the source of a MOSFET operated in saturation. As we saw in Section 5.1, a current source can easily be implemented using a saturated MOSFET with fixed source and gate voltages. This arrangement results in a circuit with two MOSFETs of the same type connected in series (Fig. 5.3(a)), where $\mathrm{M_b}$ is the current source and M_1 is the input transistor. The input voltage V_{in} is applied to the gate voltage of M_1 and the output voltage V_{out} is the source voltage of M_1. In the subthreshold domain, the output voltage adjusts itself according to Eq. 5.1.1:

$$\boxed{V_{out} = \kappa_n V_{in} - U_T \log\left(\frac{I_b}{I_{n0}}\right) = \kappa_n V_{in} - \kappa_b V_b + V_s} \tag{5.2.2}$$

where κ_b is the subthreshold slope factor of $\mathrm{M_b}$ and we assume that both transistors have the same dimensions. Note that V_{out} is linearly related to V_{in} with a positive slope of $\kappa_n < 1$ and a fixed offset that depends on the bias current I_b. The output voltage follows the input voltage with a gain of κ_n, hence the name of the circuit. The measured input-output characteristic

of this circuit is shown in Fig. 5.4. The condition for saturation of the biasing transistor, and thus proper operation of the circuit, is $V_{out} > V_s + 4U_T$: That is, $V_{in} > \kappa_n^{-1}(\kappa_b V_b + 4U_T)$. In this range, the residual non-linearity of the characteristic is mainly due to the body effect of M_1: κ_n is not constant. The dependence of κ_n on V_{in}, as obtained by differentiating the curve of Fig. 5.4, is shown in Fig. 5.5. Given Eq. 5.2.2 we see that we can alternatively use V_b as a high-impedance input terminal to obtain a negative gain of $-\kappa_b$. However, this arrangement has the disadvantage that the current flowing through the circuit and thus the transient dynamics, depend on the input signal.

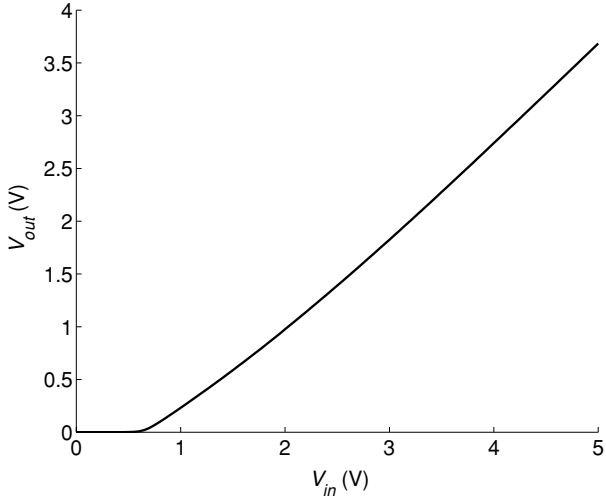


Figure 5.4
Transfer characteristic of source follower circuit.

It is possible to get around the κ reduction factor in the transfer characteristic, if the bulk potential of the input MOSFET can be controlled independently. As mentioned previously, in a CMOS process this independence is possible for only one type of MOSFET: The one that sits in a well with opposite doping from the substrate. For a source-follower circuit constructed of well MOSFETs, the wells of the transistors can be connected to their respective sources, as shown in Fig. 5.3(b) for pFETs in *n*-wells. Since the subthreshold slope

factors are equal for both MOSFETs,

$$V_{out} = V_{in} + \frac{U_T}{\kappa_p} \log\left(\frac{I_b}{I_{p0}}\right) = V_{in} + V_s - V_b\,. \tag{5.2.3}$$

The output voltage now follows the input voltage with unity gain and a fixed offset $V_s - V_b$. The bias transistor stays in saturation for $V_{in} < V_b - 4U_T$.

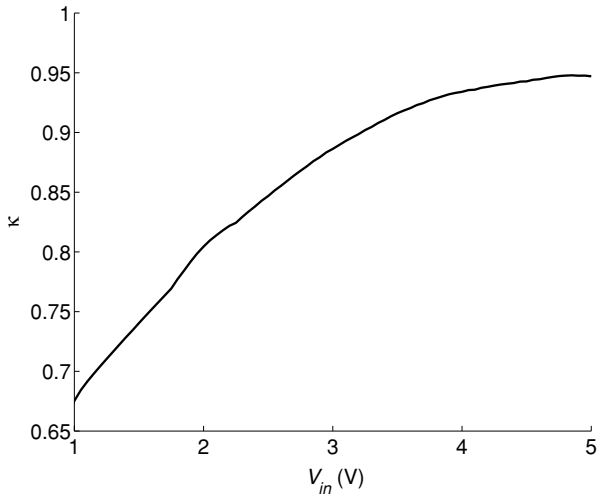


Figure 5.5
Dependence of subthreshold slope factor of source follower circuit on input voltage.

The well of M_1,V_w, can be used as an additional input terminal as shown in Fig. 5.3(c). This configuration affords an additional degree of freedom in controlling the output voltage, which is now given by

$$V_{out} = (1-\kappa_p)V_w + \kappa_p V_g + U_T \log\left(\frac{I_b}{I_{p0}}\right) \tag{5.2.4}$$

$$= (1-\kappa_p)V_w + \kappa_p V_g + \kappa_b (V_s - V_b)\,. \tag{5.2.5}$$

The additional terminal allows the offset of V_{out} to be adjusted independently of I_b and thus of the temporal dynamics of the circuit. We can either apply the input voltage at the gate terminal and adjust the offset with V_w, or apply the input voltage at the well terminal and adjust the offset with V_g. In the former case the gain is κ_p and in the latter case it is $1-\kappa_p$, which is usually smaller

because subthreshold slope factors tend to be larger than 0.5. Since the well has a smaller input impedance than the gate, due to its larger capacitance and *leakage currents*, the well-input option should only be chosen if the voltage source at the input is sufficiently strong. In either case, the circuit should be operated such that V_w is never significantly smaller than V_{out}, otherwise a large leakage current flows from the output node into the well via the forward-biased source node of M_1.

The source-follower circuit can thus be used to introduce an adjustable offset to a voltage, and to optionally reduce the voltage range with a gain factor smaller than one. Its main function, however, is that of an *impedance converter*: The circuit transforms a weakly-driven voltage signal into a more strongly-driven voltage signal.

Inverting Amplifier

A completely different function is obtained if two MOSFETs of opposite types are connected in series at their drains with their sources at fixed potentials, as shown in Fig. 5.6(a). In this common-drain configuration, both MOSFETs act as current sources, as long as they are in saturation. Since both transistors have to source the same current, a small change ∂V_g in the gate voltage of one transistor results in a large change ∂V_{out} in the common drain voltage in the opposite direction. The circuit thus acts as an *inverting voltage amplifier*. The gains A_n and A_p of the gate inputs to the two transistors can be computed from the Early effect as

$$A_n \equiv \frac{\partial V_{out}}{\partial V_{gn}} = -\frac{\kappa_n}{U_T}\frac{V_{nE}V_{pE}}{V_{nE}+V_{pE}} \tag{5.2.6}$$

$$A_p \equiv \frac{\partial V_{out}}{\partial V_{gp}} = -\frac{\kappa_p}{U_T}\frac{V_{nE}V_{pE}}{V_{nE}+V_{pE}} \tag{5.2.7}$$

where κ_n and κ_p are the subthreshold slope factors and V_{nE} and V_{pE} are the Early voltages of the nFET and the pFET respectively. The gains A_n and A_p are typically between -100 and -1000, so that both transistors will be simultaneously in saturation for variations of only a few millivolts on one of the gate terminals. In the above-threshold domain, the absolute values of the gains are smaller due to the reduced Early voltages. The operating point and the current through the circuit can be set by adjusting the source voltages of the two MOSFETs accordingly.

Now consider the circuit in Fig. 5.6(b) where an input signal V_{in} is applied to both gates. The resulting gain $A = \partial V_{out}/\partial V_{in}$ is now the sum of the gains

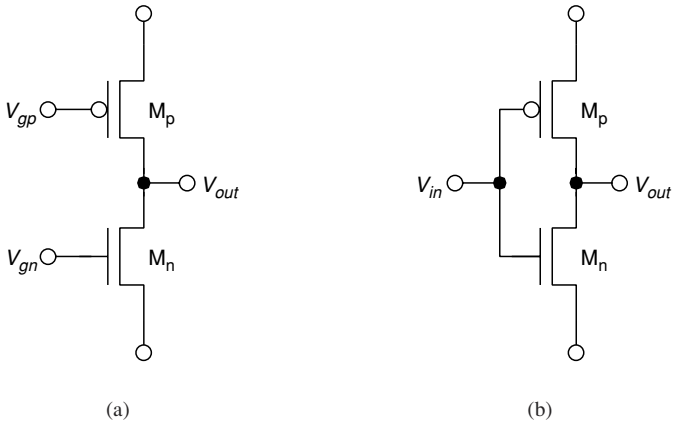


Figure 5.6
Inverting amplifier with (a) two input terminals and (b) one input terminal. The circuit in (b) is also used as an inverter in digital CMOS logic.

for the individual gate inputs. This circuit, with the source voltages connected to the respective power supply rails, is the basic building block of digital CMOS logic. The circuit is called an *inverter*, because if V_{in} is close to one of the power supply voltages representing one logic state, then V_{out} is close to the other power supply voltage representing the other logic state. While the circuit is in a given logic state the MOSFETs are turned off and nearly no power is consumed, which is one reason why CMOS logic is popular for low-power designs.

5.3 Differential Pair and Transconductance Amplifier

In the following, we will turn to some slightly more complex circuits in order to introduce the principles of the transconductance amplifier, which is used in a variety of circuit configurations in analog circuit design.

Differential Pair

The differential pair has the same basic structure as the source follower, except that the bias current I_b is now shared by two MOSFETs M_1 and M_2 whose sources are connected to the drain of the bias MOSFET M_b, as shown

in Fig. 5.7. The sharing of the current between M_1 and M_2 depends on their respective gate voltages V_1 and V_2. If all MOSFETs are operated below threshold and in saturation and we assume that M_1 and M_2 have the same subthreshold slope factor κ_n, we obtain

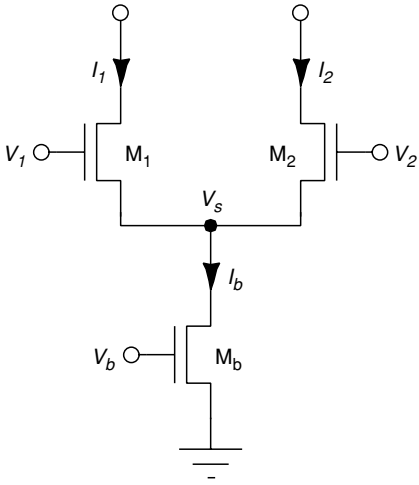


Figure 5.7
Differential pair.

$$I_1 = I_b \frac{e^{\kappa_n V_1/U_T}}{e^{\kappa_n V_1/U_T} + e^{\kappa_n V_2/U_T}} \tag{5.3.1}$$

and

$$I_2 = I_b \frac{e^{\kappa_n V_2/U_T}}{e^{\kappa_n V_1/U_T} + e^{\kappa_n V_2/U_T}}. \tag{5.3.2}$$

The dependence of these two currents on the difference of the input voltages is shown in Fig. 5.8. The curves have a sigmoidal shape. They are almost linear for small voltage differences and saturate at I_b for large voltage differences. Such compressive nonlinearities are very useful for the implementation of different functions, especially in the context of neural networks. What makes the circuit even more useful is the fact that to a first approximation (neglecting the Early effect), the output currents depend only on the difference of the input voltages: The circuit has a small *common-mode sensitivity*. Given that voltages are differential rather than absolute quantities such a property is very useful.

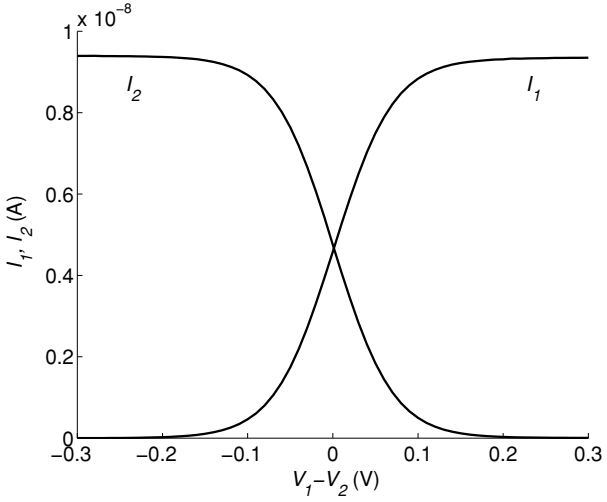


Figure 5.8
Dependence of differential pair output currents on differential input voltage.

The condition for saturation of $\mathrm{M_b}$ is

$$e^{-V_s/U_T} << 1\,. \tag{5.3.3}$$

With

$$e^{-V_s/U_T} = \frac{e^{\kappa_b V_b/U_T}}{e^{\kappa_b V_b/U_T} + e^{\kappa_n V_1/U_T} + e^{\kappa_n V_2/U_T}} \tag{5.3.4}$$

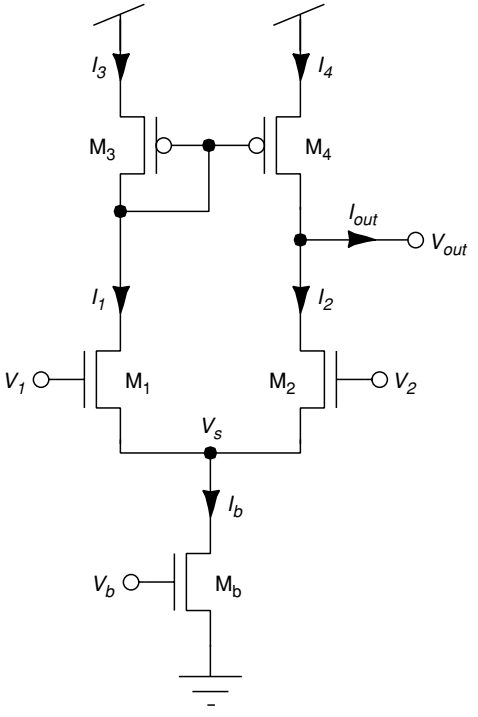
the condition for saturation of $\mathrm{M_b}$ becomes

$$\max(V_1, V_2) > \kappa_n^{-1}(4U_T + \kappa_b V_b) \tag{5.3.5}$$

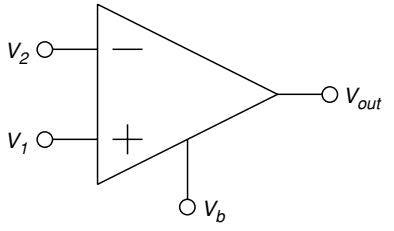
if $|V_1 - V_2| > 4U_T$. If $\mathrm{M_b}$ is not in saturation, the currents depend strongly on the common mode of the input voltages.

Transconductance Amplifier

The two output currents in the differential pair circuit can be subtracted from one another to form a single bidirectional output current. The subtraction is performed by connecting a current mirror of the complementary transistor type to the differential pair, as shown in Fig. 5.9(a). The resulting circuit is the simplest version of a *differential transconductance amplifier*, whose symbol is shown in Fig. 5.9(b). As long as all MOSFETs stay in saturation and the



(a)



(b)

Figure 5.9
Simple differential transconductance amplifier. (a) Schematic diagram. (b) Circuit symbol with inverting (-) and non-inverting (+) inputs.

differential pair is operated below threshold, the output current is given by

$$\boxed{I_{out} = I_1 - I_2 = I_b \frac{e^{\kappa_n V_1/U_T} - e^{\kappa_n V_2/U_T}}{e^{\kappa_n V_1/U_T} + e^{\kappa_n V_2/U_T}} = I_b \tanh\left(\frac{\kappa_n}{2U_T}(V_1 - V_2)\right)} \; . \tag{5.3.6}$$

This relationship is confirmed by the the measured data shown in Fig. 5.10. For small differential voltages it is approximately linear:

$$I_{out} \approx g_m (V_1 - V_2) \tag{5.3.7}$$

where

$$g_m = \frac{I_b \kappa_n}{2U_T} \tag{5.3.8}$$

is the *transconductance* of the amplifier. The reason for this name is the fact although g_m has the dimensions of a conductance, the output current is measured at a different terminal to the pair across which the input voltage gradient is applied.

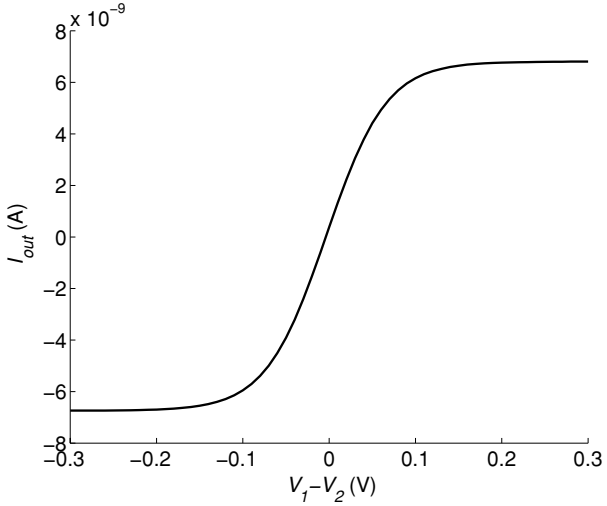


Figure 5.10
I-V transfer characteristics of simple differential transconductance amplifier.

If the differential pair is operated above threshold, it can be shown that

$$I_{out} = \frac{\beta}{2}(V_1 - V_2)\sqrt{\frac{4I_b}{\beta} - (V_1 - V_2)^2} \tag{5.3.9}$$

for $|V_1 - V_2| \leq \sqrt{2I_b/\beta}$. The transconductance is thus given by

$$g_m = \sqrt{\beta I_b}. \tag{5.3.10}$$

Because the input voltages are applied to insulated gates, the input conductances of the transconductance amplifier are negligible and the input currents are close to zero under steady state conditions. The *output conductance* depends on the Early voltages of M_2 and M_4:

$$g_d = -\frac{\partial I_{out}}{\partial V_{out}} = \frac{I_2}{V_{E2}} + \frac{I_4}{V_{E4}} \tag{5.3.11}$$

where V_{E2} and V_{E4} are the Early voltages of M_2 and M_4 respectively. For $V_{E2} = V_{E4} \equiv V_E$ this simplifies to

$$g_d \approx \frac{I_b}{V_E}\,. \tag{5.3.12}$$

In addition to the saturation condition for $\mathrm{M_b}$ in subthreshold (Eq. 5.3.5), we have to consider the saturation conditions for the other transistors. For practical purposes, M_1 and M_3 will always be in saturation because M_3 is diode-connected and the drain of M_1 is thus at a high voltage. The saturation conditions for M_4 and M_2 restrict the output voltage range for subthreshold operation to

$$\kappa_n \max(V_1, V_2) - \kappa_b V_b + 4U_T \approx V_s + 4U_T < V_{out} < V_{dd} - 4U_T\,. \tag{5.3.13}$$

Figure 5.11 shows the relationship between output current and output voltage for given equal input voltages. The output conductance is the absolute value of the slope in the linear region. The limiting regions for V_{out}, according to Eq. 5.3.13, are clearly visible.

In the mode described above, the transconductance amplifier is used as a differential-voltage-to-current converter. However, it can also be used in the open-circuit mode as a differential-voltage amplifier. In which case, for

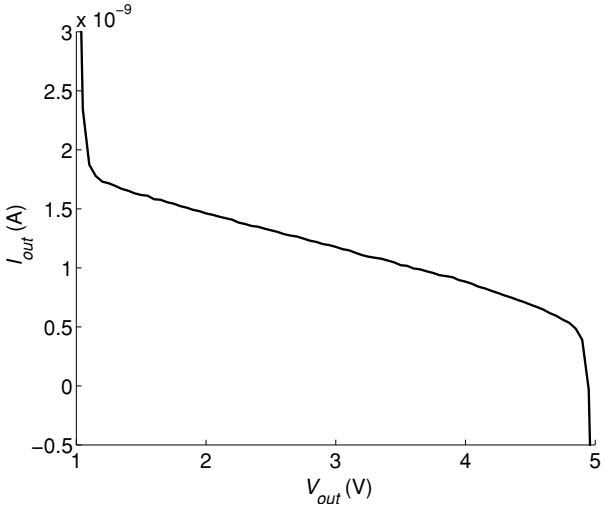


Figure 5.11
I-V characteristics at output of simple differential transconductance amplifier for fixed equal input voltages.

$V_1 \approx V_2$, we obtain

$$0 = dI_{out} = \frac{\partial I_{out}}{\partial (V_1 - V_2)} d(V_1 - V_2) + \frac{\partial I_{out}}{\partial V_{out}} dV_{out} \tag{5.3.14}$$

$$= g_m d\,(V_1 - V_2) - g_d\, dV_{out}\,. \tag{5.3.15}$$

From this result we can compute the *open-circuit voltage gain*

$$A \equiv \frac{dV_{out}}{d(V_1 - V_2)} = \frac{g_m}{g_d}\,, \tag{5.3.16}$$

which in the subthreshold regime can be expressed as

$$A = \frac{\kappa_n}{U_T} \frac{V_{E2} V_{E4}}{V_{E2} + V_{E4}} \approx \frac{\kappa_n V_E}{2U_T}\,. \tag{5.3.17}$$

Above threshold we obtain

$$A \approx \sqrt{\frac{\beta}{I_b}} V_E\,. \tag{5.3.18}$$

Since V_{out} increases with increasing V_1 and decreases with increasing V_2, the gate of M_1 is called the *non-inverting* input terminal and the gate of M_2 the *inverting input* terminal of the amplifier. In the circuit symbol, the

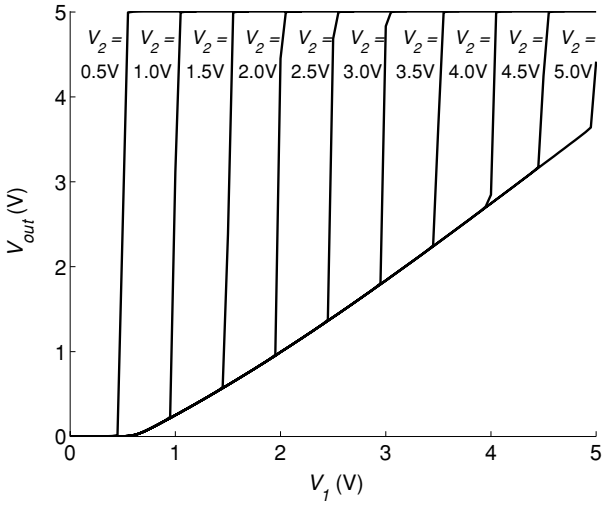


Figure 5.12
Voltage amplification characteristics of simple differential transconductance amplifier.

two input terminals are denoted by a plus and a minus sign respectively, as shown in Fig. 5.9(b). The open-circuit voltage gain increases with the Early voltages, and therefore with the length of the output transistors. Typical subthreshold values are between 100 and 1000. Given this large gain and the unavoidable transistor mismatches introduced by the fabrication process, the device is not normally used as an open-circuit voltage amplifier. In the open-loop configuration, the transconductance amplifier is used mainly as a *comparator*, which outputs a high voltage if $V_1 > V_2$ and a low voltage if $V_1 < V_2$. However, these output voltages are not independent of the input voltages and we therefore do not obtain an ideal comparator with a binary output.

Since in the open circuit V_{out} is normally at one of its limits, we will now determine where those limits lie. If V_1 is larger than V_2, M_4 goes out of saturation. For $V_1 > V_2 + 4U_T$, the current through M_2 is much smaller than the current through M_1 and M_3. Hence, V_{out} goes almost all the way to V_{dd}, shutting off M_4. If V_1 is smaller than V_2, M_2 goes out of saturation; but the current mirror acts, such that $I_1 \approx I_2 \approx I_b/2$. If V_2 is significantly larger

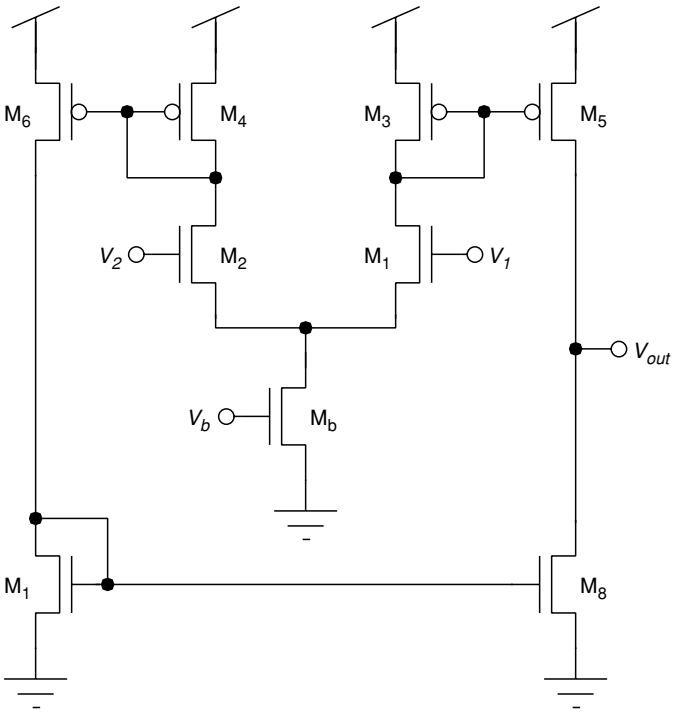


Figure 5.13
Wide-output-range differential transconductance amplifier.

than V_1, the voltage drop across M_2 is close to zero and $V_{out} \approx V_s$. With

$$e^{-V_s/U_T} = \frac{\frac{1}{2}e^{\kappa_b V_b/U_T}}{\frac{1}{2}e^{\kappa_b V_b/U_T} + e^{\kappa_n V_1/U_T}} \tag{5.3.19}$$

we see that V_{out} goes all the way to zero if $V_1 < \kappa_n^{-1}\,(\kappa_b V_b - (4 + \log 2)U_T)$ and

$$V_{out} \approx \kappa_n V_1 - \kappa_b V_b + U_T \log 2 \tag{5.3.20}$$

if $V_1 > \kappa_n^{-1}\,(\kappa_b V_b + (4 - \log 2)U_T)$. The open-circuit voltage characteristics of the simple transconductance amplifier are shown in Fig. 5.12 for different fixed values of V_2. As predicted by Eq. 5.3.20, the lower limit of the output voltage range ramps up linearly with V_1, with a slope of κ_n.

If an application requires an output range extending to both supply rails, the circuit can be expanded as shown in Fig. 5.13. The current I_2 is mirrored

twice, such that the output stage is symmetric and decoupled from the input stage. This decoupling also allows the design of output stages with large open-loop gains (long MOSFETs) or large output currents (MOSFETs with large width-to-length ratios or bipolars (see Chapter 12)). Disadvantages of this extension are the need for almost twice as many transistors as for the basic version, and the increased effect of mismatches due to fabrication tolerances.

Transconductance amplifiers are widely used as elements in circuit applications, where they are usually referred to as *operational amplifiers*. Most commercially-available transconductance amplifiers are more sophisticated and evolved than the versions presented here. They are multistage circuits, which are optimized with respect to a set of performance criteria, some of which relate to their dynamic behavior (stability, gain-bandwidth product, etc.). The treatment of such circuits is beyond the scope of this chapter, but can be found in the literature ((Gregorian, 1999; Huijsing, 2001)).

5.4 Unity-Gain Follower

Most circuit applications use the transconductance amplifier as part of a negative-feedback loop. The negative feedback ensures that the amplifier stays within its operating range. The simplest negative feedback loop is obtained by short-circuiting the output terminal and the inverting input terminal, as shown in Fig. 5.14. The transfer function of this circuit is given by

$$\frac{dV_{out}}{dV_{in}} = \frac{A}{A+1} \approx 1. \tag{5.4.1}$$

Due to the large open-loop voltage gain A, the transfer function is almost

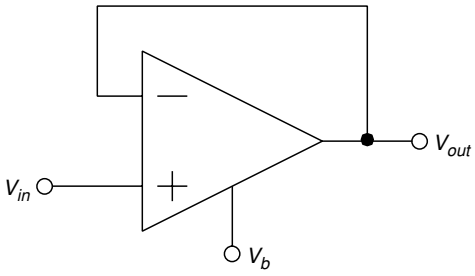


Figure 5.14
Unity-gain follower.

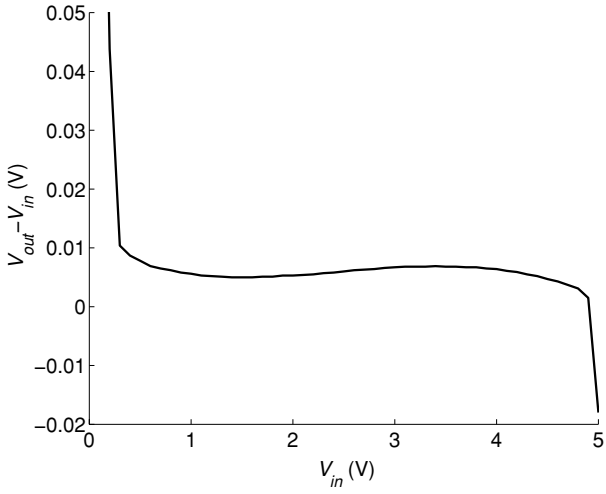


Figure 5.15
Deviation of output voltage from input voltage for simple unity-gain follower.

unity and $V_{out} \approx V_{in}$. The circuit configuration is therefore called *unity-gain follower*. It is used as an impedance converter (also called a *buffer*) and it converts a high input impedance into a lower output impedance. In contrast to the source follower presented in Section 5.2, which is also used as an impedance converter, the unity-gain follower does not introduce a large voltage offset. The measured deviation of the output signal from the input signal, $V_{out} - V_{in}$, as a function of the input voltage is shown in Fig. 5.15. The deviation is about 5 mV, except for very low voltages, where the bias current of the amplifier is practically shut off and the behavior of the circuit gets degraded by leakage effects: At very high voltages, where the current mirror is shut off. This deviation has a random component, which is due to circuit mismatches; and a systematic component, which is due to the amplifier's finite open-loop gain A, as shown by Eq. 5.4.1.

This page intentionally left blank

6 Current-Mode Circuits

During the last 40 years, the vast majority of analog circuits have used voltages to represent and process relevant signals. However, recently, current-mode signal processing circuits, in which signals and state variables are represented by currents rather than voltages (Tomazou et al., 1990), have shown advantages over their voltage-mode counterparts. Their advantages include higher bandwidth, higher dynamic range, and they are more amenable to lower power supplies.

In this chapter, we will describe some basic current-mode circuits commonly used in neuromorphic systems. Although the individual basic circuits are relatively simple, they offer very interesting signal-processing properties when connected to form networks. We start by first describing the current conveyor block which can be used to replace the traditional operational amplifier.

6.1 The Current Conveyor

In voltage-mode circuits, the main building block used to add, subtract, amplify, attenuate, and filter voltage signals is the operational amplifier. In current-mode circuits, the analogous building block is the *current conveyor* (Smith and Sedra, 1968; Wilson, 1990).

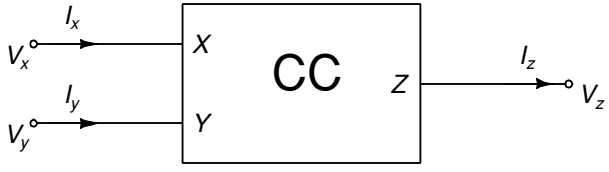


Figure 6.1
Current conveyor.

The original current conveyor (Fig. 6.1) was a three-terminal device (two input terminals X and Y and one output terminal Z) with the following properties:

1. The potential at its input terminal (X) is equal to the voltage applied at the other input terminal (Y).

2. An input current that is forced into node X results in an equal amount of current flowing into node Y.

3. The input current flowing into node X is conveyed to node Z, which has the characteristics of a high output impedance current source.

The term *conveyor* refers to the third property above: Currents are conveyed from the input terminal to the output terminal, while decoupling the circuits connected to these terminals.

The simplest CMOS implementation of a current conveyor is a single MOS transistor (Fig. 6.2(a)). When used as a current buffer, it conveys current from a low impedance input node X to a high impedance output node Z: And when used as a source-follower, its source terminal X can follow its gate Y. A more elaborate current-controlled conveyor is shown in Fig. 6.2(b). This basic two-transistor current conveyor is used in many neuromorphic circuits (Cohen and Andreou, 1992; Boahen and Andreou, 1992; Delbrück and Mead, 1994) and is a key component of the current-mode winner-take-all circuit that is analyzed in Section 6.3. It has the desirable property of having the voltage at node X controlled by the current being sourced into node Y. If the transistors are operated in the subthreshold domain, the monotonic function that links the voltage at the node X to the *current* being sourced into Y is a logarithm. As voltages at the nodes Y and X are decoupled from each other, V_x can be clamped to a desired constant value by chosing appropriate values of I_y. An example of a system level application is shown in Fig. 6.2(c): A current-controlled conveyor is used to make multiple copies of the current I_y. The input current I_y sets the value of V_x independent of V_y[1]. This property can be useful for processing time-varying signals: If the current I_y represents a time-varying signal, and if the capacitance on node X is not negligible (for example, the node X is connected to many transistors), then keeping the voltage V_x clamped to a constant value ensures that the output currents $I_{y_1}, I_{y_2}, \ldots, I_{y_n}$ faithfully follow I_y.

Sedra and Smith (1970) reformulated the definition of the current conveyor, describing a new circuit that combines both voltage and current-mode signal processing characteristics. This new type of current conveyor (denoted as conveyor of class II) is represented by the symbol shown in Fig. 6.1 and its

1 As opposed to simple current mirrors in which $V_x \equiv V_y$.

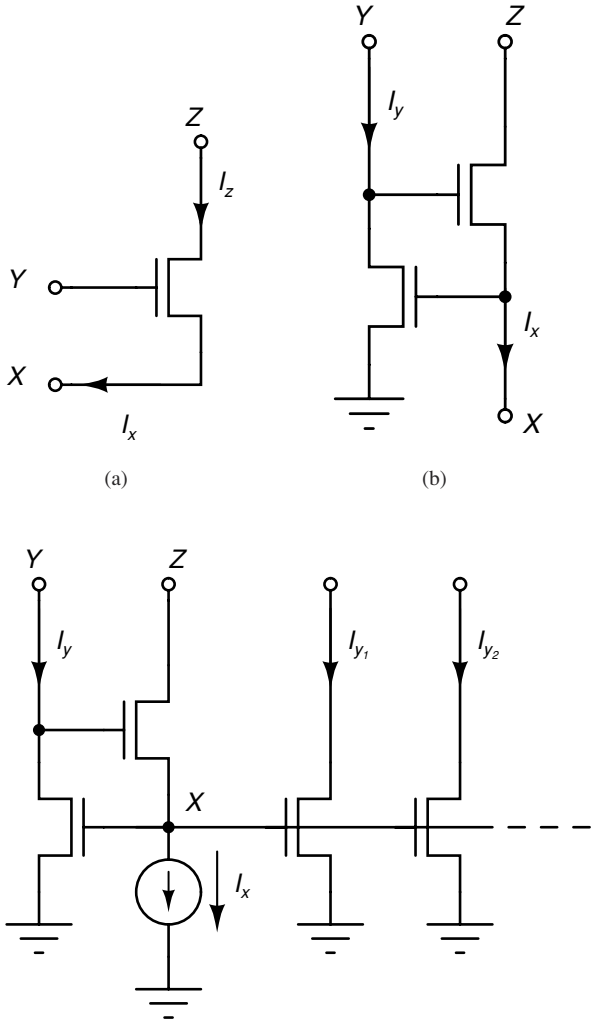
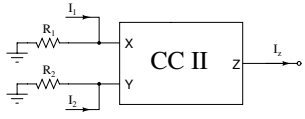
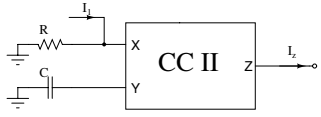
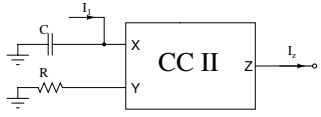
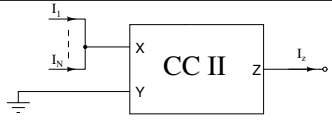


Figure 6.2
Current conveyor implementations. (a) Single MOS transistor current conveyor. (b) Two MOS transistor current-controlled conveyor. (c) System application of current conveyor for the generation of multiple copies of an input current.

Table 6.1
Applications of the class II current conveyor to current-mode signal processing circuits.

Functional Element	Function	Realization Using Current Conveyor
Current Amplifier	$I_z = \frac{R_1}{R_2} I_1$	
Current Differentiator	$I_z = CR\frac{dI_1}{dt}$	
Current Integrator	$I_z = \frac{1}{CR}\int I_1\, dt$	
Current Summer	$I_z = -\sum_j I_j$	

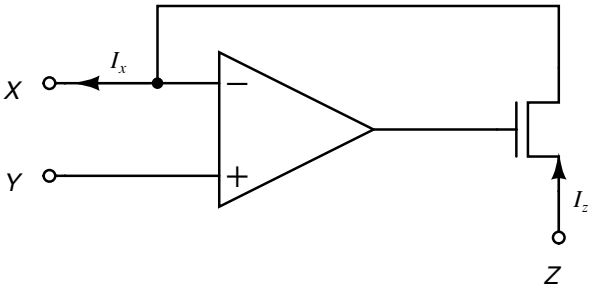
input-output characteristics are defined as

$$\begin{bmatrix} V_x \\ I_y \\ I_z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \pm 1 \end{bmatrix} \begin{bmatrix} V_y \\ V_z \\ I_x \end{bmatrix}. \tag{6.1.1}$$

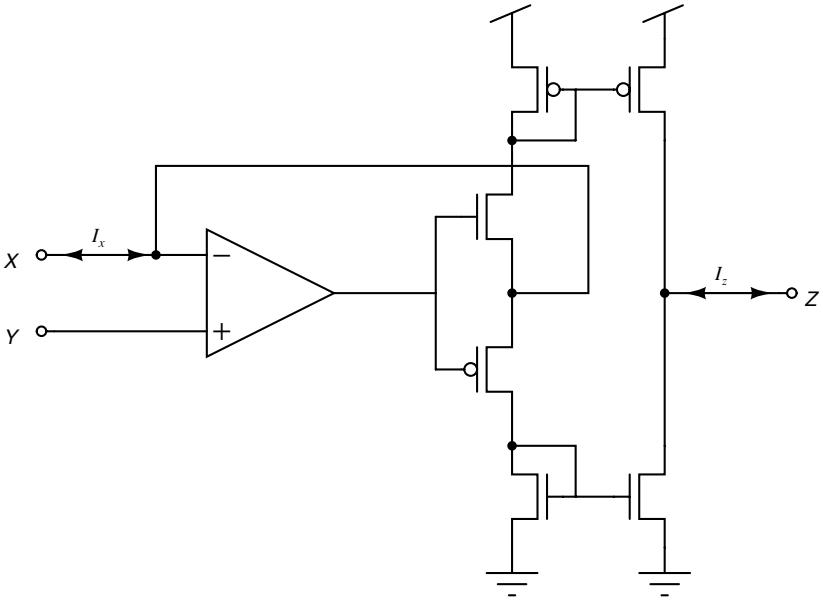
The input voltages V_x and V_y are linked by a unity gain relationship ($V_x = V_y$); the terminal Y has infinite impedance ($I_y = 0$) and the current forced into node X is conveyed to the high impedance output node Z with a ± 1 gain. Examples of CMOS circuits that comply with the definition of a class II current conveyor are shown in Fig. 6.3. These circuits are typically used as current precise rectifiers, but they (and similar circuits) can implement many other signal processing functions. Table 6.1 shows some possible applications of class II current conveyors.

6.2 The Current Normalizer

Signal normalization is another important signal processing function that can be realized using current-mode circuits. This function is useful for rescaling data from multiple sources into a standard reference frame. Normalization



(a)



(b)

Figure 6.3
CMOS implementations of class II current conveyors. (a) Uni-directional current conveyor: This circuit operates correctly only for positive values of I_z (*that is* currents that are sourced into node Z). (b) Bi-directional current conveyor: This circuit conveys both positive and negative input currents.

is often used in artificial signal processing systems and is also ubiquitous in biological nervous systems.

A current-mode circuit that normalizes its input signals is shown in Fig. 6.4. It receives analog continuous time input currents and provides normalized output currents. The circuit is based on the Gilbert normalizer circuit (Tomazou et al., 1990) that was originally designed using bipolar transistors (see Chapter 7). It is a modular circuit that can be extended to an arbitrary number of cells by simply connecting additional current mirrors to the common node V_c. If the input currents are subthreshold, the circuit is characterized by the equations

$$\begin{aligned} I_{in_i} &= I_0 e^{\kappa \frac{V_{d_i}}{U_T}} \\ I_{out_i} &= I_0 e^{\kappa \frac{V_{d_i}}{U_T} - \frac{V_c}{U_T}} \end{aligned} \tag{6.2.1}$$

where i is the index of the i^{th} cell of the circuit, U_T is the thermal voltage and κ is the subthreshold slope coefficient (see Section 3.2). By applying Kirchhoff's current law to the common node V_c we obtain

$$\sum_{i=1}^{N} I_{out_i} = I_b \tag{6.2.2}$$

where I_b is a constant current set by the control voltage V_b. We use this constraint to solve Eq. 6.2.1 for V_c, and to derive the dependence of the output current on the input currents:

$$I_{out_i} = I_b \frac{I_{in_i}}{\sum_j I_{in_j}}. \tag{6.2.3}$$

The output current of each cell I_{out_i} is directly proportional to its input current (with a proportionality constant I_b), but scaled by the sum of all the input currents $\sum_j I_{in_j}$.

6.3 Winner-Take-All Circuits

A winner-take-all (WTA) circuit is a network of competing cells (neural, software, or hardware) that reports only the response of the cell that has the strongest activation while suppressing the responses of all other cells. The circuit essentially implements a $\max()$ function. These circuits are typically used to implement and model competitive mechanisms among populations

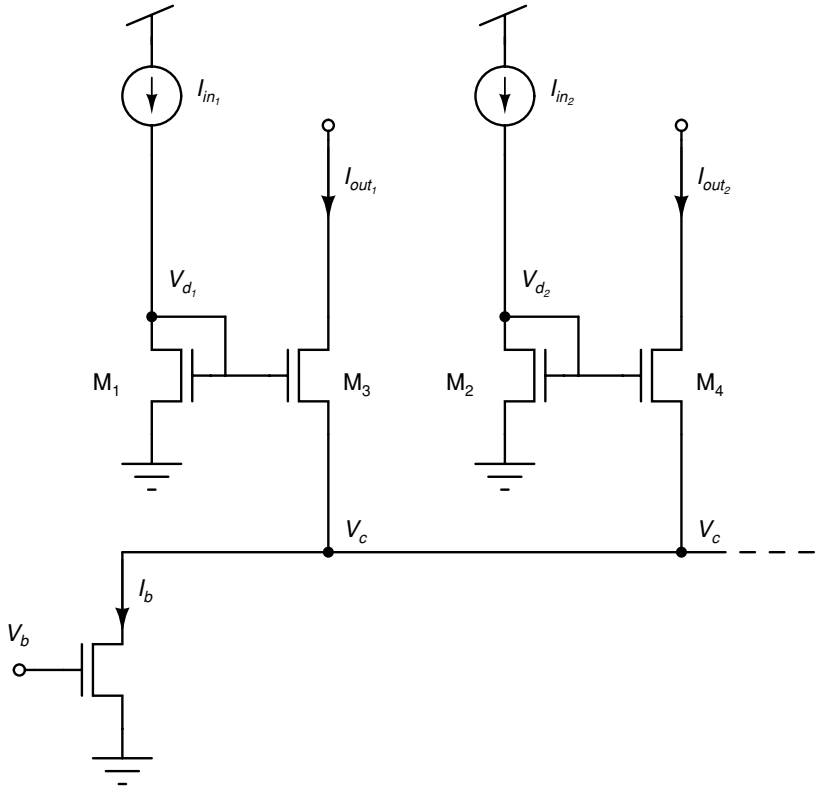


Figure 6.4
Two-input current normalizer circuit.

of neurons. For example, they are used to select specific regions of an input space (Yuille and Geiger, 1995). Many WTA networks have been implemented both in software (von der Malsburg, 1973; Nass and Cooper, 1975; Grossberg, 1978; Kaski and Kohonen, 1994) and in hardware (Lazzaro et al., 1989; Choi and Sheu, 1993; Starzyk and Fang, 1993; Serrano and Linares-Barranco, 1995; Lau and Lee, 1998; Demosthenous et al., 1998; Liu, 2000; Hahnloser et al., 2000; Indiveri, 2001a).

In this section we analyze a class of WTA networks that emulate biological networks, consisting of a cluster of excitatory neurons that innervate a global feedback inhibitory neuron. These networks have been implemented in aVLSI and applied to a wide variety of tasks, including selective attention (Bra-

jovic and Kanade, 1998; Wilson et al., 1999; Indiveri, 2000, 2001b), auditory localization (Lazzaro and Mead, 1989), visual stereopsis (Mahowald, 1994), smooth pursuit/tracking (Etienne-Cummings et al., 1996; Horiuchi and Koch, 1999), and detection of heading direction (Indiveri et al., 1996; Mudra et al., 1999).

WTA Neural Models

We shall focus on a particularly simple yet powerful model that describes a population of N homogeneous excitatory units that excites a single global inhibitory unit which feedbacks to inhibit all the excitatory units (Fig. 6.5). For sake of simplicity, we neglect the dynamics of the system and examine only the steady-state solutions. Dynamic properties of these networks and of other physiological models of competitive mechanisms are described in detail in Kaski and Kohonen (1994); Yuille and Geiger (1995); Grossberg (1978); Ben-Yishai et al. (1995).

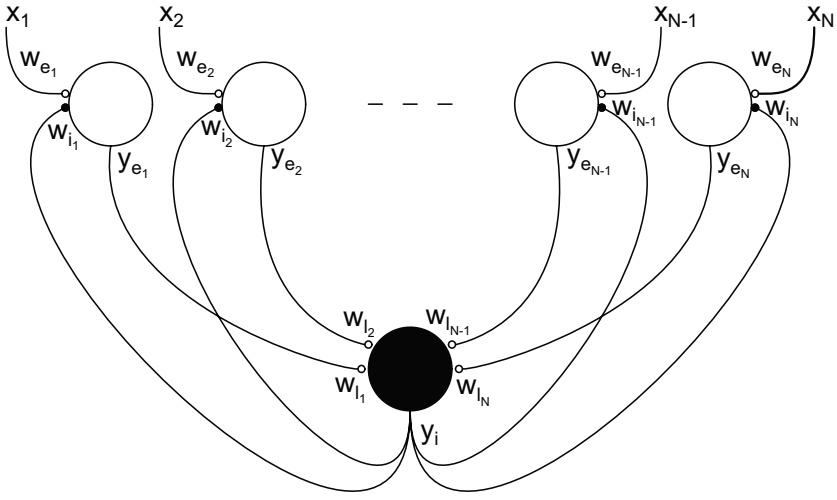


Figure 6.5
Network of N excitatory neurons (empty circles) projecting to one common inhibitory neuron (filled circle), which provides feedback inhibition. Small filled circles indicate inhibitory synapses and small empty circles indicate excitatory synapses. $x_1 \ldots x_N$ are external inputs; $y_{e_1} \ldots y_{e_N}$ are the outputs of the excitatory neurons; y_i is the output of the inhibitory neuron; $w_{e_1} \ldots w_{e_N}$ are the excitatory synaptic weights of the external inputs; $w_{l_1} \ldots w_{l_N}$ are the excitatory weights onto the global inhibitory neuron; and $w_{i_1} \ldots w_{i_N}$ are the inhibitory weights from the inhibitory neuron onto the excitatory neurons.

Consider a network (Fig. 6.5), in which the external input to the j^{th} excitatory neuron is x_j, the response of the j^{th} excitatory neuron is y_{e_j}, the response of the inhibitory neuron is y_i; and in which the weights of the synapses from the external inputs to the excitatory neurons, from the inhibitory neuron to excitatory ones and from the excitatory neuron to the inhibitory one are w_{e_j}, w_{i_j} and w_{l_j} respectively. We can write this network as

$$y_{e_j} = f(w_{e_j} x_j - w_{i_j} y_i) \tag{6.3.1}$$

$$y_i = f\left(\sum_{j=1}^{N} w_{l_j} y_{e_j}\right) \tag{6.3.2}$$

where $f(\cdot)$ denotes the transfer function of both excitatory and inhibitory neurons. This system of coupled equations describes the recurrent interactions between excitatory neurons and the inhibitory neuron. We explore the behavior of the system by considering three special cases:

1. The case in which all neurons have a linear transfer function ($f(x) = x$).
2. The case in which the neurons are linear-threshold ($f(x) = \max(0, x)$), and all external inputs are identical.
3. The case in which the neurons are linear-threshold, and one external input is much larger than all others.

More general cases using non-linear transfer functions are difficult to solve analytically; however, they can be studied using numerical simulations.

Linear Units If the neurons are fully linear ($f(x) = x$) we can solve the system analytically:

$$\begin{aligned} y_{e_j} &= w_{e_j} x_j - w_{i_j} y_i \\ y_i &= \sum_j w_{l_j} (w_{e_j} x_j - w_{i_j} y_i) \end{aligned} \tag{6.3.3}$$

which implies that

$$\begin{aligned} y_{e_j} &= w_{e_j} x_j - \frac{w_{i_j} \sum_k w_{l_k} w_{e_k} x_k}{1 + \sum_k w_{l_k} w_{i_k}} \\ y_i &= \frac{\sum_k w_{l_k} w_{e_k} x_k}{1 + \sum_k w_{l_k} w_{i_k}}. \end{aligned} \tag{6.3.4}$$

In the simplified case, we assume that all the weights of each kind are the same:

$$\begin{aligned} w_{e_j} &= w_e && \forall j \\ w_{i_j} &= w_i && \forall j \\ w_{l_j} &= w_0 && \forall j \end{aligned}$$

and so

$$y_{e_j} = w_e x_j - \frac{w_e \sum_k x_k}{\frac{1}{w_i w_0} + N} \tag{6.3.5}$$

The output of each neuron is proportional to its input, but has a normalizing term subtracted. Equation (6.3.5) shows that the response y_{e_j} of a linear excitatory neuron can have both positive and negative values, depending on the inputs x_k, on its connection weights w_e, w_i, w_0 and on the total number of excitatory neurons N.

Linear Threshold Units with Uniform External Inputs The *half-wave rectification* function ($f(x) = \max(0, x)$) is a more biologically realistic function than the linear one of the previous case. Neurons with this transfer function have a response of only positive values.

In this case, the system of linear equations (Eqs. 6.3.2) becomes a system of *non-linear* coupled equations, and it is not longer possible to obtain a general closed form solution. However if all external inputs are identical ($x_j = x_0 \ \forall j$), we can reduce the system to

$$\begin{aligned} y_{e_j} &= \max(0, w_{e_j} x_0 \ - \ w_{i_j} y_i) \\ y_i &= \max\left(0, \sum_j w_{l_j} y_{e_j}\right) \end{aligned} \tag{6.3.6}$$

and if we make the working hypothesis that $(w_{e_j} x_0 \ - \ w_{i_j} y_i) > 0 \ \ \forall \ j$, then we obtain the linear system:

$$\begin{aligned} y_{e_j} &= w_{e_j} x_0 \ - \ w_{i_j} y_i \\ y_i &= \sum_j w_{l_j} (w_{e_j} x_0 - w_{i_j} y_i) \end{aligned} \tag{6.3.7}$$

which yields

$$y_{e_j} = x_0 \frac{w_{e_j}\,(1 + \sum_k w_{l_k} w_{i_k}) - w_{i_j} \sum_k w_{e_k} w_{l_k}}{1 + \sum_k w_{l_k} w_{i_k}}$$
$$y_i = x_0 \frac{\sum_j w_{e_j} w_{l_j}}{1 + \sum_j . w_{l_j} w_{i_j}} \tag{6.3.8}$$

If the synapses from external inputs and those from the inhibitory neuron have equal strength ($w_{e_j} = w_{i_j} = w_0 \; \forall \; j$), then

$$y_{e_j} = \frac{x_0}{\frac{1}{w_0} + \sum_k w_{l_k}}$$
$$y_i = \frac{x_0 \sum_j w_{l_j}}{\frac{1}{w_0} + \sum_k w_{l_k}}. \tag{6.3.9}$$

The hypothesis used to obtain Eq. 6.3.7 is satisfied for all values of $x_0 > 0$, $w_0 > 0$, and $w_{l_j} > 0 \quad \forall \; j$. In summary, if all inputs are equal, then all excitatory linear threshold units have identical outputs which are equal to the input normalized by a term that is directly proportional to the weights w_{l_j} and inversely proportional to w_0.

Linear Threshold Units with One Input Much Greater than All Others Now consider the case in which one input (say the external input to unit j_0, x_{j0}) is much greater than all other external inputs ($x_{j0} \gg x_j \quad \forall j \neq j_0$) and the synaptic weights are as described above. Again, we assume a priori that the weighted external excitatory input to unit j_0 exceeds the inhibitory input to the same unit ($w_{e_{j0}} x_{j0} - w_{i_{j0}} y_i > 0$) and that the weighted external inputs to all other excitatory inputs don't ($w_{e_j} x_j - w_{i_j} y_i < 0 \; \forall \; j \neq j0$). Under these assumptions, Eq. 6.3.6 can be rewritten

$$y_{e_{j0}} = (w_{e_{j0}} x_{j0} - w_{i_{j0}} y_i)$$
$$y_{e_j} = 0 \; \forall j \neq j0$$
$$y_i = w_{l_{j0}} (w_{e_{j0}} x_0 - w_{i_{j0}} y_i) \tag{6.3.10}$$

which can be simplified to yield

$$\begin{aligned} y_{e_{j0}} &= \frac{w_{e_{j0}} x_{j0}}{1 + w_{l_{j0}} w_{i_{j0}}} \\ y_{e_j} &= 0 \ \ \forall j \neq j0 \\ y_i &= \frac{w_{e_{j0}} w_{l_{j0}} x_{j0}}{1 + w_{l_{j0}} w_{i_{j0}}}. \end{aligned} \tag{6.3.11}$$

This solution satisfies the assumption that $w_{e_{j0}} x_{j0} > w_{i_{j0}} y_i$ for all values of $w_{e_{j0}}$, x_{j0}, and $w_{i_{j0}}$. It also satisfies the *a priori* assumption that $w_{e_j} x_j < w_{i_j} y_i$ as long as the external input x_{j0} is sufficiently large with respect to all other x_j inputs.

Summarizing: if one external input is much greater than the other inputs, then all excitatory linear threshold units, except the one receiving the strongest input, are suppressed. The output of the winning unit is a normalized version of the input, and the normalizing factor is directly proportional to the connection weights $w_{i_{j0}}$, $w_{l_{j0}}$, and inversely proportional to $w_{e_{j0}}$.

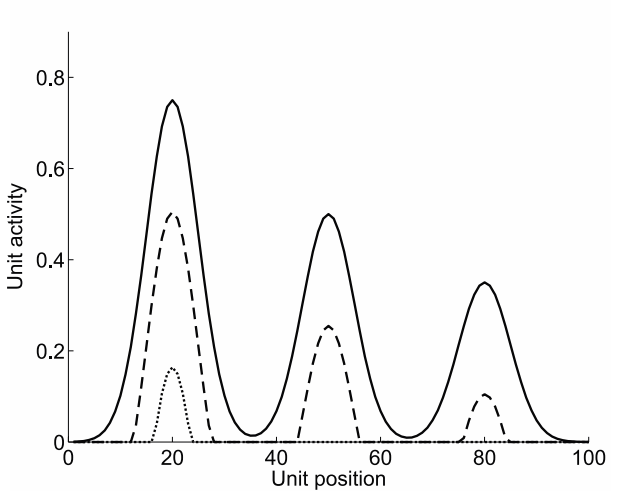


Figure 6.6
Simulations of a WTA network comprising 100 linear-threshold units ordered along one spatial dimension. The input (solid line) is composed of 3 Gaussians. The outputs are shown for two cases: $w_{e_j} = 1, w_{i_j} = 1$ and $w_{l_j} = 0.0250 \ \forall j$ (dashed line); $w_{e_j} = 1, w_{i_j} = 1$ and $w_{l_j} = 0.0325 \ \forall j$ (dotted line).

Numerical Simulations It is not possible to obtain a closed form solution for networks with linear threshold units and any arbitrary input distribution, or networks with arbitrary transfer functions, however numerical simulations are useful for providing insight into the general computational properties of the network.

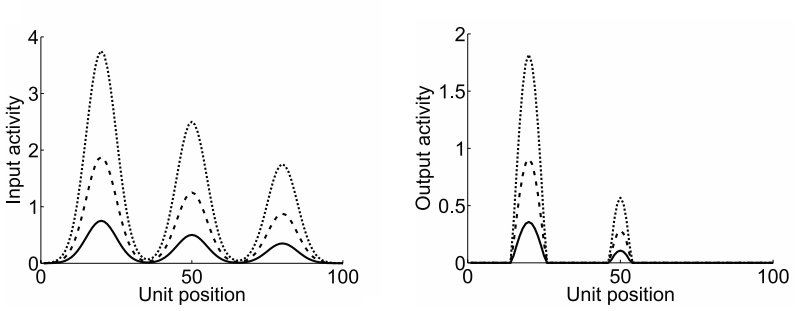


Figure 6.7
Numerical simulation the same WTA network shown in Fig. 6.6 now with weight values $w_{e_j} = 1, w_{i_j} = 1$ and $w_{l_j} = 0.0275$. (a) is the input distribution of increasing amplitude. (b) Network responses to the three inputs shown in (a).

For example, the simulations shown in Figs. 6.6 and 6.7 explore the response of a network with $f(x) = \max(0, x)$ and N=100 to a more complicated input distribution, consisting of three Gaussians centered at unit positions 20, 50, and 80, and having maximum values of 0.75, 0.5, and 0.35 respectively (see solid line of Fig. 6.6). The simulations of Fig. 6.6 show the effect of modifying the excitatory to inhibitory weights w_{l_j} (with all other weights set to one). When $w_{l_j} = 0.0250\ \forall j$, the output is a thresholded version of the input, consisting of 3 peaks of activity. However, when w_{l_j} is increased to 0.0325 $\forall j$, only the strongest input peak is reflected in the output.

In the simulations of Fig. 6.7(a), the excitatory to inhibitory weights w_{l_j} are set to an intermediate value of to 0.0275 $\forall j$, and the network responds to the two strongest peaks in the input. The form of the response is invariant to the input strength (or alternatively, the strength of the w_{e_j} weights) as shown in Fig. 6.7.

Non-linear Programming Formulation

The competitive mechanism that emerges from the neural architecture of Fig. 6.5 can also be described mathematically. The following set of non-linear equations select the largest number among N real numbers by multiplying the neuron output signals y_{e_j} by binary-valued constants (α_j, either 0 or 1):

$$\min\left(-\sum_{j=1}^{N}\alpha_j y_{e_j}\right) \text{ with constraint:}$$

$$\sum_{j=1}^{N}\alpha_j = 1 \quad (\alpha_j \in \{0,1\}). \tag{6.3.12}$$

Systems of non-linear equations with discrete constraints are difficult to solve. We can simplify the system if we extend the domain of α_j to the continuous interval $[0,1]$ and include an additional constraint that forces the continuous values of α_j to tend either toward zero or one. :

$$\min\left(-\sum_{j}\alpha_j y_{e_j}\right) \text{with constraints:}$$

$$\sum_{j}\alpha_j = 1$$

$$\sum_{j}\alpha_j \ln\ \alpha_j = 0. \tag{6.3.13}$$

These types of systems can be solved using the Lagrange multipliers method (Bertsekas, 1982). Solving Eq. 6.3.13 is equivalent to finding $\min_{\alpha_j}(L)$ and $\max_{\lambda_1,\lambda_2}(L)$, where λ_1 and λ_2 are parameters called *Lagrange multipliers*, and L is the cost function

$$L = -\sum_{j=1}^{N}\alpha_j y_{e_j} + \lambda_1\left(\sum_{j=1}^{N}\alpha_j - 1\right) + \lambda_2\sum_{j=1}^{N}\alpha_j \ln\ \alpha_j. \tag{6.3.14}$$

If we set λ_2 to a constant, then we can find $\min_{\alpha_j}(L)$ and $\max_{\lambda_1}(L)$ by solving:

$$\begin{aligned}\frac{\partial}{\partial \alpha_j} L &= -y_{e_j} + \lambda_1 + \lambda_2(\ln\ \alpha_j - 1) = 0 \\ \frac{\partial}{\partial \lambda_1} L &= \sum_j \alpha_j - 1 = 0\end{aligned} \tag{6.3.15}$$

which implies

$$\begin{aligned}\alpha_j &= e^{\frac{y_{e_j} - \lambda_1 - \lambda_2}{\lambda_2}} \\ e^{\frac{\lambda_1 + \lambda_2}{\lambda_2}} &= \sum_j e^{y_{e_j}/\lambda_2}\end{aligned} \tag{6.3.16}$$

and

$$\alpha_j = \frac{e^{y_{e_j}/\lambda_2}}{\sum_k e^{y_{e_k}/\lambda_2}}. \tag{6.3.17}$$

This equation approaches the solution of Eq. 6.3.13 when λ_2 is sufficiently small.

Systems of constrained non-linear equations can be implemented using MOS transistors in the subthreshold domain, The response of the circuit is analogous to the solution of Eq. 6.3.16. Figure 6.8 shows an example of such a circuit, which solves the system of equations (Eqs. 6.3.13). The solution can be confirmed by applying Kirchhoff's current law at node V_c of the circuit ($I_b = \sum_j I_{out_j}$), and observing that the output current of each branch of the WTA network can be expressed as a fraction α_j of the total bias current:

$$I_{out_j} = \alpha_j\ I_b = \alpha_j \sum_{k=1}^{N} I_{out_k}. \tag{6.3.18}$$

Because I_{out_j} can also be written as

$$I_{out_j} = I_0 e^{\kappa \frac{V_{d_j}}{U_T} - \frac{V_c}{U_T}} \tag{6.3.19}$$

the circuit's response and the system of equations (Eqs. 6.3.16) are equivalent, if

$$\begin{aligned}I_{out_j} &= e^{\frac{y_{e_j}}{\lambda_2}} \\ I_b &= e^{\frac{\lambda_1 + \lambda_2}{\lambda_2}}\end{aligned}$$

and

$$y_{e_j} = \lambda_2 \left(\kappa \frac{V_{d_j}}{U_T} - \frac{V_c}{U_T} + \ln(I_0) \right)$$
$$\lambda_1 = \lambda_2 \left(\ln(I_b) - 1 \right). \tag{6.3.20}$$

Current Mode WTA Circuit

The circuit of Fig. 6.8 is a continuous time, analog circuit that implements a WTA network. It was originally designed by Lazzaro et al. (1989) and is extensively used in a wide variety of applications. The circuit is extremely compact and elegant: It processes all the (continuous-time) input signals in parallel, using only two transistors per input cell, and one global transistor that is common to all cells. Collective computation and global connectivity is obtained using one single node common to all cells.

An example of a WTA circuit containing only 2 cells is shown in Fig. 6.8. Each cell comprises a current-controlled conveyor (Section 6.1). The WTA network is modular and can be extended to N cells, by connecting additional cells to the node V_c. Input currents are applied to the network through current sources which are implemented for example using subthreshold pFETs. The output signals are encoded both by the I_{out_1} and I_{out_2} currents, and the V_{d_1} and V_{d_2} voltages. The voltage V_b sets the bias current I_b. Transistors M_1 and M_2 discharge nodes V_d and so implement inhibitory feedback. Transistors M_3 and M_4 implement an excitatory feedforward path by charging node V_c. The overall circuit selects the largest input current I_{in_j} because cell j provides $I_{out_j} \approx I_b$, and so suppresses all other output voltages and currents ($V_{d_i \neq j} \approx 0$, $I_{out_i \neq j} \approx 0$). Cell j wins the competition because its voltage V_{d_j} determines V_c by the exponential characteristics of the transistor that sinks the output current I_{out_j} (for example, M_3 or M_4).

We will analyze the behavior of the circuit in the steady-state case using the methods that we applied for the network model: By providing constant input signals and measuring the outputs after the circuit has settled. We consider three cases: Both inputs are equal; one input much larger than the other; and two inputs that differ by a very small amount (small-signal regime).

Both Inputs Equal If the two input currents are equal ($I_{in_1} = I_{in_2} = I_m$) then the currents flowing into transistors M_1 and M_2 of Fig. 6.8 are also

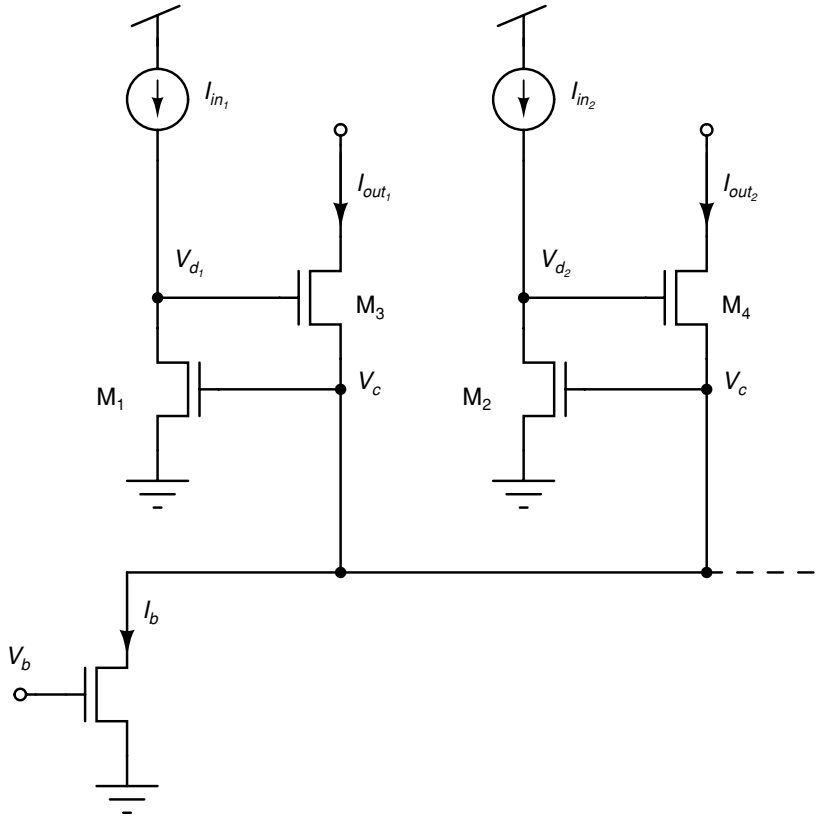


Figure 6.8
Two cells of a current mode WTA circuit.

equal. In this case, because the gates of M_1 and M_2 are tied to the same common node V_c, the drain voltages of M_1 and M_2 must take the same value ($V_{d_1} = V_{d_2} = V_m$). As a result, the output transistors M_3 and M_4 will have the same gate-to-source voltage difference ($V_{gs_3} = V_{gs_4} = V_m - V_c$). If both output transistors are in saturation then the output currents must be identical. Moreover, Kirchhoff's current law requires that, at the common node V_c, $I_{out_1} = I_{out_2} = I_b/2$ (Eq. 6.2.2).

One Input Much Greater than The Other Recall that the subthreshold current flowing through a transistor can be divided into a *forward* component, I_f,

and a *reverse* component, I_r, (Eq. 3.2.11). When the transistor's source voltage V_s is approximately equal to its drain voltage V_d, I_r becomes comparable to I_f.

With this property in mind, we can consider the case in which $I_{in_1} \gg I_{in_2}$. In this case, the drain voltage of M_1 (V_{d_1}) will be greater than the drain voltage of M_2 (V_{d_2}). If the transistor M_1 is in saturation ($V_{d_1} > 4U_T$), the dominant component of its drain current will be in the forward direction and its gate voltage V_c will increase such that $I_{d_1} = I_{f_1} = I_0 e^{\kappa \frac{V_c}{U_T}} = I_{in_1}$. Although the two input currents I_{in_1} and I_{in_2} are different, the forward component of the drain currents of M_1 and M_2 are equal ($I_{f_1} = I_{f_2}$) because the two transistors have a common gate voltage V_c, and both their sources are tied to ground. The drain current I_{d_2} of transistor M_2 can only be equal to the input current I_{in_2} under the following conditions:

$$I_{f_2} - I_{r_2} = I_{in_2}$$

which implies that

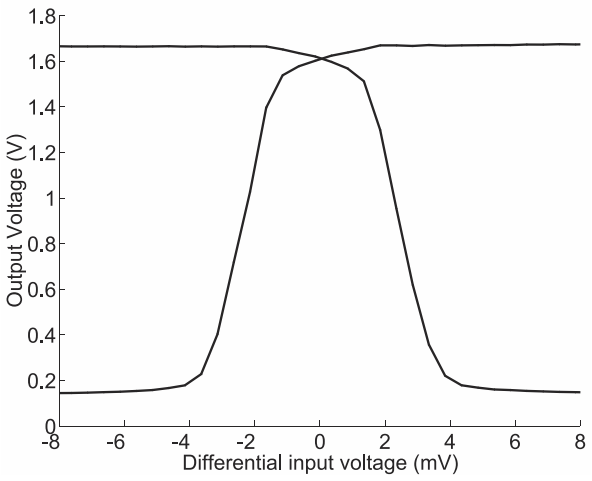
$$I_{r_2} = I_{f_2} - I_{in_2}$$

which implies that

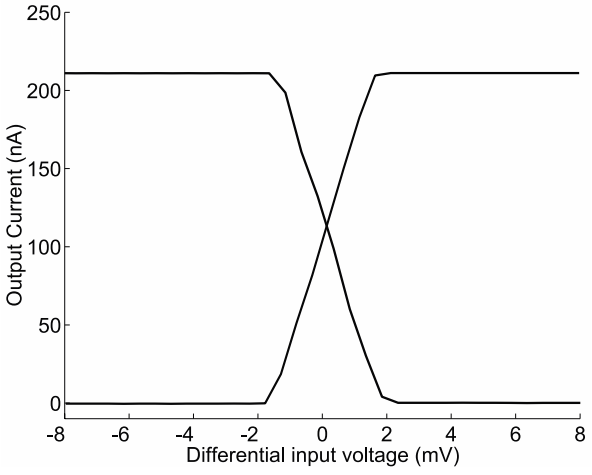
$$I_{r_2} = I_{in_1} - I_{in_2} \gg 0.$$

The reverse component of I_{d_2} becomes significant only if V_{d_2} decreases enough for M_2 to operate in its ohmic region ($V_{d_2} \leq 4U_T$). In this case, the output transistor M_4 is effectively switched off, and $I_{out_2} = 0$. Consequently, M_3 sources all the bias current ($I_{out_1} = I_b$), with V_{d_1} satisfying the equation $I_0 e^{\kappa V_{d_1} - V_c} = I_b$.

The experimental data of Fig. 6.9 shows the output voltages ($V_{d,1}$ and $V_{d,2}$) and output currents ($I_{out,1}$ and $I_{out,2}$) of the circuit, in response to the differential input voltage ΔV which encodes the ratio of the input currents. In this experiment, the input currents were provided by pFETs operating in the subthreshold regime: The gate voltage V_{in_1} of the pFET sourcing current into the first cell was set to 4.3V, while the gate voltage V_{in_2} of the pFET sourcing current into the second cell was set to $V_{in_2} = V_{in_1} + \Delta V$. The two traces in each plot show the responses of the two cells as ΔV was swept from -8mV to +8mV. When ΔV is zero (the input currents are identical), the output signals of both cells are also identical. When ΔV is large (one input current dominates), a single cell is selected.



(a)



(b)

Figure 6.9
Responses of the two-cell WTA circuit shown in Fig. 6.8. (a) Voltage output (V_{d_1} and V_{d_2}) versus the differential input voltage. (b) Current output (I_{out_1} and I_{out_2}). The bias voltage V_b = 0.7V. The small difference in the maximum output currents is due to device mismatch effects in the read-out transistors of the two cells.

If ΔV is small, the above description is not adequate. Instead, we can compute the outputs signals of the cells using *small-signal* analysis (Lazzaro, 1990).

Two Inputs Differ by a Small Amount To analyze the circuit in this regime, we must consider the Early effect of the transistor operating in the saturation region (Eq 3.5.3):

$$I_{ds} = I_{sat}(1 + \frac{V_{ds}}{V_e}) \tag{6.3.21}$$

where V_e is the Early voltage.

Assume that the two input currents I_{in_1} and I_{in_2} are initially equal. In this case, the transistors M_1 and M_2 operate in the saturation region: The output voltages V_{d_1} and V_{d_2} will settle to a common value, and the output currents I_{out_1} and I_{out_2} will both be equal to $I_b/2$. If we now increase the input current I_{in_1} by a small amount δ_I and apply Eq. 3.5.3 to transistor M_1 of Fig. 6.8, then its drain voltage V_{d_1} will increase by

$$\delta_V = \frac{\delta_I}{I_{sat}} V_e. \tag{6.3.22}$$

As V_{d_1} is also the *gate* voltage of transistor M_3, the $I_{out,1}$ will be amplified by an amount proportional to e^{δ_V}. The constraint of Eq. 6.2.2 requires that I_{out_2} decrease by the same amount in steady state. This reduction means the gate voltage V_{d_2} of M_4 must decrease by δ_V.

The gain of the competition mechanism ($\frac{\delta_V}{\delta_I}$) in the small signal regime is directly proportional to the Early voltage V_e and inversely proportional to I_{sat}. The Early voltage depends on the geometry of the transistors and is fixed at design time. On the other hand I_{sat} depends on V_c, which changes with the amplitude of the input currents.

6.4 Resistive Networks

Conventional methods of implementing resistors in VLSI technology include using single MOSFETs as described in Section 5.1; or using more complex circuits such as the transconductance amplifier of Section 5.3. These methods have the disadvantage of emulating linear resistors for only a very limited range of voltages, and of resistance values. If we consider currents, and not

voltages, to representi input and output signals of MOSFETs, then we can implement resistive networks using *single* transistors instead of resistors. In this configuration, the transistor is linear for a wide range of current values. Furthermore, if the transistor is operated in the subthreshold regime, then the resistance (or conductance) can be varied by changing its gate voltage.

A conventional conductance, G, is defined by the relationship

$$I_{ab} = G\,(V_a - V_b)$$

where I_{ab} is the current flowing from terminal a to terminal b, and V_a, V_b the voltages at the corresponding terminals. If the two terminals a and b are the source and the drain of a subthreshold nFET, the current I_{ab} can be expressed by the usual transistor relationship:

$$I_{ab} = I_0 e^{\kappa \frac{V_g}{U_T} - \frac{V_a}{U_T}} - I_0 e^{\kappa \frac{V_g}{U_T} - \frac{V_b}{U_T}} \tag{6.4.1}$$

where V_g is the transistor's gate voltage.

If we define the *pseudo-voltage* (Vittoz and Arreguit, 1993) $V^* = V_0 e^{-\frac{V}{U_T}}$ (where V_0 is an arbitrary scaling voltage), and the *pseudo-conductance* $G^* = \frac{I_0}{V_0} e^{\kappa \frac{V_g}{U_T}}$, then we can write

$$I_{ab} = G^*\,(V_a^* - V_b^*) \tag{6.4.2}$$

where the value of pseudo-conductance G^*depends exponentially on the transistor's gate voltage V_g. Using Eq. 6.4.2 we can map any resistive network into an equivalent transistor network: Each resistor R_i of the resistive network can be replaced by a single transistor M_i, provided that all the transistors share the same substrate (that is, they are all either nFETs or pFETs). If the gate voltages V_{g_i} of all the transistors are equal, then the transistor network is linear with respect to current (Vittoz, 1994). This linear behavior holds for the entire range of weak inversion, which may be as much as 6 orders of magnitude in transistor current. Because all V_{g_i} must be the same, the values of the individual conductances can only be adjusted by changing the W/L ratio (which modulates I_0) of each transistor.

An alternative intrepretation of the mapping between resistive and transistor networks uses the concept of a *current diffusor* (Boahen, 1997) illustrated in Fig. 6.10.

The currents I_{in_1} and I_{in_2} are inputs to the circuit. Assuming that the three nFETs are identical (that is, their I_0 and κ parameters are equal), and solving

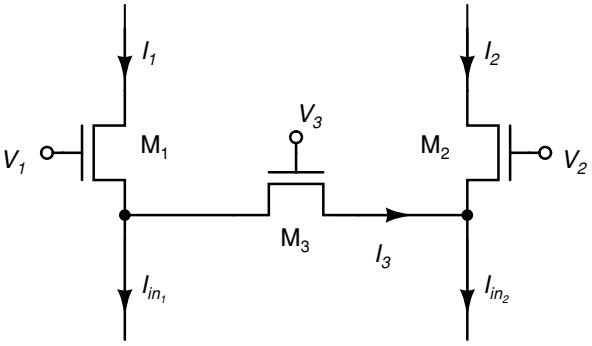


Figure 6.10
Current diffusor circuit. The current I_3, proportional to $(I_2 - I_1)$, diffuses from the source to the drain of M_3.

the circuit equations, we obtain:

$$I_3 = I_0 e^{\kappa \frac{V_3}{U_T}} \left(\frac{I_2}{I_0 e^{\kappa \frac{V_2}{U_T}}} - \frac{I_1}{I_0 e^{\kappa \frac{V_1}{U_T}}} \right) . \tag{6.4.3}$$

If $V_1 = V_2 = V_{ref}$, this relationship can be simplified to yield

$$I_3 = e^{\frac{\kappa}{U_T}(V_3 - V_{ref})} (I_2 - I_1) . \tag{6.4.4}$$

The diffusion current I_3 through M_3 is proportional to $(I_2 - I_1)$. The proportionality factor can be modulated by either V_{ref} or V_3. The current-mode diffusor network (Fig. 6.11(a)) is composed of multiple instances of the circuit of Fig. 6.10. In this network, current injected at a node j diffuses laterally and decays with distance (Mead, 1989). Consequently the network acts as spatial low-pass filter (see Section 8.1); and because the network is linear, the effects of currents injected at different nodes superimpose.

The diffusor network (Fig. 6.11(a)) has the same network response as the resistive network in Fig. 6.11(b). This equivalence can be demonstrated by comparing the transfer functions of the two circuits. Applying Kirchhoff's current law at node V_j of Fig. 6.11(a):

$$I_{out_j} - (I_j - I_{j-1}) = I_{in_j}. \tag{6.4.5}$$

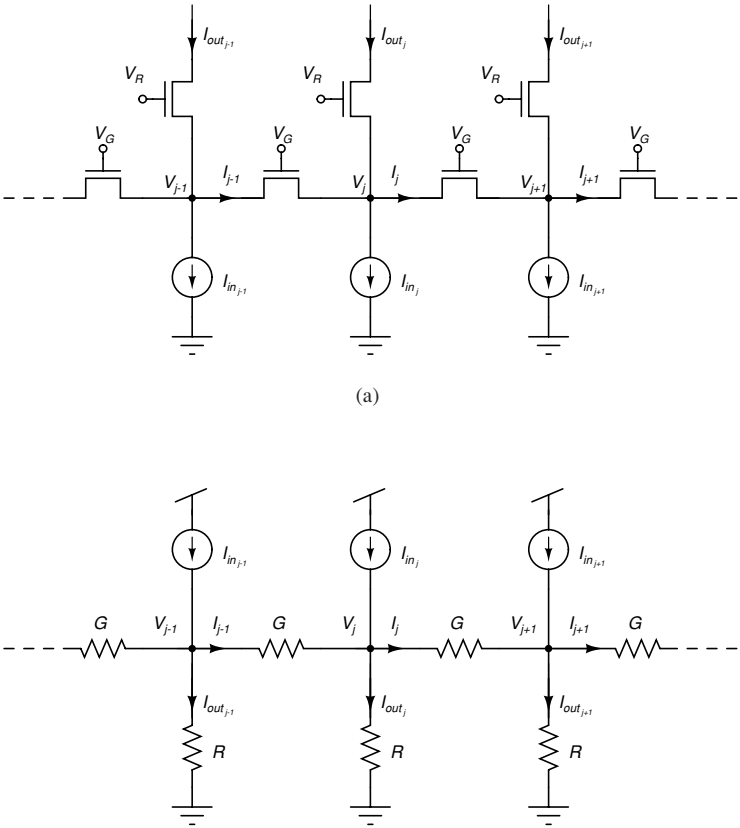


Figure 6.11
Similarities between (a) current-mode diffusor network, and (b) resistive network.

Using Eq. 6.4.4, we can express I_j and I_{j-1} in terms of the output currents:

$$I_{j-1} = e^{\frac{\kappa}{U_T}(V_G - V_R)}(I_{out_j} - I_{out_{j-1}}) \tag{6.4.6}$$

$$I_j = e^{\frac{\kappa}{U_T}(V_G - V_R)}(I_{out_{j+1}} - I_{out_j}). \tag{6.4.7}$$

Substituting these two relationships in Eq. 6.4.5 yields

$$I_{out_j} - I_{in_j} = e^{\frac{\kappa}{U_T}(V_G - V_R)}(I_{out_{j+1}} - 2I_{out_j} + I_{out_{j-1}}) \tag{6.4.8}$$

Similarly, we can apply Kirchhoff's current law at node V_j of Fig. 6.11(b):

$$I_{out_j} - (I_{j-1} - I_j) = I_{in_j}. \tag{6.4.9}$$

Because $I_j = G(V_j - V_{j+1})$ and $I_{out_j} = V_j/R$, I_j can be expressed as a function of I_{out_j} and of $I_{out_{j+1}}$:

$$I_j = \frac{1}{RG}(I_{out_j} - I_{out_{j+1}}). \tag{6.4.10}$$

Combining this equation with Eq. 6.4.9 yields

$$I_{out_j} - I_{in_j} = \frac{1}{RG}(I_{out_{j+1}} - 2I_{out_j} + I_{out_{j-1}}). \tag{6.4.11}$$

The term $(I_{out_{j+1}} - 2I_{out_j} + I_{out_{j-1}})$ in Eq. 6.4.11 is the discrete approximation of the $\frac{d^2}{dx^2}$ operator. Both circuits of Fig. 6.11 approximate the diffusion equation that characterizes the properties of a continuous resistive sheet (Mead, 1989):

$$\lambda^2 \frac{d^2}{dx^2} V_{out}(x) = V_{out}(x) - V_{in}(x) \tag{6.4.12}$$

where λ is the *diffusion length*. In the discrete resistive network of Fig. 6.11(b) the diffusion length $\lambda = 1/\sqrt{RG}$, while in the diffusor network of Fig. 6.11(a) the diffusion length is $\lambda = e^{\frac{\kappa}{2U_T}(V_G - V_R)}$.

6.5 Current Correlator and Bump Circuit

The current correlator measures the correlation between unidirectional input currents, whereas the bump circuit measures the similarity or dissimilarity of input voltages. Both of these circuits have been used in many aVLSI designs. For example, these circuits have been used in a stereoscopic vision system (Mahowald, 1994) to disambiguate between real and false targets, and in an auto-focusing system to measure the input signal power (Delbrück, 1989).

Simple Current Correlator

Mead recognized that in subthreshold operation, the *current-correlator* circuit in Fig. 6.12 computes a measure of the correlation between its two input

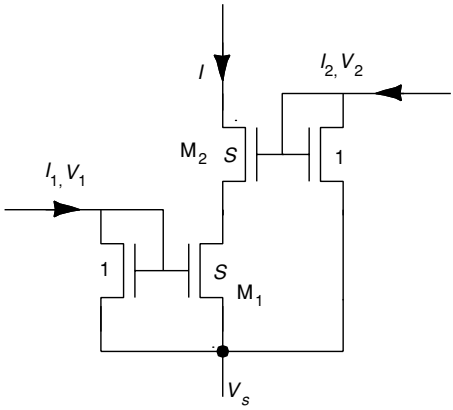


Figure 6.12
The simple current-correlator. S is the strength ratio (see text) between the transistors in the middle leg, and the transistors in the outer legs. I is large only when both I_1 and I_2 are large.

currents I_1 and I_2. Intuitively, the series-connected transistors perform an analog AND-like computation. If either of the gate voltages on these series-connected transistors is low, then the output current is shut off. Conversely, if both of the input voltages are high, then the output current is large. In the intermediate regions, the circuit computes an approximation to the product of the input currents.

We will analyze the simple current-correlator for subthreshold transistor operation. We refer to the effective W/L ratio for a transistor as the *strength* of the transistor. For circuit configurations like the simple current-correlator, the *strength ratio* is the ratio of the strength of the transistors in the middle leg, to the strength of the transistors in the outer legs. This parameter is given by

$$S = \frac{(W/L)_{\mathrm{middle}}}{(W/L)_{\mathrm{outer}}} \tag{6.5.1}$$

and is an important circuit parameter for both the simple current-correlator and the bump circuit.

To compute the output current I, we assume that the top transistor M_2 in Fig. 6.12 is saturated, and that the currents through M_1 and M_2 are identical.

Using the subthreshold equation, we obtain

$$\begin{aligned} I &= Se^{-V_s}\frac{e^{V_1}e^{V_2}}{e^{V_1}+e^{V_2}} \\ &= S\frac{I_1 I_2}{I_1+I_2}. \end{aligned} \tag{6.5.2}$$

While the transistors are operating in subthreshold, the circuit operation is symmetric in the two input currents, despite the apparent asymmetry in the stacking order of the transistors M_1 and M_2. Above threshold, the function is more complicated and is no longer symmetric in the input currents.

This simple current-correlator circuit computes a *self-normalized* correlation. The output current is proportional to the product of the two input currents, divided by the sum of the inputs.

We can extend the simple current-correlator to more than a pair of inputs. The output current for n input currents (a stack of n series-connected transistors) is

$$\frac{1}{I_{out}} = \sum_{k=1}^{n}\frac{1}{I_k}. \tag{6.5.3}$$

The n-input current correlator computes the parallel combination of the n input currents. The maximum number of inputs can be large, because the only requirement for correct circuit operation is that the top transistor in the correlator be saturated. However, the output current scales as $1/n$.

Bump-Antibump Circuit

Figure 6.13 (a) shows the *bump-antibump* circuit. It has three outputs (Fig. 6.13 (b)): I_1, I_2, and I_{mid}. Output I_{mid} is the bump output. Outputs I_1 and I_2 behave like rectifier outputs, becoming large only when the corresponding input is sufficiently larger than the other input. If I_1 and I_2 are combined, they form the *antibump* output, which is the complement of the bump output.

Intuitively, we can understand the operation of this circuit as follows: The three currents must sum to the bias current I_b; hence, the voltage V_c follows the higher of V_1 or V_2. The series-connected transistors M_1 and M_2 form the core of the same analog current correlator that is used in the current-correlator. When $\Delta V = 0$, current flows through all three legs of the circuit. When $|\Delta V|$ increases, the common-node voltage V_c begins to follow the higher of V_1 or V_2. This action shuts off I_{mid}, because the transistor whose gate is connected to the lower of V_1 or V_2 shuts off. Both V_1 and V_2 can rise together and I_{mid}

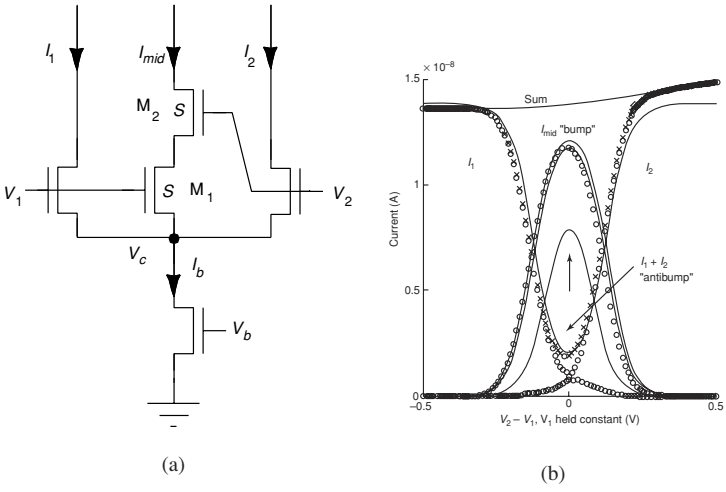


Figure 6.13
(a) The bump-antibump circuit. (b) Output characteristics of bump-antibump circuit. The plots show data points together with theoretical fits of the form given in the text. The curve pointed to by the arrow shows the fit that would result from using S=5.33 derived from the drawn layout geometry, before any process correction. The two theoretical curves shown for I_{mid} are the result of computing the best numerical fit to the entire curve (S=22.4), and using the ratio of the measured maximum to minimum current I_1 + I_2 (S=28). The two numerically fit curves are nearly indistinguishable, and both are very different from the theoretical curve derived from the layout geometry. The width and length reduction parameters from this process run were of the order of 0.5 μm. Using these parameters, we compute S= 6.6, which is still far short of the observed behavior. The curve labeled Sum is I_1 + I_2 + I_{mid}. The slope on the Sum curve is due to the drain conductance of the bias transistor.

does not increase, because the common-node voltage V_c increases along with V_1 and V_2.

Using the subthreshold transistor equation; the input-output relation for the simple current-correlator (Eq. 6.5.2); and Kirchhoff's current law applied to the common node ($I_b = I_1 + I_2 + I_{mid}$) we can compute the current I_{mid}[2]

$$I_{mid} = \frac{I_b}{1 + \frac{4}{S}\cosh^2 \frac{\kappa \Delta V}{2}} \ . \tag{6.5.4}$$

2 $\cosh(x) = \frac{e^x + e^{-x}}{2}$. $\mathrm{sech}(x) = \frac{1}{\cosh(x)}$.

Of course, the antibump outputs sum together to make the current

$$I_1 + I_2 = I_b - I_{mid} = \frac{I_b}{\frac{S}{4}\mathrm{sech}^2\frac{\kappa\Delta V}{2} + 1}. \tag{6.5.5}$$

When $\Delta V = 0$, the $\cosh^2$ term is 1 and the bump current becomes

$$I_{mid} = \frac{I_b}{1 + \frac{4}{S}} \tag{6.5.6}$$

while the antibump outputs sum to

$$I_1 + I_2 = \frac{I_b}{\frac{S}{4} + 1}. \tag{6.5.7}$$

Now we can observe the effect of the transistor strength ratio, S, on the circuit behavior. The width of the bump, measured in input voltage units, depends on this ratio. S controls the fraction of the bias current I_b that is supplied by I_{mid} when $\Delta V = 0$. By examining the denominator of Eq. 6.5.4, we see that the width of the bump scales approximately as $\log(S)$, when $S \geq 1$. If we define the width of the response as the ΔV that makes I_{mid} drop to $1/2$ of the value it takes when $\Delta V = 0$, we obtain

$$\Delta V_{1/2} \approx \frac{2}{\kappa} \log S \tag{6.5.8}$$

in the limit of very large S. The units for ΔV are kT/q.

The dynamic range of the antibump output, $I_1 + I_2$ also depends on S. The antibump output ranges from the minimum value in Eq. 6.5.7, to I_b. The ratio of maximum to minimum output (the dynamic range) is $\frac{S}{4} + 1$. It appears that making S very large gives a very wide dynamic range. This is true, but if S is very large, the bottom of the antibump output characteristic becomes very flat. To make an output characteristic that is maximally parabolic on a linear scale requires a suitable compromise on S. A suitable value for S, and thus the actual transistor dimensions, should be carefully chosen using simulation of circuit performance.

Figure 6.13(b) shows measured operating curves for the bump circuit. The theoretical form for I_{mid} fits the data quite well; there is a discrepancy between the measured and expected value for S. The bump-antibump circuits behaves as though S is much larger than it should be given the layout of the circuit. This effect is fortuitous, because it means the circuit layout need not be as bulky as would be required for very large S. In practice, if oneuses the bump circuit output (I_{mid}), the effective value of S is not particularly relevant unless

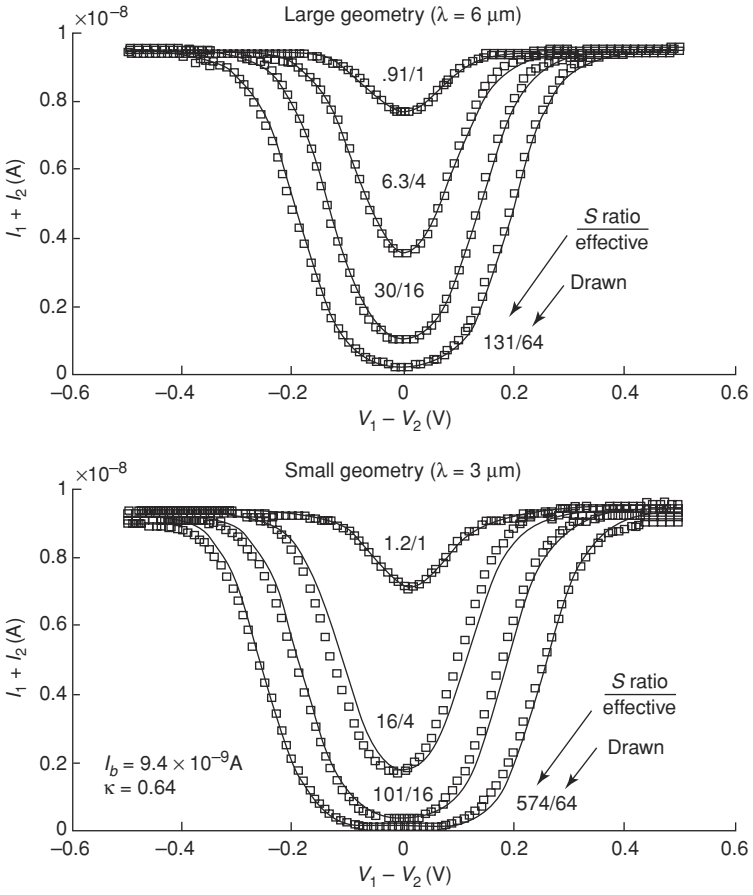


Figure 6.14
Antibump outputs from 8 bump-antibump circuits with different geometries. The solid curves show the theoretical fits derived from Eq. 6.5.5. The numbers beside each curve are the actual and expected S values for the circuit. The different bump circuits in each set of graphs all had transistors of the same width in the outer legs, and transistors of the same length in the middle leg. For the top set of curves: Minimum transistor dimension was 6 μm; correlating transistors had widths 6, 12, 24, and 48 μm; and outer transistors had lengths 6, 12, 24, and 48 μm. For the bottom set of curves, all transistor dimensions were halved. The discrepancy between measured and expected S values are larger for the circuits with smaller dimensions. The curves were fit by minimizing the total squared error, using a single common I_b and κ.

S is really large, because the width of the bump is only weakly dependent on S. On the other hand, if the antibump outputs are used, the output current at zero input difference depends critically S. Since the dynamic range of output

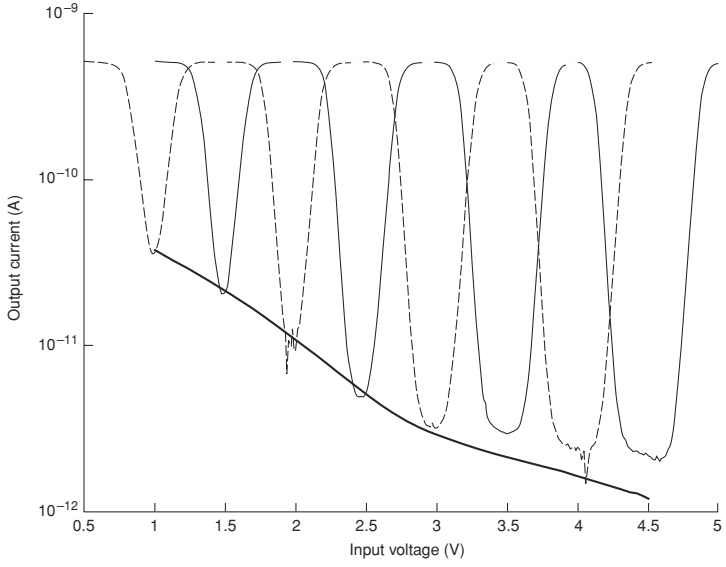


Figure 6.15
The back-gate effect on antibump circuit operation. Each curve shows the antibump output current for a different setting of the V_1 input. The output current is minimum when $V_1 = V_2$. The bottom curve shows the output when $V_1 = V_2$.

depend on S, it is important that S should be correct. Obtaining the correct S value is not easy since S depends on two transistor effects: The *short* and *narrow* channel effects, which arise from the electrostatics of the fields in the channel[3]. These effects make the channel dimensions behave smaller than they are drawn. For logic designs, the channel effects show up primarily as shifts of the threshold voltage: Minimum-length transistors have a lower threshold voltage than longer ones, and minimum-width transistors have a higher threshold voltage than wider ones. The threshold may vary by as much as several hundred millivolts. In subthreshold operation, a threshold voltage decrease of 100 mV is equivalent to a 10-fold increase in I_0. The best method we have found to estimate the actual threshold shift for a given geometry and operating point is to use the most accurate SPICE transistor model available (presently BSIM3v3) to carefully simulate the design. Even then, the models are not very good at predicting the actual value of S. Figure 6.14 shows

3 Sec. 13.1 shows process formation of the bird's beaks and lateral diffusion of the field implant that lead to the narrow channel effect.

measured and calculated S values for a number of different bump circuits built in a MOSIS 2 μm process.

The short and narrow channel effects are affected by back-gate bias. The back-gate voltage acts differentially on the short and narrow channel effects; consequently S changes with operating point. Figure 6.15 shows this effect: this effective S increases with common-mode voltage. This effect can be used to advantage for tuning the response if one has control over the common mode input voltage.

This page intentionally left blank

7 Analysis and Synthesis of Static Translinear Circuits

A Short History of Translinear Circuits In 1975, Barrie Gilbert coined the word *translinear* to describe a class of circuits whose large-signal behavior hinges on the extraordinarily precise exponential current–voltage characteristic of the bipolar transistor; and the intimate thermal contact and close matching of monolithically integrated devices (Gilbert, 1975). The unusual functions performed by these circuits—including multiplication (Paterson, 1963; Gilbert, 1968a,c; Brüggemann, 1970; Schlotzhauer and Viswanathan, 1972; Gilbert, 1974), wideband signal amplification (Gilbert, 1968a,b), and various power-law relationships (Barker and Hart, 1974)—were utterly incomprehensible in terms of linear current amplifiers composed of bipolar transistors. At the same time, Gilbert enunciated a general circuit principle, the *translinear principle* (TLP), whereby the (steady-state) large-signal characteristics of these circuits can be quickly analyzed, usually with only a few lines of algebra, by considering only the currents flowing in the circuits.

The word *translinear* refers to the exponential current–voltage characteristic of the bipolar junction transistor that is central to the functioning of these circuits: The bipolar junction transistor's *trans*conductance is *linear* in its collector current. Gilbert also used the word to refer to analysis and design techniques (for example, the translinear principle) that bridge the gap between the well-established domain of linear-circuit design and the largely uncharted domain of nonlinear-circuit design, for which precious little can be said in general (Gilbert, 1990, 1993, 1996). As we shall see in Sect. 7.3, the translinear principle is a *trans*lation of a *linear* constraint on the voltages in a circuit (Kirchhoff's voltage law), into a product-of-power-law constraint on collector currents flowing in the circuit, using the exponential current–voltage relationship.

Since Gilbert coined the word *translinear*, the translinear principle has been the basis of a plethora of useful nonlinear circuits, including wideband analog multipliers (Gilbert and Holloway, 1980; Huijsing et al., 1982; Gilbert, 1983), translinear current conveyors (Fabre, 1984, 1985; Normand, 1985), translinear frequency multipliers (Ashok, 1976a; Genin and Konn, 1979; Konn and Genin, 1979; Surakampontorn et al., 1988), operational current amplifiers (Fabre, 1986; Fabre and Rochegude, 1987; Fabre, 1988; Toumazou et al., 1989), RMS–DC converters (Gilbert and Counts, 1976; Seevinck et al., 1984; Wassenaar et al., 1988; Frey, 1996a; Mulder et al., 1996, 1997c), and vector-

magnitude circuits (Ashok, 1976b; Gilbert, 1976; Doorenbosch and Goinga, 1976; Wong et al., 1983; Seevinck et al., 1984). In the 1980s, Seevinck (1981, 1988) made significant contributions to translinear-circuit design by developing systematic techniques for the analysis and synthesis of these circuits.

Since the mid-1990s, there has been a growing interest in translinear circuits, primarily because of the development of the class of *dynamic translinear circuits*, which had its origins in 1979 in the work of Adams (1979). Although he does not appear to have made a connection between his own ideas and the growing body of work on translinear circuits, Adams proposed a method of implementing large-signal–linear, continuous-time filters using linear capacitors, constant current sources, and translinear devices. He called this method *log-domain filtering*, because all the filtering occurred on log-compressed voltage state variables using translinear devices. The concept of log-domain filtering remained in obscurity for over a decade, only to be independently rediscovered by Seevinck. In 1990, Seevinck presented a first-order filter, which he dubbed a *companding current-mode integrator* (Seevinck, 1990). Unfortunately, it appears that neither Seevinck nor Adams had a clear idea of how to generalize their ideas to implement filters of higher order. In 1993, encouraged by Adams to pursue the idea of log-domain filtering, Doug Frey introduced a general method for synthesizing log-domain filters of arbitrary order using a state-space approach; and he presented a highly modular technique for implementing such filters (Frey, 1993).

Jan Mulder *et al*. coined the phrase *dynamic translinear circuits* and made the clearest connection between translinear circuits and log-domain filters (Serdijn et al., 1997b; Mulder et al., 1997b; Serdijn et al., 1999; Mulder et al., 1999). They extended Seevinck's translinear analysis and synthesis methodology to encompass dynamic constraints based on what they called the *dynamic translinear principle*, with which capacitive currents embedded within translinear loops can be expressed directly in terms of products of the currents flowing through translinear devices, and their time derivatives. Dynamic translinear circuit techniques have been successfully applied to the structured design of both linear dynamical systems (e.g., log-domain filters (Frey, 1993, 1996b; Perry and Roberts, 1996; Liu, 1996; Drakakis et al., 1997; Mulder et al., 1997a; Drakakis et al., 1998, 1999; Enz and Punzenberger, 1999)) and nonlinear dynamical systems (for example, RMS-DC converters (Frey, 1996a; Mulder et al., 1996, 1997c), oscillators (Pookaiyaudom and Mahattanakul, 1995; Thanachayanont et al., 1995; Serdijn et al., 1997a, 1998), phase detectors (Payne and Thanachayanont, 1997), and phase-locked

loops (Thanachayanont et al., 1997; Payne et al., 1998)). Although dynamic translinear circuits are beyond the scope of this chapter, the principles that we shall develop in this chapter directly extend to, and form a foundation for, the analysis and synthesis of this emerging class of circuits.

7.1 The Ideal Translinear Element

Figure 7.1(a) shows a circuit symbol for an ideal *translinear element* (TE). This symbol, has a gate, an emitter, and a collector. It is commonly used in power electronics to represent an *insulated-gate bipolar transistor* (IGBT), which is a hybrid bipolar/MOS device that combines the high input impedance of an MOSFET and the larger current-handling capabilities of a power bipolar transistor. Although it may be possible to build translinear circuits with IGBTs, that possibility is *not* what we are presently considering. Instead, we use the circuit symbol shown in Fig. 7.1(a) for two reasons: Firstly, the ideal TE should have the nearly inviolate exponential current–voltage relationship of the bipolar transistor and the infinite input impedance of the MOSFET; the hybrid symbol of Fig. 7.1(a) is highly suggestive of precisely this mixture of bipolar and MOS qualities. Secondly, even though translinear circuits were originally implemented with bipolar transistors, they can also be implemented using subthreshold MOSFETs. By using a symbol for the ideal TE that resembles both types of transistors, we remind ourselves continually of these properties.

We shall assume that the ideal TE produces a collector current I that is exponential in its gate-to-emitter voltage V and is given by

$$I = \lambda I_{\mathrm{s}} e^{\eta V/U_{\mathrm{T}}} \tag{7.1.1}$$

where I_{s} is a pre-exponential scaling current, λ is a dimensionless constant that scales I_{s} proportionally, η is a dimensionless constant that scales the gate-to-emitter voltage V, and U_{T} is the thermal voltage kT/q. To demonstrate that the ideal TE is *translinear* in the sense that its transconductance is linear in its collector current, we can calculate its transconductance by differentiating Eq. 7.1.1 with respect to V:

$$g_m = \frac{\partial I}{\partial V} = \lambda I_{\mathrm{s}} e^{\eta V/U_{\mathrm{T}}} \times \frac{\eta}{U_{\mathrm{T}}} = \frac{\eta I}{U_{\mathrm{T}}}.$$

Figures 7.1(b) through 7.1(e) show five practical circuit implementations of the ideal TE. The first of these TEs is the *pn* junction diode, shown

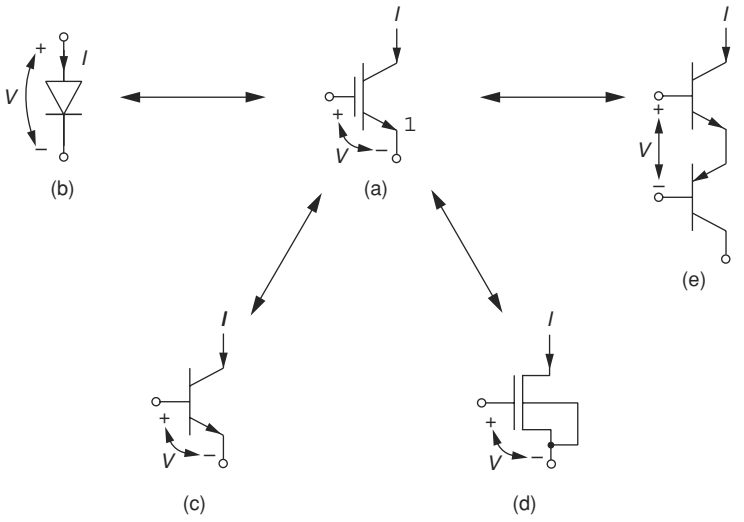


Figure 7.1
Translinear elements (TEs). (a) Circuit symbol for an ideal TE. Such a device produces a current I that is exponential in its controlling voltage V. Parts (b) through (f) show five practical TE implementations comprising (b) a diode, (c) an *npn* bipolar transistor, (d) a subthreshold *n*FET with its source and bulk connected together, and (e) a compound TE comprising an *npn* and a *pnp* with their emitters connected together. Of course, for the TEs shown in parts (c) and (d), the appropriate complementary transistors are also TEs.

in Fig. 7.1(b). Although the forward-biased diode does have an exponential current–voltage characteristic, it is a two-terminal device and does not, strictly speaking, have a transconductance. Moreover, diodes seldom actually appear in translinear circuits. Instead, for the sake of device matching, we almost invariably use diode-connected transistors in place of diodes. Nonetheless, for simplicity, many presentations of the translinear principle begin by considering a loop of diodes. For the diode, λ corresponds to the relative area of the *pn* junction and η is typically close to unity.

Most people consider the bipolar transistor, biased into its forward–active region, to be the quintessential TE (Fig. 7.1(c)). This transistor usually has a precise exponential relationship between its collector current and its base-to-emitter voltage, over more than eight decades of current. For the bipolar transistor, λ corresponds to the relative area of the emitter–base junction and η is typically close to one. The main limitation of the bipolar transistor as a TE is its finite base current, which is often what limits the range of usable current levels in bipolar translinear circuits.

The subthreshold MOSFET with its source and bulk connected together, as shown in Fig. 7.1(d), and biased into saturation, also has an exponential current–voltage characteristic. In this case, λ corresponds to the W/L ratio of the MOSFET and η is equal to κ. In the majority of CMOS technologies, we fabricate one type of MOSFET (either *n*FET or *p*FET) in a global substrate that is maintained at a single common potential. The other type is fabricated inside isolated local wells that may be biased at different potentials. In such technologies, we can only connect the source and bulk together for the type of transistor that is fabricated inside the well. For instance, in an *n*-well CMOS technology, the *p*FETs are fabricated in *n*-wells. By fabricating *p*FETs in separate wells, their individual sources and bulks can be connected together, and different TEs can operate simultaneously with different source potentials. All of the *n*FETs are fabricated inside a global *p*-type substrate. Consequently, if all of the sources and bulks must be shorted together, different TEs cannot have their sources at different potentials. Fortunately, for certain translinear-loop topologies, the source and bulk of all of the MOSFETs within the translinear loops need not be connected together (see Sect. 7.3).

A variety of *compound* TEs can be constructed by combining two or more transistors in various ways. Figure 7.1(e) shows one such compound TE, comprising an *npn* transistor and a *pnp* bipolar transistor with their emitters connected together. For this TE, the controlling voltage is the voltage difference between the base of the *npn* and that of the *pnp*; the output current is available at the collector of either transistor. In this device λ corresponds to the geometric mean between the relative emitter area of the *npn* and the relative emitter area of the *pnp* and $\eta = \frac{1}{2}$.

7.2 Translinear Signal Representations

Translinear circuits are *current-mode* circuits: Their input and output signals are represented as currents. More precisely, we represent a dimensionless quantity z as the ratio of a *signal current*, I_z, to a *unit current*, I_u. In other words, we define a number:

$$z \equiv \frac{I_z}{I_\mathrm{u}}.$$

We call I_u the *unit current* precisely because it is the current level that represents unity in our number system: $z = 1$ if and only if $I_z = I_\mathrm{u}$. The value of the unit current ultimately determines the power dissipation, computational

throughput, and the precision, of a translinear analog information-processing system. By allowing the unit current level to change over time, we can build computing systems that can adaptively trade computational throughput and precision, for power dissipation (Andreou and Boahen, 1994).

The representation of signals by currents in translinear circuits is analogous to a floating-point number representation in the digital domain. For translinear circuits, current signals span a large dynamic range (typically several decades), but noise restricts the precision available at any given signal-current level to a certain finite fraction of that current level. This fraction is characterized by the *noise-to-signal ratio* (NSR). The manner in which the noise current scales with the signal current depends on the type of noise that dominates the system (Sarpeshkar, 1998). If white noise dominates, then NSR scales inversely with the square root of the signal current; so we can trade power consumption for precision. On the other hand, if $1/f$ noise dominates, then the NSR is roughly independent of the signal-current level; so we cannot obtain additional precision by increasing power consumption. In the digital domain, floating-point numbers also span a large dynamic range, but because the mantissa is represented by a finite number of bits, the precision for any given exponent is limited to a constant fraction of two raised to that exponent.

Usually voltages in translinear circuits are considered to be irrelevant. However, these voltage signals are implicitly logarithmic representations of the various input and output signals; and so they are directly analogous to *logarithmic number systems* in the digital domain (Mitchell, 1962; Swartzlander and Alexopoulos, 1975; Edgar and Lee, 1979; Arnold et al., 1990), where numbers are represented internally by their binary logarithms to some finite precision. Many researchers (Hall et al., 1970; Kingsbury and Rayner, 1971; Swartzlander et al., 1983; Lang et al., 1985; LaMaire and Lang, 1986; Vainio and Neuvo, 1986; Taylor et al., 1988; Kurokawa and Mizukoshi, 1991; Lai, 1991; Lai and Wu, 1991; Yu and Lewis, 1991; Lewis, 1995) have proposed using logarithmic number systems in special-purpose digital signal processors for applications such as digital filtering, fast Fourier transform computation, and computer graphics; which typically require a large dynamic range, and in which operations such as multiplication, division, squaring, and square-rooting occur more frequently than do addition and subtraction.

The appeal of such logarithmic number systems is that operations of multiplication and division can be implemented using only fixed-point adders; and the operations of squaring and square-rooting using only shifters. However, in such digital arithmetic units, conversion between conventional number for-

mats and the logarithmic number format is quite expensive, typically involving large lookup tables. Also, whereas multiplication and division are relatively inexpensive in such systems, addition and subtraction can only be approximated and involve additional lookup tables, making them quite cumbersome. By contrast, in translinear analog information-processing systems, conversion between the logarithmic (that is, voltage) signal representation and the linear (that is, current) signal representation is extremely inexpensive, requiring only a single translinear device. Consequently, with translinear circuits, operations like multiplication, division, squaring, and square-rooting are inexpensive *and* the operations of addition and subtraction, which we can implement simply using Kirchhoff's current law on a single wire, are also inexpensive.

Because of the logarithmic relationship that exists between the controlling voltage and the output current of a translinear device, we must ensure that the currents flowing through all translinear devices remain strictly positive at all times. In order to represent both positive and negative quantities by currents, we can follow one of two basic approaches. Firstly, we can add an offset current I_y to a signal current I_z so that their sum $I_z + I_y$ remain positive at all times, as shown in Fig. 7.2(a). In this case, the signal current is a *bidirectional current*. Note that the condition that $I_z + I_y > 0$ implies that $I_z > -I_y$. Thus, while there is no restriction on the magnitude of positive values of I_z, negative values of I_z cannot exceed the magnitude of the offset current I_y. Secondly, we can represent a quantity that can be both positive and negative as a *differential current*: The difference between two strictly positive components, each of which is represented independently as the ratio of a signal current to the unit current, as shown in Fig. 7.2(b). In other words, we represent a number $z \equiv z^+ - z^-$, where

$$z^+ \equiv \frac{I_z^+}{I_\mathrm{u}} > 0 \text{ and } z^- \equiv \frac{I_z^-}{I_\mathrm{u}} > 0.$$

7.3 The Translinear Principle

In this section, we shall derive the *translinear principle* for a loop of ideal TEs and illustrate its use in analyzing translinear circuits. We shall then consider a loop of subthreshold MOSFETs with their bulks all connected to the common substrate potential to determine how the translinear principle is modified by the body effect.

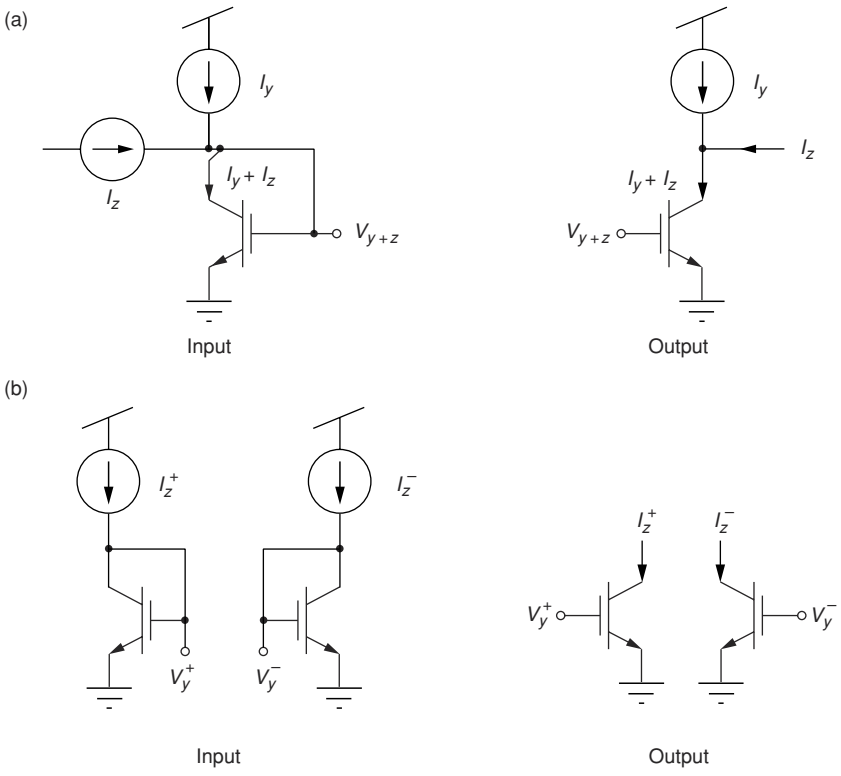


Figure 7.2
Translinear representations for quantities that can take on both positive and negative values. (a) The quantity z is represented by a *bidirectional current* I_z offset by another current I_y, so that $I_z + I_y > 0$. (b) The quantity z is represented by a *differential current* $I_z \equiv I_z^+ - I_z^-$, where $I_z^+ > 0$ and $I_z^- > 0$.

Translinear Loops of Ideal Translinear Elements

Consider the closed loop of N ideal TEs, shown in Fig. 7.3. The large arrow shows the clockwise direction around the loop. If the emitter arrow of a TE points in the clockwise direction, we classify that TE as a *clockwise element*. If the emitter arrow of a TE points in the counterclockwise direction, we classify that TE as a counterclockwise element. We denote by CW the set of clockwise-element indices, and by CCW the set of counterclockwise-element indices.

As we proceed around the loop in the clockwise direction, the gate-to-emitter voltage of a counterclockwise element corresponds to a voltage in-

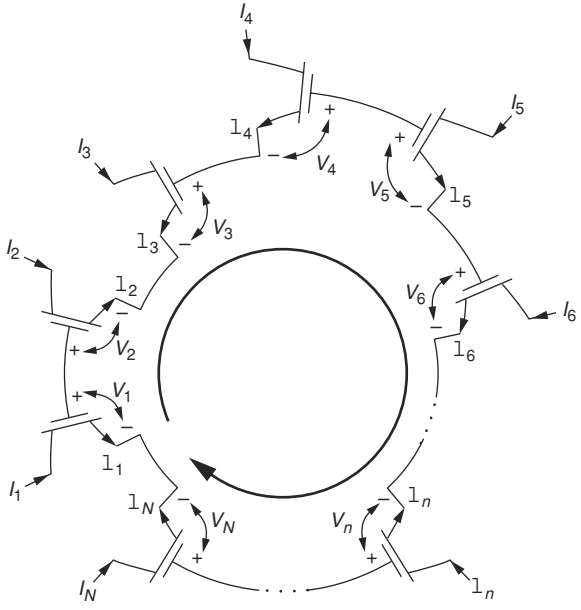


Figure 7.3
A conceptual translinear loop comprising N ideal TEs. The large arrow indicates the clockwise direction around the loop. If a TE symbol's emitter arrow points in the direction opposite to that of the arrow, then we consider the element a *counterclockwise element*. If a TE symbol's emitter arrow points in the same direction as the large arrow, then the element is a *clockwise element*. The translinear principle states that the product of the currents flowing through the clockwise elements is equal to the product of the currents flowing through the counterclockwise elements.

crease, whereas the gate-to-emitter voltage of a clockwise element corresponds to a voltage drop. One way of stating Kirchhoff's voltage law is that the sum of the voltage increases around a closed loop is equal to the sum of the voltage drops around the loop. Consequently, by applying Kirchhoff's voltage law around the loop of TEs shown in Fig. 7.3, we have that

$$\sum_{n \in \mathrm{CCW}} V_n = \sum_{n \in \mathrm{CW}} V_n. \tag{7.3.1}$$

By solving Eq. 7.1.1 for V in terms of I and substituting the resulting expression for each V_n in Eq. 7.3.1, we obtain

$$\sum_{n \in \mathrm{CCW}} \frac{U_\mathrm{T}}{\eta} \log \frac{I_n}{\lambda_n I_\mathrm{s}} = \sum_{n \in \mathrm{CW}} \frac{U_\mathrm{T}}{\eta} \log \frac{I_n}{\lambda_n I_\mathrm{s}}. \tag{7.3.2}$$

Assuming that all TEs are operating at the same temperature, we can cancel the common factor of U_{T}/η in all the terms of Eq. 7.3.2 to obtain

$$\sum_{n\in\mathrm{CCW}} \log \frac{I_n}{\lambda_n I_{\mathrm{s}}} = \sum_{n\in\mathrm{CW}} \log \frac{I_n}{\lambda_n I_{\mathrm{s}}}. \tag{7.3.3}$$

Because $\log x + \log y = \log xy$, Eq. 7.3.3 can be rewritten as

$$\log \prod_{n\in\mathrm{CCW}} \frac{I_n}{\lambda_n I_{\mathrm{s}}} = \log \prod_{n\in\mathrm{CW}} \frac{I_n}{\lambda_n I_{\mathrm{s}}}. \tag{7.3.4}$$

Exponentiating both sides of Eq. 7.3.4 and rearranging yields

$$\prod_{n\in\mathrm{CCW}} \frac{I_n}{\lambda_n} = I_{\mathrm{s}}^{N_{\mathrm{CCW}}-N_{\mathrm{CW}}} \prod_{n\in\mathrm{CW}} \frac{I_n}{\lambda_n} \tag{7.3.5}$$

where N_{CCW} and N_{CW} denote the number of counterclockwise elements, and the number of clockwise elements respectively. It is easy to see that, if $N_{\mathrm{CW}} = N_{\mathrm{CCW}}$, then Eq. 7.3.5 reduces to

$$\prod_{n\in\mathrm{CCW}} \frac{I_n}{\lambda_n} = \prod_{n\in\mathrm{CW}} \frac{I_n}{\lambda_n} \tag{7.3.6}$$

which has no remaining dependence on temperature or device parameters. Equation 7.3.6 is the *translinear principle*, which can be stated as follows:

In a closed loop of ideal TEs comprising an equal number of clockwise and counterclockwise elements, the product of the (relative) current densities flowing through the counterclockwise elements is equal to the product of the (relative) current densities flowing through the clockwise elements.

Further, if each TE in the loop has the same value of λ, then Eq. 7.3.6 reduces to

$$\boxed{\prod_{n\in\mathrm{CCW}} I_n = \prod_{n\in\mathrm{CW}} I_n.} \tag{7.3.7}$$

Equation 7.3.7 is an important special case of the translinear principle that can be stated as follows:

In a closed loop of identical ideal TEs comprising an equal number of clockwise and counterclockwise elements, the product of the currents flowing through the counterclockwise elements is equal to the product of the currents flowing through the clockwise elements.

This derivation of the translinear principle can be characterized as a *trans*lation of a *linear* set of algebraic constraints on the voltages in the circuit (that is, Kirchhoff's voltage law applied around the loop of Fig. 7.3) into a product-of-power-law constraint on the currents flowing in the circuit. This characterization of the translinear principle is one way to state the second sense of the word *translinear* as originally used by Gilbert (1990, 1993, 1996, 1975).

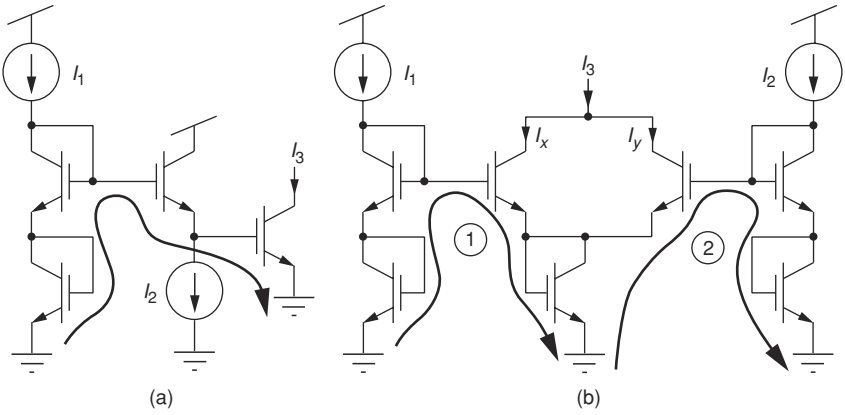


Figure 7.4
Two translinear-loop circuits. (a) A simple circuit with one translinear loop comprising two clockwise elements and two counterclockwise elements arranged in a stacked topology. (b) A circuit with two overlapping translinear loops each of which comprises two clockwise elements and two counterclockwise elements arranged in a stacked topology.

To illustrate the use of the translinear principle, we shall analyze the two translinear-loop circuits shown in Fig. 7.4. First, consider the circuit of Fig. 7.4(a). This circuit has a single translinear loop comprising four identical TEs, two of which face in the clockwise direction and two of which face in the counterclockwise direction. Input current I_1 passes through both counterclockwise elements. Input current I_2 passes through one of the clockwise elements. The output current I_3 passes through the other clockwise element. Consequently, to analyze this circuit, we apply the translinear principle, as stated in Eq. 7.3.7, and write that

$$I_1^2 = I_2 I_3$$

which can be rearranged to obtain

$$I_3 = \frac{I_1^2}{I_2}.$$

Thus, the circuit of Fig. 7.4(a) is a squaring–reciprocal circuit.

Next, consider the circuit shown in Fig. 7.4(b). This circuit has two overlapping translinear loops, each of which comprises four identical TEs with two in the clockwise direction and two in the counterclockwise direction. By Kirchhoff's current law, we have that the output current I_3 is given by

$$I_3 = I_{\mathrm{x}} + I_{\mathrm{y}}. \tag{7.3.8}$$

In the first loop, input current I_1 passes through both counterclockwise elements. Intermediate current I_{x} passes through the first clockwise element and the output current I_3 passes through the second clockwise element. So, by the translinear principle,

$$I_1^2 = I_{\mathrm{x}} I_3$$

which implies that

$$I_{\mathrm{x}} = \frac{I_1^2}{I_3}. \tag{7.3.9}$$

By a similar argument involving the second translinear loop,

$$I_{\mathrm{y}} = \frac{I_2^2}{I_3}. \tag{7.3.10}$$

Substituting Eqs. 7.3.9 and 7.3.10 into Eq. 7.3.8 and solving for I_3:

$$I_3 = \sqrt{I_1^2 + I_2^2}.$$

Thus, the circuit of Fig. 7.4(b) computes the length of a two-dimensional vector. This circuit can be extended to handle an arbitrary number of inputs and could be useful to compute vector magnitudes.

Translinear Loops of Subthreshold MOSFETs

Consider the closed loop of N saturated subthreshold MOSFETs whose bulks are all connected to a common substrate potential, shown in Fig. 7.5(a). Here, V_n represents the gate-to-source voltage of the nth MOSFET, and U_n is the voltage on the nth node relative to the substrate potential. Again, the large

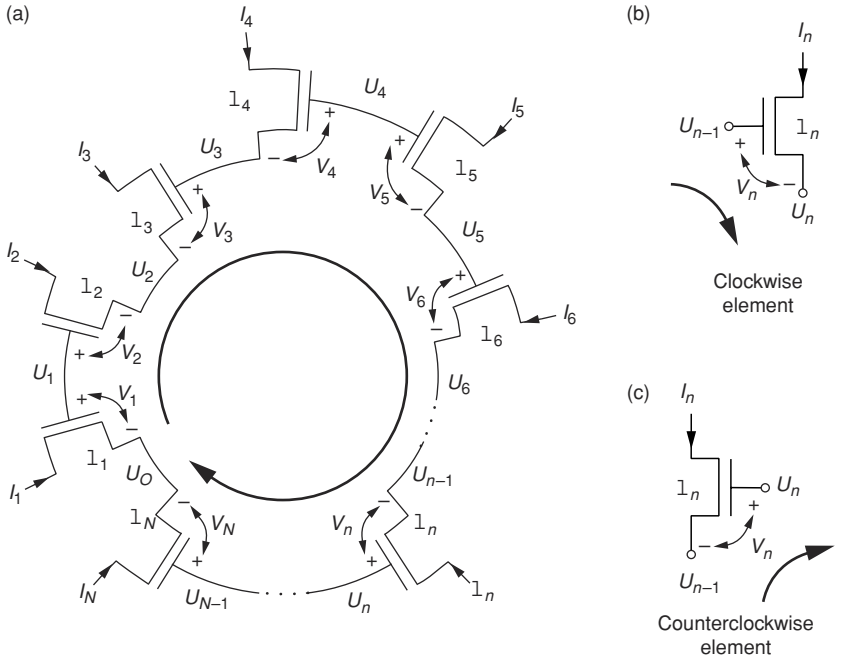


Figure 7.5
A translinear loop of subthreshold MOSFETs with their bulks tied to a common substrate potential. Here V_n refers to the gate-to-source voltage of the nth MOSFET and U_n refers to the voltage on the nth node referenced to the common substrate potential. (a) A conceptual translinear loop comprising N subthreshold MOSFETs with their bulks tied to a common substrate. (b) A clockwise element is one whose gate-to-source voltage is a voltage drop in the clockwise direction around the loop. (c) A counterclockwise element is one whose gate-to-source voltage is a voltage increase in the clockwise direction around the loop.

arrow in Fig. 7.5(a) indicates the clockwise direction around the loop. As shown in Fig. 7.5(b), we shall consider a clockwise element to be one whose gate-to-source voltage is a voltage drop in the clockwise direction around the loop, and we shall consider a counterclockwise element to be one whose gate-to-source voltage is a voltage increase in the clockwise direction around the loop(Fig. 7.5(c)).

Recall that the channel current, of a saturated *n*MOSFET, operating in subthreshold, is given by

$$I = \lambda I_0 e^{(\kappa V_g - V_s)/U_T}.$$

From this equation, if the nth MOSFET is a clockwise element, we have that

$$I_n = \lambda_n I_0 e^{(\kappa U_{n-1} - U_n)/U_T}$$

which can be rearranged to yield

$$e^{U_n/U_T} = \left(e^{U_{n-1}/U_T}\right)^{\kappa} \left(\frac{\lambda_n I_0}{I_n}\right). \tag{7.3.11}$$

Equation 7.3.11 expresses a recurrence relationship between the nth node voltage to the $(n-1)$st node voltage for clockwise elements. On the other hand, if the nth MOSFET is a counterclockwise element, we have that

$$I_n = \lambda_n I_0 e^{(\kappa U_n - U_{n-1})/U_T}$$

which can be rearranged:

$$e^{U_n/U_T} = \left(e^{U_{n-1}/U_T}\right)^{1/\kappa} \left(\frac{I_n}{\lambda_n I_0}\right)^{1/\kappa}. \tag{7.3.12}$$

Equation 7.3.12 likewise expresses a recurrence relationship between the nth node voltage and the $(n-1)$st node voltage for counterclockwise elements.

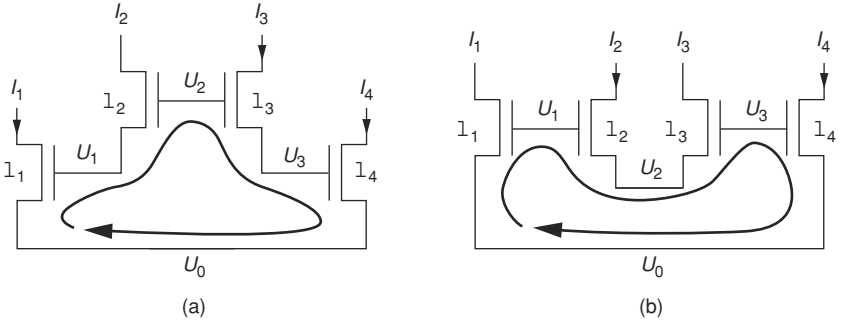


Figure 7.6
Two subthreshold MOS translinear loops comprising two clockwise transistors and two counterclockwise transistors. In each case, the bulks of all four transistors are tied to a common potential. (a) A *stacked* loop topology. (b) An *alternating* loop topology.

We can use the recurrence relationships, expressed in Eq. 7.3.11 and Eq. 7.3.12, to build up the translinear-loop constraint equation for the subthreshold MOS translinear loop (Fig. 7.5(a)), as follows. We begin at one of the nodes in the loop, say U_0, and proceed sequentially around the loop in the

clockwise direction, recursively applying Eq. 7.3.11 or Eq. 7.3.12 to get to the next node, depending on whether the current element is clockwise or counterclockwise. When we encounter a clockwise element, we raise the partially formed translinear-loop equation to the κ power and multiply it by $\lambda_n I_0/I_n$, as expressed in Eq. 7.3.11. When we encounter a counterclockwise element, we raise the partially formed translinear-loop equation to the $1/\kappa$ power and multiply it by $(I_n/\lambda_n I_0)^{1/\kappa}$, as expressed in Eq. 7.3.12. Finally, when we return to the node from which we started, we stop and simplify the resulting expression.

To illustrate this process, we shall consider two simple subthreshold MOS translinear loops, as shown in Fig. 7.6. Each of these loops comprises four transistors, two of which face in the clockwise direction and two of which face in the counterclockwise direction. The first loop, shown in Fig. 7.6(a), has a *stacked* topology: All of the gate-to-source voltage drops are stacked up. The second loop, shown in Fig. 7.6(b) has an *alternating* topology: We alternate between clockwise and counterclockwise elements, as we go around the loop.

First, consider the stacked MOS translinear loop, shown in Fig. 7.6(a). Starting with node U_0 and proceeding around the loop in the clockwise direction, we encounter two counterclockwise elements followed by two clockwise elements before we finish back at node U_0. Following the procedure just described, we have that

$$\Bigg(\Bigg(\Bigg(\underbrace{\underbrace{\underbrace{\left(e^{U_0/U_\mathrm{T}}\right)^{1/\kappa}\left(\frac{I_1}{\lambda_1 I_0}\right)^{1/\kappa}}_{e^{U_1/U_\mathrm{T}}}\Bigg)^{1/\kappa}\left(\frac{I_2}{\lambda_2 I_0}\right)^{1/\kappa}}_{e^{U_2/U_\mathrm{T}}}\Bigg)^{\kappa}\left(\frac{\lambda_3 I_0}{I_3}\right)}_{e^{U_3/U_\mathrm{T}}}\Bigg)^{\kappa} \cdot \left(\frac{\lambda_4 I_0}{I_4}\right) = e^{U_0/U_\mathrm{T}}$$

which can be simplified to obtain

$$\left(e^{U_0/U_\mathrm{T}}\right)^{1/\kappa}\left(\frac{I_1}{\lambda_1 I_0}\right)^{1/\kappa}\left(\frac{I_2}{\lambda_2 I_0}\right)\left(\frac{\lambda_3 I_0}{I_3}\right)\left(\frac{\lambda_4 I_0}{I_4}\right)^{1/\kappa} = \left(e^{U_0/U_\mathrm{T}}\right)^{1/\kappa}.$$

By canceling common factors and grouping the counterclockwise currents on the left-hand side of the equation and grouping the clockwise currents on the

right-hand side of the equation, we obtain the translinear-loop equation:

$$\underbrace{\left(\frac{I_1}{\lambda_1}\right)^{1/\kappa}\left(\frac{I_2}{\lambda_2}\right)}_{\mathrm{CCW}} = \underbrace{\left(\frac{I_3}{\lambda_3}\right)\left(\frac{I_4}{\lambda_4}\right)^{1/\kappa}}_{\mathrm{CW}} \tag{7.3.13}$$

which has no remaining temperature dependence and no dependence on I_0, but does depend on the subthreshold slope factor, κ.

Next, consider the alternating MOS translinear loop, shown in Fig. 7.6(b). Again, starting with node U_0 and proceeding around the loop in the clockwise direction, we first encounter a counterclockwise element followed by a clockwise element, followed by another counterclockwise element, followed by another clockwise element. Following the procedure just described, we have that

$$\underbrace{\left(\underbrace{\left(\underbrace{\left(\underbrace{\left(e^{U_0/U_\mathrm{T}}\right)^{1/\kappa}\left(\frac{I_1}{\lambda_1 I_0}\right)^{1/\kappa}}_{e^{U_1/U_\mathrm{T}}}\right)^{\kappa}\left(\frac{\lambda_2 I_0}{I_2}\right)}_{e^{U_2/U_\mathrm{T}}}\right)^{1/\kappa}\left(\frac{I_3}{\lambda_3 I_0}\right)^{1/\kappa}}_{e^{U_3/U_\mathrm{T}}}\right)^{\kappa}}_{\cdot\left(\frac{\lambda_4 I_0}{I_4}\right) = e^{U_0/U_\mathrm{T}}}$$

which can be simplified to obtain

$$\left(e^{U_0/U_\mathrm{T}}\right)\left(\frac{I_1}{\lambda_1 I_0}\right)\left(\frac{\lambda_2 I_0}{I_2}\right)\left(\frac{I_3}{\lambda_3 I_0}\right)\left(\frac{\lambda_4 I_0}{I_4}\right) = \left(e^{U_0/U_\mathrm{T}}\right).$$

By canceling common factors and grouping the counterclockwise currents on the left-hand side of the equation and grouping the clockwise currents on the right-hand side of the equation, we obtain the translinear-loop equation:

$$\underbrace{\left(\frac{I_1}{\lambda_1}\right)\left(\frac{I_3}{\lambda_3}\right)}_{\mathrm{CCW}} = \underbrace{\left(\frac{I_2}{\lambda_2}\right)\left(\frac{I_4}{\lambda_4}\right)}_{\mathrm{CW}} \tag{7.3.14}$$

which has no remaining temperature dependence, no dependence on I_0, *and* no dependence on κ. From these examples, we can make a number of observations about subthreshold MOS translinear loops. Firstly, if the number of clockwise

elements is equal to the number of counterclockwise elements in a closed loop, then the e^{U_0/U_T} factor will cancel on both sides of the equation. In general, each time a clockwise element is traversed, this factor is raised to the κ power; and each time a counterclockwise element is traversed, this factor is raised by $1/\kappa$. Thus, if there are N_CW clockwise elements and N_CCW counterclockwise elements, the factor e^{U_0/U_T} will appear on the left-hand side of the translinear-loop equation raised to the $\kappa^{N_\mathrm{CW}-N_\mathrm{CCW}}$ power, which is equal to unity if $N_\mathrm{CW} = N_\mathrm{CCW}$. Therefore, this factor, which appears on both sides of the final translinear-loop equation will cancel.

Secondly, if the number of clockwise elements is equal to the number of counterclockwise elements, then the translinear-loop equation will be independent of I_0. To see why, consider first the power to which I_0 is raised after we traverse a clockwise element followed by a counterclockwise element. Suppose that I_0 appears in the initial translinear-loop equation raised to the α power. After traversing a clockwise element, I_0 is raised to the $\kappa\alpha + 1$. Then, after traversing a counterclockwise element, I_0 will be raised to the $(\kappa\alpha + 1)/\kappa - 1/\kappa = \alpha$ power. Thus, the power to which I_0 is raised in the translinear-loop equation remains unchanged when we pass from a sequence of clockwise elements to a sequence of counterclockwise elements. Next, consider what happens to the power to which I_0 is raised after we traverse a counterclockwise element followed by a clockwise element. Again, suppose I_0 appears in the initial translinear-loop equation raised to the α power. After traversing a counterclockwise element, we will find I_0 raised to the $\alpha/\kappa - 1/\kappa$ power. Then, after traversing a clockwise element, we will have I_0 raised to the $\kappa\,(\alpha/\kappa - 1/\kappa) + 1 = \alpha$ power. Thus, the power to which I_0 is raised in the translinear-loop equation also remains unchanged when we pass through a boundary from a sequence of counterclockwise elements to a sequence of clockwise elements.

Now, if there are an equal number of clockwise and counterclockwise elements, then there will be at least two such boundaries; this lower bound is achieved when the loop has a purely stacked topology. If there are $2N$ elements in the loop (that is, $N_\mathrm{CW} = N_\mathrm{CCW} = N$), then there can be at most N such boundaries; this upper bound is achieved when the loop has a purely alternating topology. Because the elements at the boundaries between runs of clockwise and counterclockwise elements do not alter the power to which I_0 is raised in the translinear-loop expression, we can omit the elements at each of these boundaries in determining the overall power to which I_0 is raised. However, the elements adjacent to those that we have omitted then

form another boundary. Consequently, if there is a clockwise element in the loop for each counterclockwise element, then the overall power to which I_0 is raised as we go around the loop remains unaltered. Moreover, when we begin accumulating the translinear-loop expression, I_0 does not appear in the expression (that is, it is raised to the zeroth power). Therefore, because the power to which I_0 is raised begins at zero and it remains unaltered for a loop comprising an equal number of clockwise and counterclockwise elements, the overall translinear loop expression remains independent of I_0. An identical argument holds if all the MOSFETs have the same W/L ratio (that is, $\lambda_1 = \cdots = \lambda_N = \lambda$): If all of the transistors have the same W/L ratio, and if the number of clockwise elements is equal to the number of counterclockwise elements, then the translinear-loop equation is independent of the transistors' common W/L ratio.

Finally, consider a purely alternating subthreshold MOS translinear loop comprising an equal number of clockwise and counterclockwise elements with their bulks connected to a common potential. For such a loop, the translinear-loop equation will be independent of κ. To see why, consider the loop-equation construction procedure. If we begin by traversing a counterclockwise element, the first current factor in the equation will be raised to the $1/\kappa$ power. After traversing the second element, which by hypothesis is a clockwise element, the first current factor will be raised to the κ power, canceling the initial $1/\kappa$. The second current factor will be added to the product raised to the first power. After traversing the third element, the first two current factors will again be raised to the $1/\kappa$ power as will the third current factor. After going through the fourth element, we raise the first three current factors to the κ power, again canceling the $1/\kappa$ power on each factor from the previous step. A fourth current factor is then added to the partially formed product, raised to the first power. Thus, in an alternating loop, we alternate between raising each factor to the $1/\kappa$ power and raising each one to the κ power, effectively removing the κ dependence from the partially formed product after every other step. Moreover, if there are an equal number of clockwise and counterclockwise elements in such a loop, then it follows that the final translinear-loop equation will be independent of κ, because there must be an even number of steps. A similar argument would hold if we begin with a clockwise element. This result has been demonstrated previously using other arguments by Vittoz (1996) and by Andreou and Boahen (1996).

With these considerations in mind, we can simplify our translinear-loop equation construction procedure for loops that comprise an equal number of

clockwise and counterclockwise subthreshold MOSFETs whose bulks are all tied to a common potential. First, we do not have to write down the initial and final e^{U_0/U_T} factors. Instead, we can replace them by unity, effectively dividing out the factor that will be common to both sides of the equation ahead of time. Next, we can dispense with keeping track of all of the I_0 factors, because we have shown that the final expression will be independent of I_0. To summarize this reduced procedure: We pick a starting point in the loop and begin the translinear loop expression with unity. Then, we proceed around the loop from node to node in the clockwise direction. If we traverse a clockwise element, we raise the partially formed expression to the κ power and then we multiply it by λ_n/I_n. If we traverse a counterclockwise element, we raise the partially formed expression to the $1/\kappa$ power and multiply by $(I_n/\lambda_n)^{1/\kappa}$. When we arrive back at the starting node, we equate the product that we have accumulated to unity and simplify it as necessary.

7.4 ABC's of Translinear-Loop–Circuit Synthesis

In this section, we give an overview of the basics of translinear-loop circuit synthesis. As with all synthesis problems, the task of constructing a translinear circuit that implements some desired functionality is underconstrained. For any given function, there will be a variety of circuit solutions with design decisions to be made and many trade-offs to consider. Thus, our brief discussion cannot, by any means, be exhaustive. Rather, we shall attempt to make the basic procedure clear and we shall illustrate it with a simple example. In the process, we shall discuss some design decisions and trade-offs.

Synthesizing Static Translinear-Loop Circuits

The basic procedure for synthesizing translinear-loop circuits can be described as follows: Firstly, we *acquire* a set of translinear-loop equations from the relationship(s) that we want to implement. Secondly, we *build* a translinear loop for each of the translinear-loop equations. Then, we *bias* each of the translinear loops. Finally, if possible, we *consolidate* the resulting circuits by merging some of the loops. We shall discuss each of these steps in turn, and then we shall use them to synthesize a simple two-quadrant translinear-loop multiplier circuit.

Acquiring a Set of Translinear-Loop Equations The starting point for synthesizing a translinear-loop circuit is a static linear or nonlinear mapping between dimensionless variables. The class of translinear circuits is capable of embodying a wide range of useful linear and nonlinear relationships. However, not all functions are directly implementable by translinear circuits; we can directly realize products, quotients, power-law relationships, polynomials, rational functions, and various combinations of such relationships. In many cases, we may need to find an acceptable approximation for one or more nonlinear functions in terms of polynomials, rational functions, continued fractions, or some other suitable mathematical form before we can realize the required relationship with translinear circuits. In his book on translinear circuits, Seevinck (1988) provides a good discussion of suitable approximation techniques and approximations of various transcendental functions.

Once we have obtained a set of relationships that can be realized with translinear circuits, we then represent each of the dimensionless quantities as the ratio of a signal current to the unit current or as the ratio of a differential current to the unit current, as described in Sect. 7.2. Then, we decompose the resulting equations into a set of translinear-loop equations of the form of Eq. 7.3.6. In the decomposition process, it is sometimes convenient to introduce intermediate currents, which serve to parameterize some of the relationships, allowing us to further decompose the system of equations.

Building the Translinear Loops Once we have a set of translinear-loop equations, we construct a closed loop of TEs for each one. In general, there will be more than one translinear loop that implements any given translinear-loop equation. Our choices of loop topology and current ordering must be guided by experience and other system-level design considerations. For instance, loops with an alternating topology are better than stacked loops for systems that require a low power-supply voltage. On the other hand, stacked loops are easier to bias in the case that the same current must pass through multiple TEs that face in the same direction. In a stacked loop, we only have to supply a single copy of the input current to the circuit and we can pass the same current from one element to the next in a stack of TEs. In an alternating loop, there are no runs of clockwise or counterclockwise TEs, so if we must pass the same current through N TEs facing in the same direction, then we must supply N matched copies of the input current to the circuit.

Biasing the Translinear Loops The process of biasing a translinear loop involves forcing input currents into the emitter or collector of each input TE in the loop and arranging some type of local negative feedback around the TE, thereby adjusting its gate-to-emitter voltage so that the TE passes the input current. Once again, there are many possible feedback arrangements that will properly bias each TE in any given loop. In some cases, operational amplifiers may be used to bias a translinear loop. Here, we shall consider only the three simplest possibilities, which are shown in Fig. 7.7.

Figure 7.7(a) shows the ubiquitous *diode connection*. Here, we force a current into the collector of a counterclockwise TE, which corresponds to a voltage increase, and we feed the collector voltage back to the gate. Any mismatch between the input current and the TE's collector current causes the gate voltage to charge up or down in such a way to reduce the mismatch. If the collector current is bigger than the input current, the gate will discharge, reducing the gate-to-emitter voltage and thereby reducing the collector current. If the collector current is smaller than the input current, the gate will charge up, increasing the gate-to-emitter voltage, and thereby increasing the collector current. Of course, we can introduce additional circuitry between the collector and the gate, such as one or more buffer stages, as long as changes in the collector voltage induce like changes in the gate voltage. In some cases, we actually use other TEs in the loop with their input-current sources to form emitter-follower buffer stages in making a diode connection. In some cases, we may introduce buffer stages to eliminate base-current errors in bipolar translinear-loop circuits.

Figure 7.7(b) shows an *emitter-follower connection*. Here, we force a current out of the emitter of a clockwise TE, which corresponds to a voltage drop. In such an arrangement, for any given gate voltage, the emitter voltage will adjust itself up or down so that the emitter current just balances the input current. If the emitter current is larger than the input current, the emitter voltage will charge up, reducing the gate-to-emitter voltage, thereby reducing the emitter current, until the currents match. If the input current is larger than the emitter current, the emitter will be discharged, increasing the gate-to-emitter voltage, and thereby increasing the emitter current until the currents balance.

Figure 7.7(c) shows a simple alternative to the emitter-follower connection for biasing clockwise TEs. Here, we force a current into the collector of a clockwise TE and we feed the collector voltage back through another transistor to adjust the emitter current. An *n*FET transistor is shown in Fig. 7.7(c), but a bipolar transistor could be used instead. A similar feedback arrangement (that

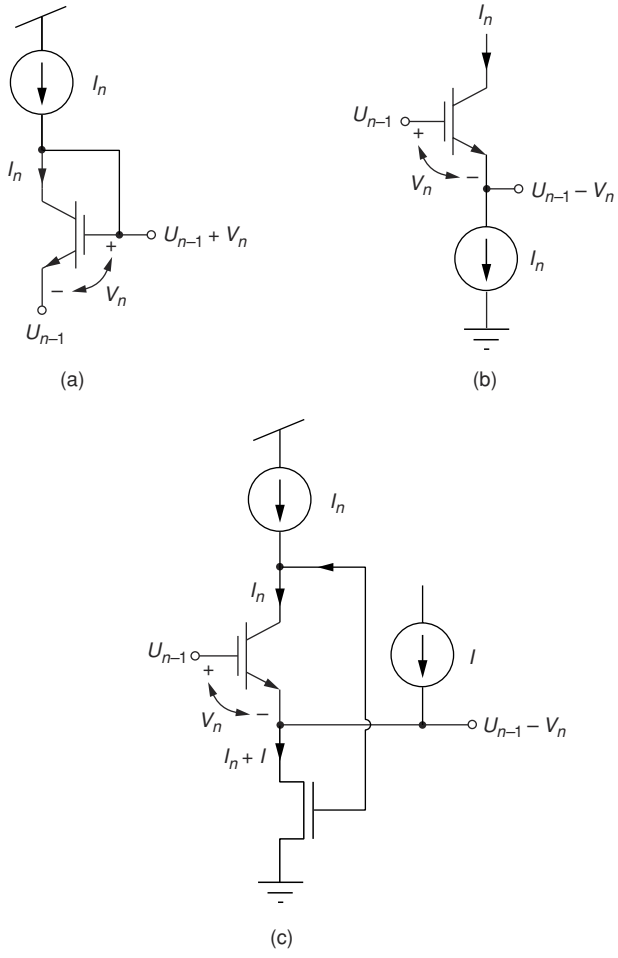


Figure 7.7
Three simple biasing arrangements for TEs. (a) Collector current forcing with the diode connection. (b) Emitter current forcing with the emitter-follower connection. (c) Collector current forcing with the Enz–Punzenberger (EP) connection.

is, forcing the collector current by feedback to the emitter terminal) involving operational amplifiers has been used in biasing translinear circuits for many years (Gilbert, 1990). The use of this particularly elegant implementation of such a feedback arrangement in the context of translinear-circuit biasing was first proposed by Punzenberger and Enz (1996) to bias low-voltage log-domain

filters. Consequently, this connection is referred to as the *Enz–Punzenberger* (EP) connection. For a given gate voltage, any imbalance between the collector current and the input current will cause the feedback transistor to adjust the gate-to-emitter voltage in such a way as to reduce the imbalance. If the input current is too large, the collector voltage will increase, causing the feedback transistor to pull a larger current out of the emitter. Conversely, if the collector current is too large, the collector voltage will decrease, reducing the amount of current flowing through the feedback transistor. This decreased current, in turn, will cause the emitter to charge up, reducing the gate-to-emitter voltage, and thereby reducing the collector current. Note that if some additional current I is injected into the emitter node, the collector voltage will adjust itself so that the feedback transistor sinks both the current flowing through the TE and the additional current. This feature of the EP connection greatly facilitates the biasing of translinear loops with an alternating topology.

In general, any translinear loop can be biased using only two of the three simple arrangements of Fig. 7.7. We can use the diode connection for biasing counterclockwise TEs, and we can use *either* the emitter-follower connection *or* the EP connection to bias the clockwise TEs. In general, the emitter-follower connection involves only one node, whereas the EP connection involves two nodes, which can cause more complicated settling behavior with the EP connection. On the other hand, in an alternating translinear loop, by using the diode connection to bias the counterclockwise elements and the EP connection to bias the clockwise elements, we are always forcing collector currents and taking outputs from collector currents. This uniform use of collector currents allows us to freely use translinear devices whose emitter current and collector current are not nearly the same, such as the *compatible lateral bipolar transistor* (CLBT), which is always available in *any* CMOS process (Vittoz, 1983). Such devices have at least two collectors, a lateral collector and a parasitic vertical collector, each of which collects only a finite fraction of the emitter current.

Consolidating the Circuit In some cases, after we have biased each of the translinear loops in the circuit, we will recognize some redundancy between the loops in the circuit. For example, if two TEs in different loops pass the same current and are at the same voltage level, then these devices are redundant and may be shared between the loops. Such consolidation is a good idea, because it usually results in smaller circuits, and fewer opportunities for errors resulting from device mismatch.

Synthesis of a Two-Quadrant Translinear Multiplier

Suppose that we want to implement a circuit that multiplies two quantities, x and y. Further, suppose that x can be either positive or negative and that y is strictly positive. Their product

$$z = xy \tag{7.4.1}$$

can be either positive or negative. We shall represent y by I_y/I_u and we use a differential representation for x and z, as described in Sect. 7.2: We represent x by

$$x = x^+ - x^-,$$

where $x^+ \equiv I_x^+/I_\mathrm{u}$ and $x^- \equiv I_x^-/I_\mathrm{u}$. Likewise

$$z = z^+ - z^-$$

where $z^+ \equiv I_z^+/I_\mathrm{u}$ and $z^- \equiv I_z^-/I_\mathrm{u}$.

Next, we substitute these definitions for x, y, and z into Eq. 7.4.1 to get

$$\left(\frac{I_z^+}{I_\mathrm{u}} - \frac{I_z^-}{I_\mathrm{u}}\right) = \left(\frac{I_x^+}{I_\mathrm{u}} - \frac{I_x^-}{I_\mathrm{u}}\right)\left(\frac{I_y}{I_\mathrm{u}}\right)$$

which we can rearrange to obtain

$$I_\mathrm{u} I_z^+ - I_\mathrm{u} I_z^- = I_y I_x^+ - I_y I_x^- .$$

One straightforward way to decompose this equation into a pair of translinear-loop equations is to equate individually the positive and negative terms on each side of the equation. Using this decomposition, we obtain the pair of translinear-loop equations

$$\underbrace{I_\mathrm{u} I_z^+}_{\mathrm{CW}} = \underbrace{I_y I_x^+}_{\mathrm{CCW}} \qquad \text{and} \qquad \underbrace{I_\mathrm{u} I_z^-}_{\mathrm{CW}} = \underbrace{I_y I_x^-}_{\mathrm{CCW}} . \tag{7.4.2}$$

Figure 7.8(a) shows a pair of translinear loops with alternating topologies that implements Eq. 7.4.2. Figure 7.8(b) shows one possible biasing arrangement for these loops; we use diode connections for each of the counterclockwise elements, and EP connections for each of the clockwise elements. Again, the exact value of the voltage V_ref is not critical: It should be high enough allow an adequate swing on the common-emitter node while keeping the feedback transistor acting properly. Note that in each loop, there is a copy of I_y forced into a diode-connected TE whose emitter is fixed at V_ref. Thus, we can eliminate one of these stages and connect the diode-voltage to both loops.

Additionally, we supply a copy of I_{u} to both loops, and as we just observed, the gate of both TEs will be at the same potential. Thus, one of the I_{u} states can also be eliminated. The consolidated two-quadrant multiplier is shown in Fig. 7.8(c).

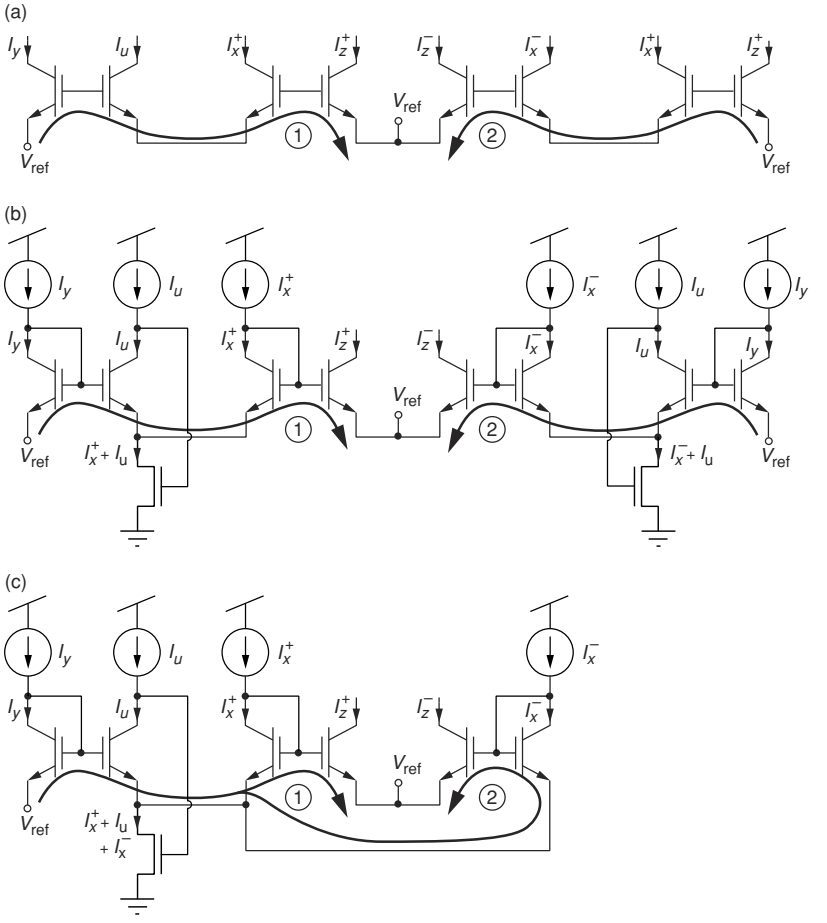


Figure 7.8
Synthesis of a two-quadrant translinear multiplier based on two alternating translinear loops. (a) A pair of alternating translinear loops that implement $I_{\mathrm{u}} I_z^+ = I_y I_x^+$ and $I_{\mathrm{u}} I_z^- = I_y I_x^-$. (b) A biasing scheme based on collector-current forcing with diode connections and EP connections. (c) The final consolidated two-quadrant translinear multiplier circuit. Here the I_y and I_{u} circuitry could be shared between the two translinear loops.

7.5 The Multiple-Input Translinear Element

Inspired originally by Shibata and Ohmi's neuron MOSFET concept (Shibata and Ohmi, 1992), we recently introduced a new translinear circuit primitive, called the *multiple-input translinear element* (MITE) (Minch, 1997; Minch et al., 1998). Such an element produces an output current I that is exponential in a weighted sum of its K input voltages, V_1 through V_K:

$$I = \lambda I_{\mathrm{s}} e^{\left(\sum_{k=1}^{K} w_k V_k - U\right)/U_{\mathrm{T}}} \tag{7.5.1}$$

where V_k is the kth input voltage, w_k is a dimensionless positive weight that scales V_k, and U is the emitter voltage of the MITE. Here I_{s}, λ, and U_{T} are the same as for the ideal TE. So defined, the MITE has K different *trans*conductances, each of which is *linear* in the MITE's output current.

Figure 7.9(a) shows a circuit symbol for an ideal K-input MITE. This symbol looks like an ideal TE whose gate voltage is set by a K-input capacitive voltage divider, where the kth divider ratio is denoted by w_k. The inputs do not need to be capacitive, but we shall assume that the input terminals draw a negligible amount of DC current and that we can control the values of the weights proportionally. In many cases, we shall be interested primarily in the number of identical unit weights, each with value w, coupling an input voltage into a MITE rather than the actual weight values involved. In such cases, each weight is an integer multiple of w and we shall omit the w associated with each of the inputs.

Figures 7.9(b) through 7.9(g) show six different practical circuit implementations of the MITE. For the first of these MITEs, shown in Fig. 7.9(b), we use a resistive voltage divider to implement the weighted voltage summation, and a bipolar transistor to implement the exponential voltage-to-current transformation. In this case, the weight associated with each input is proportional to the conductance through which that input couples into the base of the bipolar. Here, we must buffer the input voltages into the resistive network so the network neither supplies current to nor sinks current from the input nodes.

In subthreshold, the drain current of the K-input floating-gate MOS (FGMOS) transistor is proportional to the exponential of a weighted sum of its K control-gate voltages (Andreou and Boahen, 1994). Consequently, a MITE can be implemented using a single subthreshold FGMOSFET (FGMOS transistor), as shown in Fig. 7.9(c). In this case, the weight of each input is proportional to the capacitance through which that input couples into the floating gate.

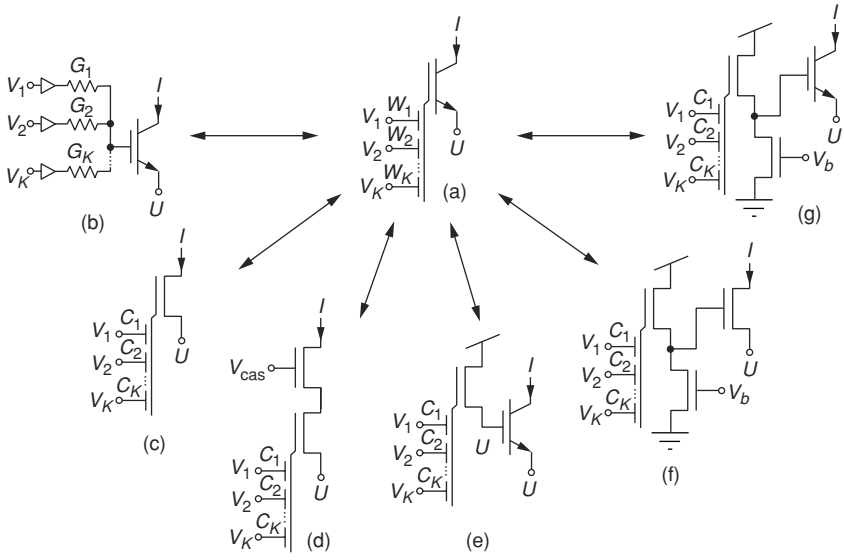


Figure 7.9
Multiple-input translinear elements (MITEs). (a) Circuit symbol for an ideal K-input MITE. Such an element produces an output current that is exponential in a weighted sum of its input voltages. Parts (b) through (g) show six different MITE implementations comprising (b) a resistive voltage divider and a bipolar transistor, (c) a single subthreshold floating-gate MOS (FGMOS) transistor, (d) a cascoded subthreshold FGMOSFET, (e) a subthreshold FGMOSFET and a bipolar transistor, (f) a floating-gate source follower and a subthreshold MOSFET, and (g) a floating-gate source follower and a bipolar transistor. For each of the five FGMOS MITE implementations, shown in parts (c) through (g), we can use the amount of floating-gate charge to store electronically adjustable, non-volatile multiplicative scale factors that we can use to build adaptive information-processing systems, or to compensate for device mismatch.

The subthreshold FGMOSFET is a good MITE implementation over the entire range of subthreshold currents. The main limitation of the subthreshold FGMOSFET as a MITE is the existence of a parasitic gate-to-drain capacitance, which results from a small region of overlap between the polysilicon gate and the drain diffusion region that arises during processing. Because the gate of a FGMOSFET is floating, an increase in the drain voltage couples into the floating gate through this overlap capacitance, thereby increasing the subthreshold drain current exponentially. In principle, this coupling can be decreased by making the FGMOSFET narrower (thereby decreasing the overlap capacitance); or by making the control-gate capacitances larger (thereby increasing the total floating-gate capacitance and so decreasing the drain capacitive-divider ratio); or by using both techniques. However, in practice, neither of

these solutions are attractive. A better solution to this problem is to cascode the subthreshold FGMOSFET, as shown in Fig. 7.9(d). We can think of the cascode transistor as a source follower with a constant input voltage V_{cas}: It reduces the swing of the drain voltage of the FGMOSFET (that is, the source follower's output voltage), and so decreases the change in current through both transistors. The cascoded subthreshold FGMOSFET is an excellent MITE implementation over the subthreshold range of currents.

Figure 7.9(e) depicts a two-transistor MITE comprising a K-input subthreshold FGMOSFET and a bipolar transistor. Intuitively, this bipolar-FGMOS MITE works as follows: The subthreshold FGMOSFET produces a current that is exponential in the weighted sum of the input voltages; again, the weight of each input is proportional to the capacitance through which that input couples into the floating gate. The bipolar transistor then acts as a current-gain stage by multiplying the subthreshold FGMOSFET current by the bipolar's forward current gain. Consequently, the upper end of the current range over which this two-transistor circuit is a good MITE implementation is extended beyond that of the subthreshold MOSFET by the bipolar's current gain, β. Because the drain of the FGMOSFET is held at a fixed potential, this MITE is insensitive to the parasitic drain-overlap capacitance.

The final two MITEs, shown in Figs. 7.9(f) and 7.9(g), are similar: Each comprises a two-transistor FGMOS source follower and a third transistor that has an exponential current–voltage characteristic. Intuitively, the floating-gate voltage develops as a weighted sum of the K input voltages via a capacitive voltage divider. In the source-follower configuration, the FGMOSFET's source voltage is approximately a linear function of the floating-gate voltage. Consequently, the source voltage is also a weighted sum of the input voltages. The third transistor then generates a current that is exponential in this source voltage. In the MITE of Fig. 7.9(f), the exponential element is a subthreshold MOSFET: Whereas, in that of Fig. 7.9(g), the exponential element is a bipolar transistor. Because the drains of the FGMOSFETs are held at a fixed potential, these MITEs also do not suffer from the drain-overlap capacitance problem.

The source-follower circuit configuration does not depend on the form of the current–voltage relationship of the MOSFET. Consequently, these three-transistor circuits are good MITE implementations even when the FGMOS source follower is biased with an above-threshold current. For the circuit of Fig. 7.9(f), biasing the FGMOS source follower with an above-threshold current allows us to make the output MOSFET as wide as necessary to obtain a larger range of exponential currents, without having to make the FGMOSFET,

and hence the floating-gate capacitance, large. The above-threshold bias gives the FGMOS source follower enough bandwidth to drive the large gate capacitance of a wide output MOSFET. The circuit of Fig. 7.9(g) is a good MITE implementation only when the base current is negligible compared with the source-follower bias current. Thus, biasing the FGMOS source follower with above-threshold currents allows this MITE to operate at high current levels and so with high bandwidth.

7.6 Multiple-Input Translinear Element Networks

In this section, we introduce three basic circuit stages, each constructed from a single MITE. These three circuit stages are the bricks from which we build a class of low-voltage translinear circuits, which we call *MITE networks*, that are equivalent to the class of translinear-loop circuits. Then, we shall examine how we can compose these stages to make translinear circuits.

Basic MITE Circuit Stages

Consider the three basic MITE circuit stages that are depicted in Fig. 7.10. The first of these circuits is a *voltage-in, current-out* (VICO) stage, shown in Fig. 7.10(a). Here, we apply input voltages V_i and V_k to two different input terminals of Q_n, which generates an output current I_n. To see how I_n depends on V_i and V_k, we use Eq. 7.5.1 to write

$$I_n \propto e^{(w_{ni}V_i + w_{nk}V_k + \cdots)/U_\mathrm{T}}. \tag{7.6.1}$$

The second of the three basic MITE stages, shown in Fig. 7.10(b), is a *current-in, voltage-out* (CIVO) stage. Here, we source an input current I_i into the output of Q_i, and we feed the output voltage V_i back through the self-coupling weight w_{ii}. This feedback configuration adjusts V_i, so that the current sunk by Q_i just balances the input current I_i. A MITE in this feedback configuration is analogous to a diode-connected transistor; so we say that it is *diode connected through* w_{ii}. To determine how the output voltage V_i depends on the input current I_i, we begin with Eq. 7.5.1 and solve for V_i in terms of I_i:

$$I_i \propto e^{(w_{ii}V_i + \cdots)/U_\mathrm{T}}$$

which can be rearranged to obtain

$$V_i = \frac{U_\mathrm{T}}{w_{ii}} \log I_i - \cdots. \tag{7.6.2}$$

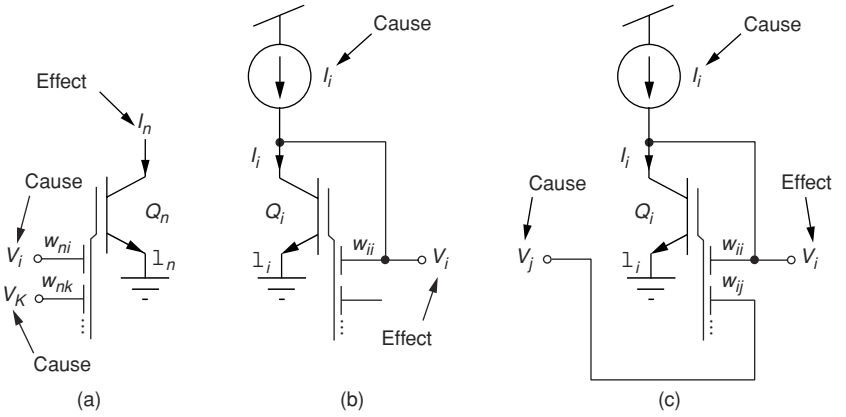


Figure 7.10
Three basic circuit stages, each comprising a single MITE. (a) A voltage-in, current-out (VICO) stage. (b) A current-in, voltage-out (CIVO) stage. (c) A voltage-in, voltage-out (VIVO) stage.

The third basic MITE stage is a *voltage-in, voltage-out* (VIVO) stage, shown in Fig. 7.10(c). This configuration is identical to the CIVO stage of Fig. 7.10(b), except that we now hold the current I_i fixed, and we are concerned instead with how the output voltage V_i depends on an input voltage V_j, which we apply to another of the input terminals of Q_i. Beginning with Eq. 7.5.1, we write

$$I_i \propto e^{(w_{ii}V_i + w_{ij}V_j + \cdots)/U_\mathrm{T}}$$

which can be rearranged to solve for V_i in terms of V_j:

$$V_i = -\frac{w_{ij}}{w_{ii}}V_j - \cdots. \tag{7.6.3}$$

We can use the circuit stage of Fig. 7.10(c) both as a CIVO stage and as a VIVO stage simultaneously. In this case, it is easy to see that V_i depends on V_j and I_i through a linear combination of of Eqs. 7.6.2 and 7.6.3 as follows:

$$V_i = \frac{U_\mathrm{T}}{w_{ii}}\log I_i - \frac{w_{ij}}{w_{ii}}V_j - \cdots. \tag{7.6.4}$$

Elementary MITE Networks

In this section, we shall examine two simple current-mode MITE circuits, each comprising two CIVO stages and a single VICO stage. These two basic current-mode circuits illustrate all of the basic intuition behind the operation

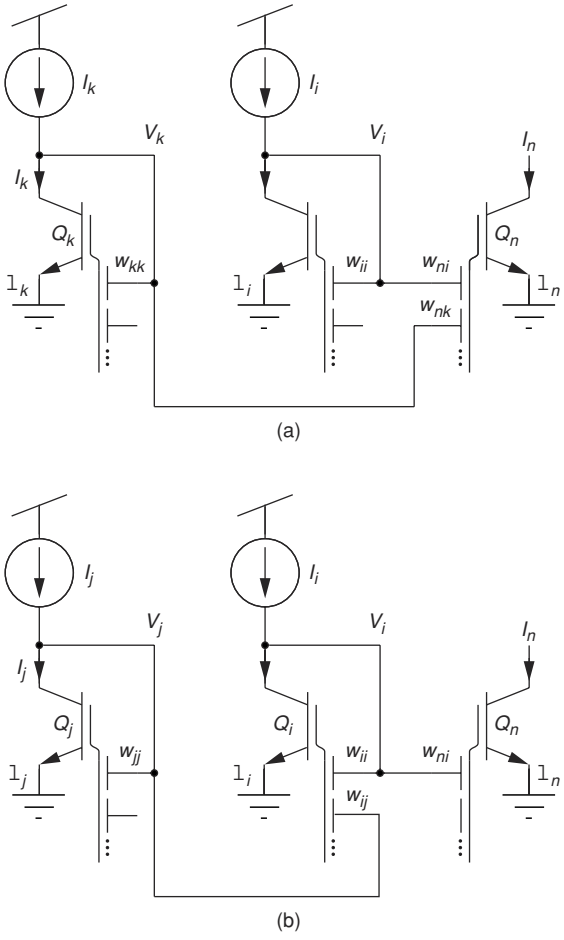


Figure 7.11
Two basic current-mode circuits comprising two CIVO stages and one VICO stage. These two circuits illustrate all of the intuition underlying the class of MITE networks. (a) A product-of-power-law circuit. (b) A quotient-of-power-law circuit.

of MITE networks.

In the first current-mode circuit, shown in Fig. 7.11(a), the outputs of two different CIVO stages are connected directly to a single VICO stage through separate inputs. To analyze this circuit, we apply Eq. 7.6.1 to the output stage:

$$I_n \propto e^{w_{ni} V_i / U_\mathrm{T}} e^{w_{nk} V_k / U_\mathrm{T}}. \tag{7.6.5}$$

Substituting Eq. 7.6.2 into Eq. 7.6.5 for each of V_i and V_k, we obtain

$$I_n \propto \exp\left[\frac{w_{ni}}{U_\mathrm{T}}\left(\frac{U_\mathrm{T}}{w_{ii}}\log I_i - \cdots\right)\right]\exp\left[\frac{w_{nk}}{U_\mathrm{T}}\left(\frac{U_\mathrm{T}}{w_{kk}}\log I_k - \cdots\right)\right].$$

Using the first term in each of the two summations and regrouping, we obtain

$$I_n \propto \exp\left[\frac{U_\mathrm{T}}{U_\mathrm{T}}\frac{w_{ni}}{w_{ii}}\log I_i\right]\exp\left[\frac{U_\mathrm{T}}{U_\mathrm{T}}\frac{w_{nk}}{w_{kk}}\log I_k\right]. \tag{7.6.6}$$

Note that, if MITEs Q_i, Q_k, and Q_n are operating at the same temperature, then the primary temperature dependence of the relationship among I_i, I_k, and I_n disappears from Eq. 7.6.6. In this intuitive analysis, we have not kept track of the scaling currents I_s, which can be strongly temperature dependent. As we have demonstrated previously (Minch, 1997), when properly designed, the relationship between the output current and the input currents for one of these circuits is generally insensitive to isothermal variations. Consequently Eq. 7.6.6 can be rewritten as

$$I_n \propto I_i^{w_{ni}/w_{ii}} \times I_k^{w_{nk}/w_{kk}}. \tag{7.6.7}$$

Thus, the output current is proportional to the product of the two input currents, each of which is raised to a power that is set by a ratio of weights.

For the second basic current-mode MITE circuit, instead of connecting the output of the second CIVO stage directly to a second input of the output VICO stage (as in the circuit of Fig. 7.11(a)), we now connect the output of the second CIVO stage to the output VICO stage through the first CIVO stage, as shown in Fig. 7.11(b). This first CIVO stage both generates a voltage that is logarithmic in the input current I_i, and serves as a VIVO stage for the second CIVO stage. This connection allows us to obtain negative powers. To illustrate this property, we apply Eq. 7.5.1 to the output VIVO stage:

$$I_n \propto e^{(w_{ni}V_i + \cdots)/U_\mathrm{T}}. \tag{7.6.8}$$

Substituting Eq. 7.6.4 into Eq. 7.6.8, we obtain

$$I_n \propto \exp\left[\frac{w_{ni}}{U_\mathrm{T}}\left(\frac{U_\mathrm{T}}{w_{ii}}\log I_i - \frac{w_{ij}}{w_{ii}}V_j - \cdots\right)\right].$$

Substituting Eq. 7.6.2 into the preceding equation for V_j and rearranging yields

$$I_n \propto I_i^{w_{ni}/w_{ii}} \times I_j^{-(w_{ni}/w_{ii})(w_{ij}/w_{jj})}$$

which becomes

$$I_n \propto \frac{I_i^{w_{ni}/w_{ii}}}{I_j^{(w_{ni}/w_{ii})(w_{ij}/w_{jj})}}. \tag{7.6.9}$$

Thus, the output current is proportional to the quotient of the two input currents, each of which is raised to a power that is set by ratios of weights. Here, the powers are not completely independent of each other. However, for any value of w_{ni}/w_{ii}, we can adjust the value of w_{ij}/w_{jj} to set the power of I_j to any value. This quotient-of-power-law relationship is also insensitive to isothermal variations.

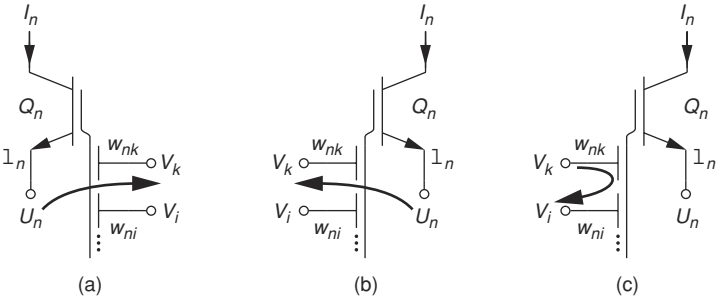


Figure 7.12
Three possible ways of traversing a MITE embedded in a MITE network. We can go (a) from an emitter to a control gate, (b) from a control gate to an emitter, or (c) from a control gate to another control gate.

These two basic current-mode circuits capture all of the intuition underlying MITE-network operation: Voltages that are logarithmic in the input currents are generated using diode-connected MITEs. Power laws are set through ratios of weights, and negative powers are obtained through voltage-inversion stages. Products are obtained by summing two or more logarithmic voltages on MITEs.

We have formalized this intuitive analysis and have obtained systematic analysis and synthesis procedures for this class of nonlinear circuits (Minch, 1997; Minch et al., 1999).

7.7 Analysis of MITE Networks

In this section, we shall develop a by-inspection analysis procedure for MITE networks by extending the analysis procedure for translinear loops of subthreshold MOSFETs discussed in Sect. 7.3. We have previously published analysis procedures for MITE networks (Minch et al., 1996; Minch, 1997). The analysis method that we develop here is somewhat more general and requires fewer initial definitions. Additionally, we can use this procedure directly to analyze subthreshold MOS translinear circuits that make use of the back gate (that is, the substrate) in addition to the front gate (van der Gevel and Kuenen, 1994; Mulder et al., 1995; Andreou and Boahen, 1996; Fried and Enz, 1996; Serrano-Gotarredona et al., 1999). To do so, we simply view the four-terminal subthreshold MOSFET as a two-input MITE with a weight of κ for the front gate and of $1 - \kappa$ for the back gate and apply the procedure that we shall develop.

As we go around loops in a MITE network, a MITE can be traversed in the three possible ways depicted in Fig. 7.12. Q_n can be traversed by going from its emitter U_n to one of its control gates V_k (Fig. 7.12(a)); this traversal corresponds to going through a counterclockwise element in a translinear-loop circuit. In this case, Eq. 7.5.1 can be rearranged to obtain the following recursion relation:

$$e^{V_k/U_\mathrm{T}} = \left(e^{U_n/U_\mathrm{T}}\right)^{1/w_{nk}} \left(\frac{I_n}{\lambda_n I_\mathrm{s}}\right)^{1/w_{nk}} \prod_{j \neq k} \left(e^{V_j/U_\mathrm{T}}\right)^{-w_{nj}/w_{nk}}. \quad (7.7.1)$$

Conversely, Q_n can be traversed by going from one of its control gates V_k to its emitter, U_n (Fig. 7.12(b)); this traversal corresponds to going through a clockwise element in a conventional translinear-loop circuit. In this case, Eq. 7.5.1 can be rearranged to obtain the following recursion relation:

$$e^{U_n/U_\mathrm{T}} = \left(e^{V_k/U_\mathrm{T}}\right)^{w_{nk}} \left(\frac{\lambda_n I_\mathrm{s}}{I_n}\right) \prod_{j \neq k} \left(e^{V_j/U_\mathrm{T}}\right)^{w_{nj}}. \quad (7.7.2)$$

Finally, Q_n can be traversed by going from one of its control gates V_k to another of its control gates V_i (Fig. 7.12(c)); this transition has no analog in conventional translinear-loop circuits. In this case, Eq. 7.5.1 can be rearranged

to obtain another recursion relationship:

$$e^{V_i/U_\mathrm{T}} = \left(e^{V_k/U_\mathrm{T}}\right)^{-w_{nk}/w_{ni}} \left(\frac{I_n}{\lambda_n I_\mathrm{s}}\right)^{1/w_{ni}} \left(e^{U_n/U_\mathrm{T}}\right)^{1/w_{ni}} \cdot \prod_{j\neq i,k} \left(e^{V_j/U_\mathrm{T}}\right)^{-w_{nj}/w_{ni}}. \tag{7.7.3}$$

The final product in each of these three recursion relationships accounts for the contributions of peripheral control gates, which do not lie directly on the particular path through the MITE network that we have chosen. Multiple translinear loops can flow together through these extra control gates, like tributaries joining to form a river. Translinear loops can split into multiple paths and merge back together again. We shall call such translinear loops *confluent*. The existence of confluent translinear loops makes the analysis of MITE networks slightly more involved than conventional translinear-loop circuits, because we may have to traverse several confluent translinear loops to analyze a given circuit completely. To analyze a network, we first identify a loop through the circuit that traverses most of the MITEs. We proceed around the loop from node to node, building up a translinear-loop expression as we go, by applying the recursion relationship appropriate to each transition. If there is a confluence of translinear loops, we trace through each one until we have built a complete translinear-loop expression.

We shall now illustrate this analysis procedure by applying it to several simple networks. Consider the network shown in Fig. 7.13(a), which comprises three two-input MITEs. Note that all of the emitters are grounded in this circuit. In this case, all of the e^{U_n/U_T} factors in the recursion relationships evaluate to unity. Consequently, we can ignore them in applying the recursion relationships. Moreover, we shall show by construction in Sect. 7.8 that any translinear-loop equation can be realized by a network with grounded emitters. However, in some cases, it may prove beneficial to have some emitters at a potential other than ground. In such cases, we would have to keep track of the emitter factors. To analyze the circuit of Fig. 7.13(a), we first identify a loop through the circuit that traverses as many of the MITEs as possible. We begin at the emitter of Q_1, which is grounded, and proceed to node V_1 through Q_1. Then, we move to node V_2 through Q_3. Finally, we return to ground by moving to the emitter of Q_2. This single loop traverses each MITE in the circuit; there

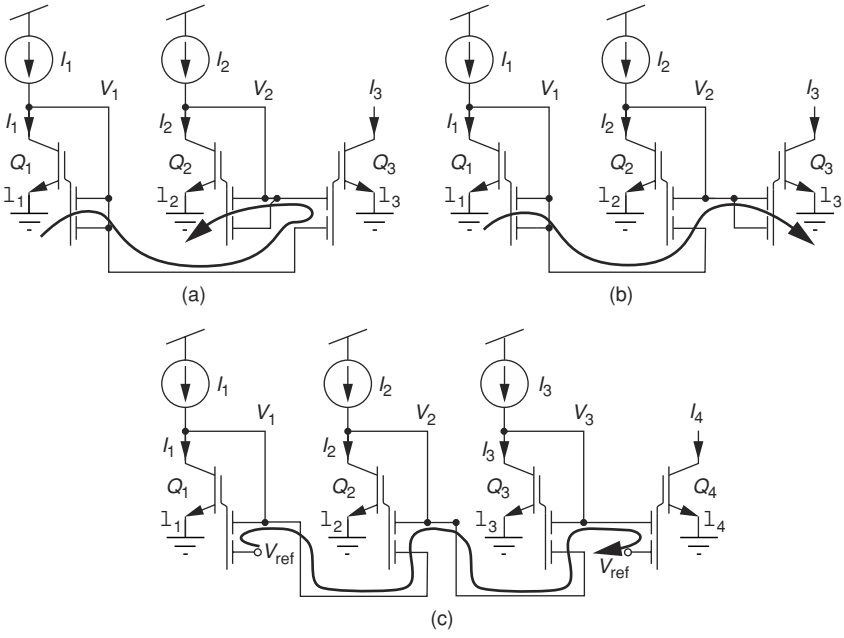


Figure 7.13
Three networks comprising two-input MITEs that can be analyzed completely by tracing a single loop. (a) A two-input geometric-mean circuit. (b) A squaring-reciprocal circuit. (c) A multiply-reciprocal circuit.

are no confluent loops. By following the procedure just described, we have that

$$\underbrace{\left(\underbrace{(1)^{1/2w}\left(\frac{I_1}{\lambda_1 I_\mathrm{s}}\right)^{1/2w}}_{e^{V_1/U_\mathrm{T}}}\right)^{-w/w}\left(\frac{I_3}{\lambda_3 I_\mathrm{s}}\right)^{1/w}}_{e^{V_2/U_\mathrm{T}}}{}^{2w}\left(\frac{\lambda_2 I_\mathrm{s}}{I_2}\right)=1$$

which can be simplified to obtain

$$\left(\frac{\lambda_1}{I_1}\right)\left(\frac{I_3}{\lambda_3}\right)^2\left(\frac{\lambda_2}{I_2}\right)=1.$$

By rearranging the preceding equation, the following translinear-loop expression is obtained:

$$\left(\frac{I_3}{\lambda_3}\right)^2=\left(\frac{I_1}{\lambda_1}\right)\left(\frac{I_2}{\lambda_2}\right)$$

which can be solved for the output current

$$I_3 = \frac{\lambda_3}{\sqrt{\lambda_1 \lambda_2}} \sqrt{I_1 I_2}.$$

Thus, the circuit of Fig. 7.13(a) is a two-input geometric-mean circuit. If each MITE has the same value of λ (that is, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$), then the output current is simply given by

$$I_3 = \sqrt{I_1 I_2}.$$

Next, consider the network shown in Fig. 7.13(b), which also comprises three two-input MITEs. To analyze this circuit, we first identify a loop through the circuit that traverses as many of the MITEs as possible. We begin at the emitter of Q_1, which is grounded, and proceed to node V_1 through Q_1. Then, we move to node V_2 through Q_2. Finally, we return to ground by moving to the emitter of Q_3. This single loop traverses each MITE in the circuit; once again, there are no confluent loops for us to trace. If we go around this loop, applying the recursion relationship appropriate to each move, we find that

$$\underbrace{\left(\underbrace{(1)^{1/2w} \left(\frac{I_1}{\lambda_1 I_\mathrm{s}} \right)^{1/2w}}_{e^{V_1/U_\mathrm{T}}} \right)^{-w/w} \left(\frac{I_2}{\lambda_2 I_\mathrm{s}} \right)^{1/w}}_{e^{V_2/U_\mathrm{T}}}{}^{2w} \left(\frac{\lambda_3 I_\mathrm{s}}{I_3} \right) = 1$$

which can be simplified to obtain

$$\left(\frac{\lambda_1}{I_1} \right) \left(\frac{I_2}{\lambda_2} \right)^2 \left(\frac{\lambda_3}{I_3} \right) = 1.$$

By rearranging the preceding expression, we obtain the translinear-loop equation:

$$\left(\frac{I_2}{\lambda_2} \right)^2 = \left(\frac{I_1}{\lambda_1} \right) \left(\frac{I_3}{\lambda_3} \right) \tag{7.7.4}$$

which, apart from a simple renumbering of the currents, is identical to the equation we derived for the circuit of Fig. 7.13(a). This result should not be too surprising, because the two networks shown in Fig. 7.13 have the same basic topology; they are merely biased differently. We can rearrange Eq. 7.7.4

to obtain the output current

$$I_3 = \frac{\lambda_1 \lambda_3}{\lambda_2^2} \frac{I_2^2}{I_1}.$$

Thus, the circuit of Fig. 7.13(b) is a squaring-reciprocal circuit. Again, if each MITE has the same value of λ (that is, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$), then the output current is simply given by

$$I_3 = \frac{I_2^2}{I_1}.$$

Next, consider the network shown in Fig. 7.13(c), which comprises four two-input MITEs. To analyze this circuit, we first identify a loop through the circuit that traverses as many of the MITEs as possible. We begin at one of the control gates of Q_1, which is connected to V_{ref}, and proceed to node V_1 through Q_1. Then, we move to node V_2 through Q_2. Then, we move to V_3 through Q_3. Finally, we return to V_{ref} through Q_4. This single loop traverses each MITE in the circuit; once again, there are no confluent loops to trace. If we go around this loop, applying the recursion relationship appropriate to each move, we find that

$$\underbrace{\left(\underbrace{\left(\underbrace{\left(\underbrace{\left(e^{V_{\text{ref}}/U_{\text{T}}}\right)^{-\frac{w}{w}} \left(\frac{I_1}{\lambda_1 I_{\text{s}}}\right)^{\frac{1}{w}}}_{e^{V_1/U_{\text{T}}}} \right)^{-\frac{w}{w}} \left(\frac{I_2}{\lambda_2 I_{\text{s}}}\right)^{\frac{1}{w}}}_{e^{V_2/U_{\text{T}}}} \right)^{-\frac{w}{w}} \left(\frac{I_3}{\lambda_3 I_{\text{s}}}\right)^{\frac{1}{w}}}_{e^{V_3/U_{\text{T}}}} \right)^{-\frac{w}{w}}} \cdot \left(\frac{I_4}{\lambda_4 I_{\text{s}}}\right)^{\frac{1}{w}} = e^{V_{\text{ref}}/U_{\text{T}}}$$

which can be simplified to obtain

$$\left(e^{wV_{\text{ref}}/U_{\text{T}}}\right) \left(\frac{\lambda_1}{I_1}\right) \left(\frac{I_2}{\lambda_2}\right) \left(\frac{\lambda_3}{I_3}\right) \left(\frac{I_4}{\lambda_4}\right) = e^{wV_{\text{ref}}/U_{\text{T}}}.$$

Rearranging the preceding expression, we obtain the translinear-loop equation:

$$\left(\frac{I_1}{\lambda_1}\right) \left(\frac{I_3}{\lambda_3}\right) = \left(\frac{I_2}{\lambda_2}\right) \left(\frac{I_4}{\lambda_4}\right). \tag{7.7.5}$$

We can rearrange Eq. 7.7.5 to obtain the output current

$$I_4 = \frac{\lambda_2 \lambda_4}{\lambda_1 \lambda_3} \frac{I_1 I_3}{I_2}.$$

Thus, the circuit of Fig. 7.13(c) is a multiply-reciprocal circuit. Again, if each MITE has the same value of λ (that is, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda$), then the output current is simply

$$I_4 = \frac{I_1 I_3}{I_2}.$$

For each of the networks that we have analyzed so far, we only had to trace a single loop around the network to fully characterize the circuit. We shall now consider a simple example where we must trace at least two confluent translinear loops to fully analyze the circuit. Figure 7.14 shows a network comprising four two-input MITEs. We cannot identify a single loop through this circuit that traverses each MITE: We must consider at least two loops that are confluent with one another. To analyze this circuit, we first identify a loop through the circuit that traverses as many of the MITEs as possible. As shown in Fig. 7.14(a), we begin at the emitter of Q_1, which is grounded, and proceed to node V_1 through Q_1. Then, we move to node V_3 through Q_3. Then, we return to ground through Q_4. We have not traversed Q_2 at all in this loop. If we go around this loop, applying the recursion relationship appropriate to each move, we find that

$$\underbrace{\left(\underbrace{(1)^{1/2w} \left(\frac{I_1}{\lambda_1 I_\mathrm{s}} \right)^{1/2w}}_{e^{V_1/U_\mathrm{T}}} \right)^{-w/w} \left(\frac{I_3}{\lambda_3 I_\mathrm{s}} \right)^{1/w}}_{e^{V_3/U_\mathrm{T}}}{}^{w} \left(\frac{\lambda_4 I_\mathrm{s}}{I_4} \right) \left(e^{V_2/U_\mathrm{T}} \right)^w = 1 \tag{7.7.6}$$

which has a factor of e^{V_2/U_T}, that we would like to express in terms of the collector currents. We can derive a suitable expression for e^{V_2/U_T} by traversing the confluent loop shown in Fig. 7.14(b). If we go around this second loop, applying the recursion relationship appropriate to each move and substitute the resulting expression for e^{V_2/U_T} directly into Eq. 7.7.6, we find that

$$\underbrace{\left(\underbrace{\left(1\underbrace{^{\frac{1}{2w}}\left(\frac{I_1}{\lambda_1 I_\mathrm{s}}\right)^{\frac{1}{2w}}}_{e^{V_1/U_\mathrm{T}}}\right)^{-\frac{w}{w}}\left(\frac{I_3}{\lambda_3 I_\mathrm{s}}\right)^{\frac{1}{w}}}_{e^{V_3/U_\mathrm{T}}}\right)^{w}\left(\frac{\lambda_4 I_\mathrm{s}}{I_4}\right)}$$
$$\cdot\underbrace{\left(\underbrace{\left(1\underbrace{^{\frac{1}{2w}}\left(\frac{I_1}{\lambda_1 I_\mathrm{s}}\right)^{\frac{1}{2w}}}_{e^{V_1/U_\mathrm{T}}}\right)^{-\frac{w}{w}}\left(\frac{I_2}{\lambda_2 I_\mathrm{s}}\right)^{\frac{1}{w}}}_{e^{V_2/U_\mathrm{T}}}\right)^{w}}=1.$$

The preceding equation can be simplified:

$$\left(\frac{\lambda_1}{I_1}\right)^{1/2}\left(\frac{I_3}{\lambda_3}\right)\left(\frac{\lambda_4}{I_4}\right)\left(\frac{\lambda_1}{I_1}\right)^{1/2}\left(\frac{I_2}{\lambda_2}\right)=1$$

and can be rearranged to obtain the translinear-loop equation:

$$\left(\frac{I_3}{\lambda_3}\right)\left(\frac{I_2}{\lambda_2}\right)=\left(\frac{I_1}{\lambda_1}\right)\left(\frac{I_4}{\lambda_4}\right). \tag{7.7.7}$$

We can rearrange Eq. 7.7.7 to obtain the output current

$$I_4=\frac{\lambda_4\lambda_1}{\lambda_2\lambda_3}\frac{I_2 I_3}{I_1}.$$

Thus, the circuit of Fig. 7.14 is also a multiply-reciprocal circuit. Again, if each MITE has the same value of λ (that is, $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\lambda$), then the output current is simply

$$I_4=\frac{I_2 I_3}{I_1}.$$

7.8 ABC's of MITE-Network Synthesis

In this section, we shall consider the basics of MITE-network synthesis. As was the case with translinear-loop circuit synthesis, the problem of synthesizing MITE networks is underconstrained and there are design trade-offs involved in the process. Once again, in our brief discussion, we cannot be exhaustive and we shall endeavour to make the basic procedure clear by illus-

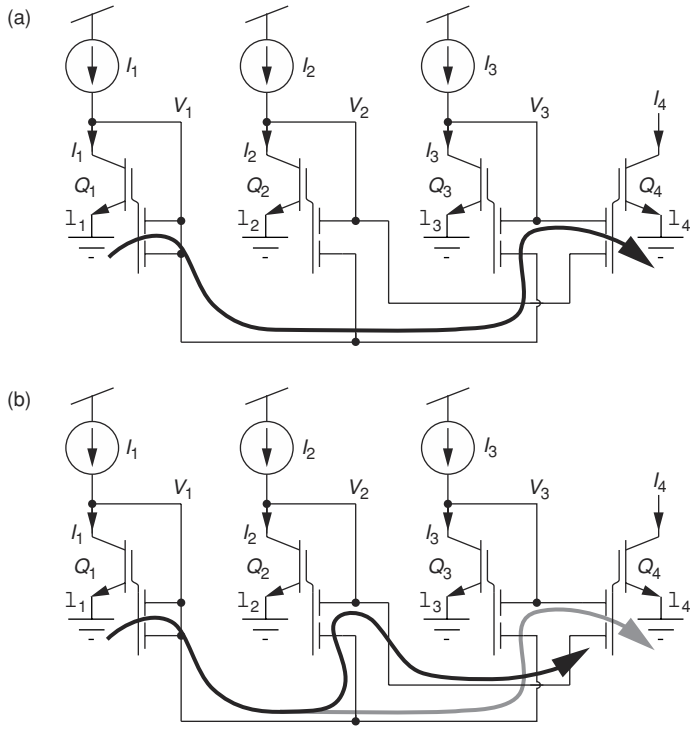


Figure 7.14
A network comprising four two-input MITEs. To analyze this circuit, at least two loops that are confluent with one another must be considered: (a) The primary loop used to analyze the circuit, and (b) the confluent loop that we trace to complete the analysis.

trating it with some simple examples. The starting point for MITE-network synthesis is identical to that of translinear-loop circuit synthesis: A set of translinear-loop equations is derived from some functional or behavioral description of the system to be implemented. The final steps are also similar: The networks can often be consolidated in the same way that translinear-loop circuits can, by merging redundant parts of the circuits that we have synthesized. Because the initial and final steps are similar, we shall focus on the middle steps in the synthesis procedure: The construction of MITE networks from translinear-loop equations.

Synthesizing Static MITE Networks

As with the synthesis of translinear-loop circuits, we can summarize the synthesis of MITE networks as follows: First, we *acquire* a set of translinear-loop equations from a behavioral or functional description of the system that we want to implement. Next, we *build* a network for each of the translinear-loop equations. This construct involves a *building* phase, a *balancing* phase, a *biasing* phase, and a *completion* phase. Finally, if possible, we *consolidate* the resulting networks by merging parts of them where possible. We shall discuss each of these steps in turn, then we shall use them to synthesize two simple multiplier circuits.

Acquiring a Set of Translinear-Loop Equations We start the construction process with a set of translinear-loop equations, each of the form

$$\prod_{n\in\text{“CW”}} I_n^{k_n} = \prod_{n\in\text{“CCW”}} I_n^{k_n} \tag{7.8.1}$$

where “CW” denotes a set of clockwise currents and “CCW” denotes a set of counterclockwise currents[1], and the k_n are positive integer powers to which the currents are raised, such that

$$\sum_{n\in\text{“CW”}} k_n = \sum_{n\in\text{“CCW”}} k_n. \tag{7.8.2}$$

With translinear-loop circuits, the reason for restricting the powers to be integers is obvious: A current is raised to a given power because it passes through an integer number of TEs facing in the same direction around a loop. A current cannot pass through a fractional number of TEs. However, with MITE networks, it is certainly possible to allow these powers to be positive real numbers subject to the constraint expressed in Eq. 7.8.2, but we shall restrict our attention here to the case of integer powers for two reasons: Firstly, integer powers suffice for many practical purposes. Secondly, with MITE networks these powers are set by ratios of weights and we obtain the most accurate ratios by connecting an integer number of identical unit cells in parallel with one another. The procedure by which we obtain such translinear-loop equations is identical to the one described in Sect. 7.4.

1 In the context of MITE networks, such designations are not as meaningful as they are in translinear-loop circuits.

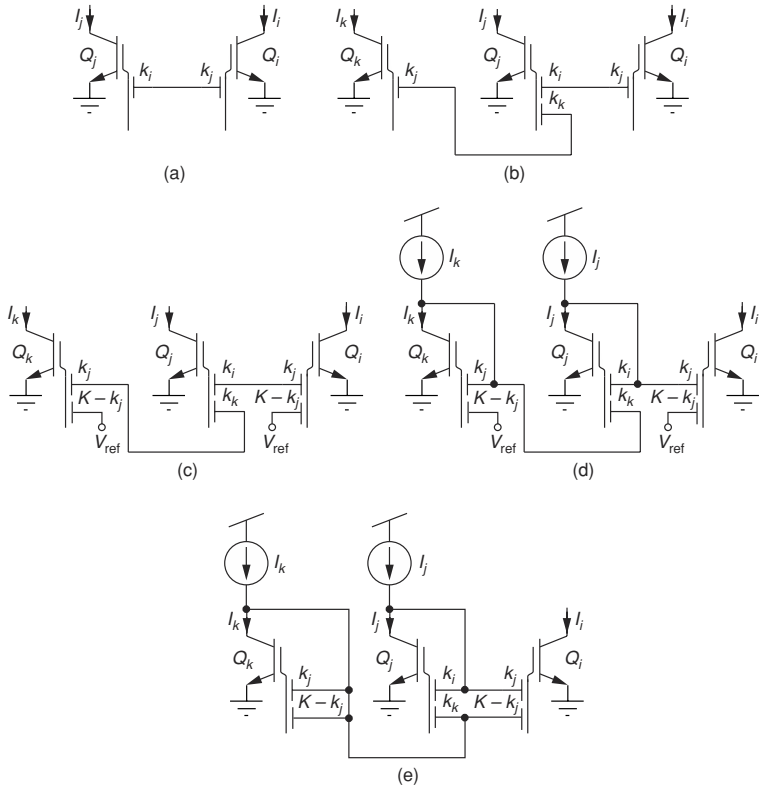


Figure 7.15
Steps in the construction of MITE networks. Here k_i, k_j, and k_k are integers that each denote some number of identical inputs and K denotes the total number of control gates for each MITE. (a) Beginning the network. (b) Building the network. (c) Balancing the network. (d) Biasing the network. (e) Completing the network.

Building MITE Networks For each translinear-loop equation, we build a MITE network. We begin a network by picking a current from each set (for example, current I_i from the "clockwise" set and current I_j from the "counterclockwise" set). We create a new MITE for each one and make a new node in the circuit, coupling it into Q_i through k_j unit inputs and into Q_j through k_i unit inputs, as shown in Fig. 7.15(a). If k_i and k_j have a factor in common, we can divide both by that factor in determining the number of unit inputs for each connection. For each additional current in the translinear-loop equation (for example, current I_k from the "clockwise" set), we create a new MITE, and we

make a new node in the circuit, connecting it to an existing MITE whose current is from the *opposite* set (for example, Q_j) through k_k unit inputs and to Q_k through k_j unit inputs, as shown in Fig. 7.15(b). Once again, if k_j and k_k have a factor in common, we can divide both by that factor in determining the number of unit inputs for each connection. We continue adding MITEs in this way until we have exhausted all of the currents in the translinear-loop equation. The order in which we add MITEs and the existing MITEs to which we connect them affect the structure of the final network and the number of inputs that fan-in to each MITE.

Once we have built the basic network for a translinear-loop equation, as just described, we then balance the fan-in of all MITEs in the network. Suppose that the largest MITE fan-in is K. We then add a sufficient number of unit inputs to each MITE, connected to an appropriate voltage V_{ref}, so they each have a fan-in of K, as shown in Fig. 7.15(c). As long as the translinear-loop equation from which we started conforms to Eq. 7.8.2, the exact value of V_{ref} is not critical: The quiescent collector voltages in the MITE network will depend on the value of V_{ref}, but as long as all of the collector voltages stay sufficiently far away from the power supply rails, the network's behavior is independent of the value of V_{ref}.

We need to balance the number of inputs to each MITE in the network because of the way in which we implement the weighted voltage summation. If we implement the weighted voltage summation using a capacitive voltage divider, as discussed in Sect. 7.5, then each weight is equal to a coupling capacitance divided by a total floating-gate capacitance. The power-law relationships implemented by a network are given by ratios of weights. As designers, we would like these powers to be independent of the total floating-gate capacitances, because they include (nonlinear) parasitic capacitances. By requiring the total floating-gate capacitance of each MITE to be the same, the total floating-gate capacitances will cancel in the weight ratios, making them depend only on ratios of coupling capacitors. The best way to ensure that the total floating-gate capacitances are the same is to require that each MITE have an identical complement of inputs. In the context of integer numbers of unit inputs, we would give each MITE the same number of unit inputs.

Biasing MITE Networks The process of biasing a MITE network is considerably simpler than that of biasing a translinear-loop circuit. We simply force input currents into the collectors of some of the MITEs and diode-connect them by connecting some of their control gates to their collectors, as shown

in Fig. 7.15(d). Those MITEs that are diode connected become inputs, while those that are not diode connected are outputs. Other biasing schemes are certainly possible, but are never needed.

Completing MITE Networks It may seem that by adding unused MITEs in the process of balancing the fan-in in a network, we are wasting resources. Indeed, such unused inputs can account for a significant fraction of the total transconductance of a MITE. Leaving them unused leads to larger collector-voltage swings and a higher required power-supply voltage. It so happens that, as long as the translinear-loop equation from which we started conforms to Eq. 7.8.2, we can utilize all of the extra control gates that we add during the balancing phase (Minch, 1997). Intuitively, because the behavior of a MITE network is unaffected by the value of V_{ref}, we can short V_{ref} to one of the collector voltages in the network without affecting the behavior of the circuit. This MITE-network transformation is called *completion* (Minch, 1997). We connect all of the unused inputs to one of the collector voltages, as shown in Fig. 7.15(e). In doing so, we should generally avoid the creation of feedback loops in the network that could affect its stability. We can always do this by choosing a MITE that only has self connections.

Consolidating MITE Networks In some cases, as with translinear-loop circuits, after we have biased each of the MITE networks in the circuit, we will recognize some redundancy between them. For example, if two MITEs in different networks pass the same current and their control gates are connected in the same manner, then these MITEs are redundant and may be shared between the networks. Such consolidation is usually a good idea, because it usually results in smaller circuits and fewer opportunities for errors resulting from device mismatch. Other, more subtle forms of MITE-network consolidation are possible (Minch, 2000b), but are beyond the scope of this chapter.

Synthesis of One-Quadrant MITE-Network Multipliers

Suppose that we want to implement a multiplication operation with strictly positive inputs using a MITE network. That is, we want to find a network that implements the relationship

$$z = xy \tag{7.8.3}$$

where $x > 0$, $y > 0$, and $z > 0$ are dimensionless quantities. Here x and y are the independent variables (that is, the inputs) and z is the dependent variable

(that is, the output). First, we represent x by $I_x/I_{\rm u}$, y by $I_y/I_{\rm u}$ and z by $I_z/I_{\rm u}$, where $I_{\rm u}$ is the unit current. Then, we substitute these definitions of x, y, and z into Eq. 7.8.3, obtaining

$$\frac{I_z}{I_{\rm u}} = \frac{I_x}{I_{\rm u}}\frac{I_y}{I_{\rm u}}$$

which can easily be rearranged to obtain the translinear-loop equation

$$\underbrace{I_z I_{\rm u}}_{\text{“CW”}} = \underbrace{I_x I_y}_{\text{“CCW”}}. \tag{7.8.4}$$

Starting from Eq. 7.8.4, we shall synthesize two different MITE networks to illustrate how the building order can influence the structure of the final network. We begin the first network by selecting I_z from the “CW” set and I_x from the “CCW” set and make a MITE for each one. Then, we make a new node in the circuit and couple it into Q_z through one unit input, and into MITE Q_x through one unit input, as shown in Fig. 7.16(a). Next, we select I_y from the “CW” set and make another MITE for it. We make a new node and couple it into Q_z through one unit input, and into Q_y through one unit input, as shown in Fig. 7.16(b). Next, we select $I_{\rm u}$ from the “CCW” set and make another MITE for it. We make a new node and couple it into MITE Q_y through one unit input, and into $Q_{\rm u}$ through one unit input, as shown in Fig. 7.16(c). Next, we balance the fan-in of all MITEs. In this case, MITEs Q_z and Q_y have two inputs, whereas MITEs Q_x and $Q_{\rm u}$ have only one. To balance the fan-in in the network, we add another unit input to MITEs Q_x and $Q_{\rm u}$, each connected to $V_{\rm ref}$, as shown in Fig. 7.16(d). Next, we bias the network by forcing $I_{\rm u}$ into $Q_{\rm u}$, I_y into Q_y, and I_x into MITE Q_x; and diode connect each one through one control gate, as shown in Fig. 7.16(e). This network implements Eq. 7.8.3, but it has two unused inputs. We can utilize these two inputs and complete the network by connecting $V_{\rm ref}$ to the collector of MITE $Q_{\rm u}$, as shown in Fig. 7.16(f). This network also implements Eq. 7.8.3, but has no unused inputs. Finally, no consolidation is possible.

We begin the second network by selecting I_z from the “CW” set and I_x from the “CCW” set and make a MITE for each one. Then, we make a new node in the circuit and couple it into Q_z through one unit input, and into Q_x through one unit input, as shown in Fig. 7.17(a). Next, we select $I_{\rm u}$ from the “CCW” set and make another MITE for it. We make a new node and couple it into Q_x through one unit input, and into $Q_{\rm u}$ through one unit input, as shown in Fig. 7.17(b). Next, we select I_y from the “CW” set and make another MITE

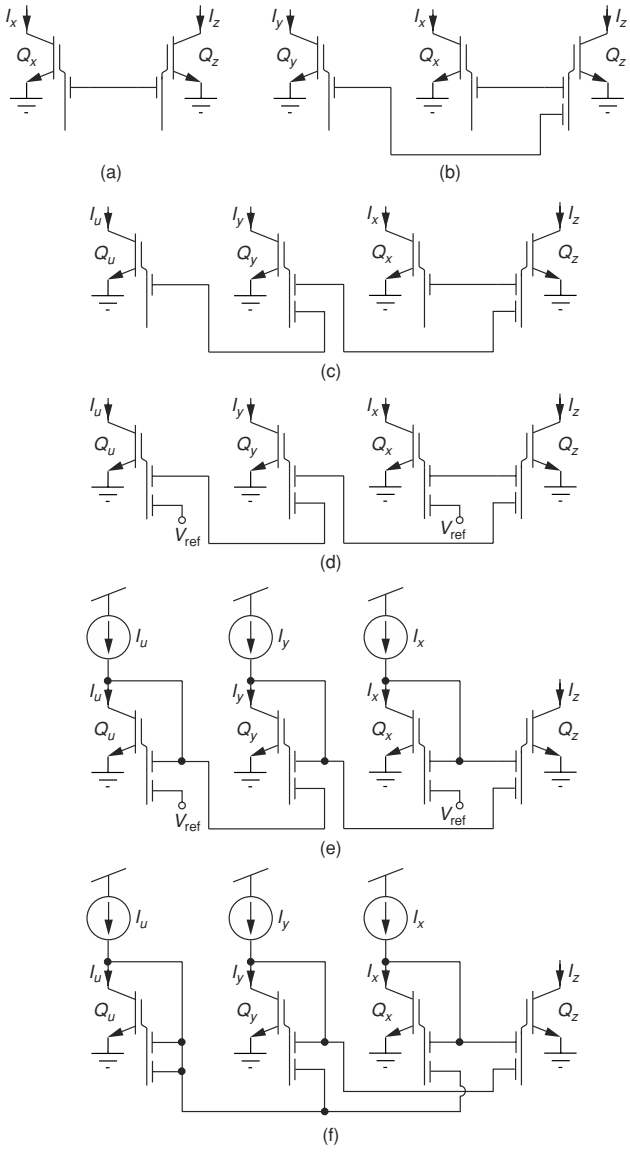


Figure 7.16
Synthesis of a one-quadrant MITE-network multiplier. (a) Beginning the network. (b, c) Building the network. (d) Balancing the network. (e) Biasing the network. (f) Completing the network.

for it. We make a new node and couple it into Q_y through one unit input, and into Q_{u} through one unit input, as shown in Fig. 7.17(c). Next, we balance the fan-in of all MITEs. In this case, MITEs Q_x and Q_{u} have two inputs, whereas MITEs Q_y and Q_z have only one. To balance the fan-in in the MITE network, we add another unit input to MITEs Q_y and Q_z, each connected to V_{ref}, as shown in Fig. 7.17(d). Next, we bias the network by forcing I_{u} into Q_{u}, I_y into Q_y, and I_x into Q_x; and diode connect each one through one control gate, as shown in Fig. 7.17(e). This network implements Eq. 7.8.3, but it has two unused inputs. We can utilize these two unused inputs and complete the network by connecting V_{ref} to the collector of Q_y, as shown in Fig. 7.17(f). Note that, if we had connected V_{ref} to the collector of Q_{u}, then we would have introduced a positive feedback loop into the network with a loop gain of unity, making the desired network equilibrium an unstable one. If we had connected V_{ref} to the collector of Q_x, then we would have introduced a negative feedback loop, creating the potential for instability. This network also implements Eq. 7.8.3, but has no unused inputs. Finally, no consolidation is possible.

We have synthesized four different MITE networks (Figs. 7.16(e), 7.16(f), 7.17(e), and 7.17(f)) that each implement Eq. 7.8.3. The circuits of Fig. 7.16 are more symmetric with respect to how many stages separate the x and y inputs from the z output than those of Fig. 7.17. Intuitively, we should expect that a network with fewer stages on average between the inputs and an outputs would be less sensitive to mismatch in MITE weight values, than would be a network with more stages. We have demonstrated this fact for these four one-quadrant multiplier circuits elsewhere (Minch, 1997). The circuits shown in Fig. 7.16 differ from those shown in Fig. 7.17 in the order in which we selected the currents in the building process and where we chose to connect their MITEs. Because the number of ways in which currents can be chosen from a translinear-loop equation grows rapidly in the number of currents, it is difficult to say general things about how the chosen order affects the performance of the final network. However, we shall make some observations: The more MITEs that we connect to any given MITE, the larger the required fan-in per MITE in the network as a whole, but the fewer the average number of intermediate stages between any two MITEs. We have shown previously (Minch, 1997) that any translinear-loop equation can be implemented as a MITE network with a maximum of one MITE between any pair of MITEs. Networks with fewer intermediate stages should be less sensitive to offset and noise accumulation than networks with more intermediate stages. Additionally, because of para-

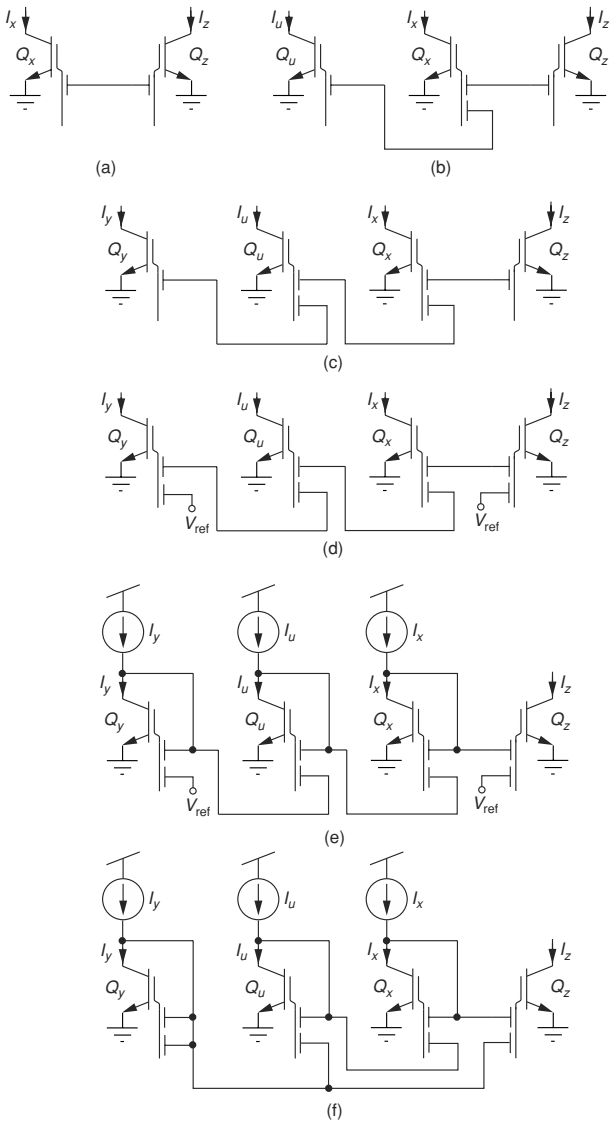


Figure 7.17
Synthesis of a one-quadrant MITE-network multiplier. (a) Beginning the network. (b, c) Building the network. (d) Balancing the network. (e) Biasing the network. (f) Completing the network.

sitic node capacitances, the response time of a network with fewer intermediate stages will be faster than that of a network with more intermediate stages.

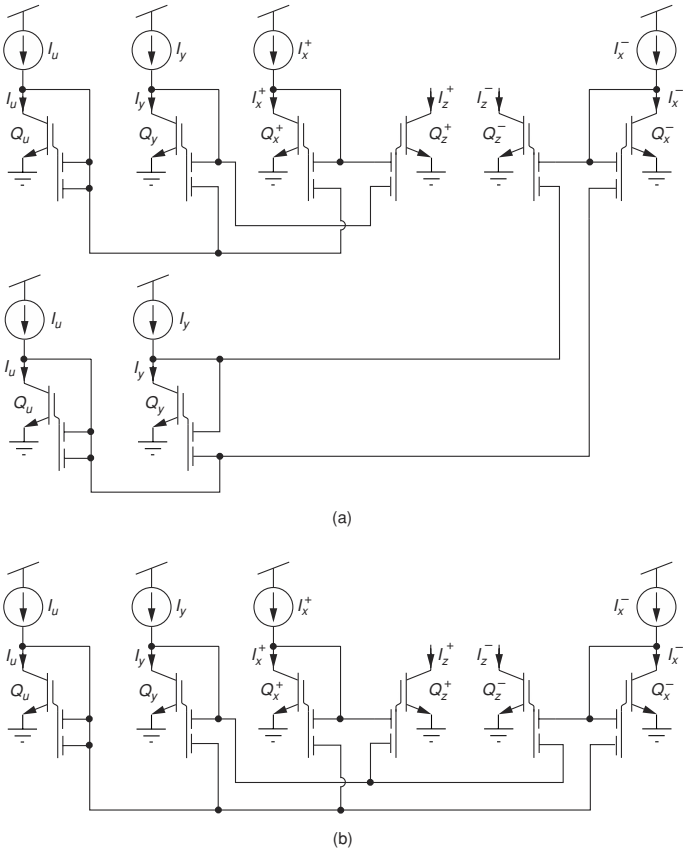


Figure 7.18
Synthesis of a two-quadrant MITE-network multiplier. (a) Two independent copies of the one-quadrant multiplier of Fig. 7.16(f). (b) The final consolidated two-quadrant MITE-network multiplier circuit. Here the I_y and I_{u} circuitry are shared between the two MITE networks.

Synthesis of a Two-Quadrant MITE-Network Multiplier

Suppose that we want to implement a circuit that multiplies two quantities, x and y, where x can be either positive or negative and that y is strictly positive.

Thus, their product

$$z = xy, \tag{7.8.5}$$

can be either positive or negative. We shall represent y by I_y/I_u and we shall use a differential representation for x and z, as described in Sect. 7.2. That is, we represent x by

$$x = x^+ - x^-$$

where $x^+ \equiv I_x^+/I_u$ and $x^- \equiv I_x^-/I_u$. Likewise, we represent z by

$$z = z^+ - z^-$$

where $z^+ \equiv I_z^+/I_u$ and $z^- \equiv I_z^-/I_u$.

Next, we substitute these definitions for x, y, and z into Eq. 7.8.5:

$$\left(\frac{I_z^+}{I_u} - \frac{I_z^-}{I_u}\right) = \left(\frac{I_x^+}{I_u} - \frac{I_x^-}{I_u}\right)\left(\frac{I_y}{I_u}\right)$$

which can be rearranged to obtain

$$I_u I_z^+ - I_u I_z^- = I_y I_x^+ - I_y I_x^-.$$

One straightforward way to decompose this equation into a pair of translinear-loop equations is to equate individually the positive and negative terms on each side of the equation. Using this decomposition, we obtain the pair of translinear-loop equations:

$$\underbrace{I_u I_z^+}_{\text{CW}} = \underbrace{I_y I_x^+}_{\text{CCW}} \qquad \text{and} \qquad \underbrace{I_u I_z^-}_{\text{CW}} = \underbrace{I_y I_x^-}_{\text{CCW}}.$$

By following the procedure shown in Fig. 7.16 for each of these translinear-loop equations, we obtain the two independent MITE networks shown in Fig. 7.18(a). Note that in each MITE network, we supply a copy of I_u to a MITE that is diode connected through two unit inputs. The collector voltages of these two MITEs should be identical, so a single Q_u can be shared between the two networks. Also, in each network, we supply a copy of I_y to a MITE that is diode connected through a single control gate and its other control gate is connected to an identical voltage. Thus, their collector voltages are also identical, and we should be able to share a single Q_y between the two circuits. The consolidated two-quadrant MITE-network multiplier is shown in Fig. 7.18(b).

This page intentionally left blank

III DYNAMICS

This page intentionally left blank

8 Linear Systems Theory

In this chapter we will review some properties of linear time-invariant systems. We consider their input/output relationship in the time domain, the impulse response, and the convolution theorem. We also review basic concepts of complex number theory; the Laplace transform, system's transfer function, and frequency domain analysis. More extended descriptions of this material can be found in many standard textbooks (Carlson, 1986; Poularikas and Seely, 1994; Oppenheim et al., 1996).

8.1 Linear Shift-Invariant Systems

In linear systems theory, a system is treated as a *black box* that does not reveal its internal states, and is characterized only by the relationship between its input and output (see Fig. 8.1). If a system has no internal stored energy, then its output response $y(t)$ is forced entirely by the input $x(t)$:

$$y(t) = F[x(t)] \tag{8.1.1}$$

where $F[\cdot]$ is the transfer function.

Linearity A system is linear if it obeys the two fundamental principles: **Homogeneity**, and **additivity**.

The principle of homogeneity states that output scales linearly with the input:

$$F[\alpha x(t)] = \alpha F[x(t)]. \tag{8.1.2}$$

Usually, linear systems theory is applied to time-varying signals. However the same methods can be applied to input and output signals that are distributed

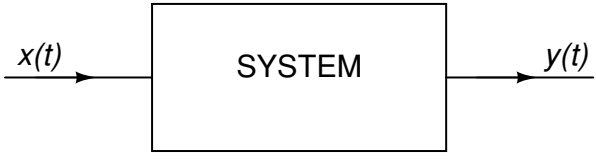


Figure 8.1
Typical black-box representation of a linear system. Its input is the signal $x(t)$ and its output is the signal $y(t)$.

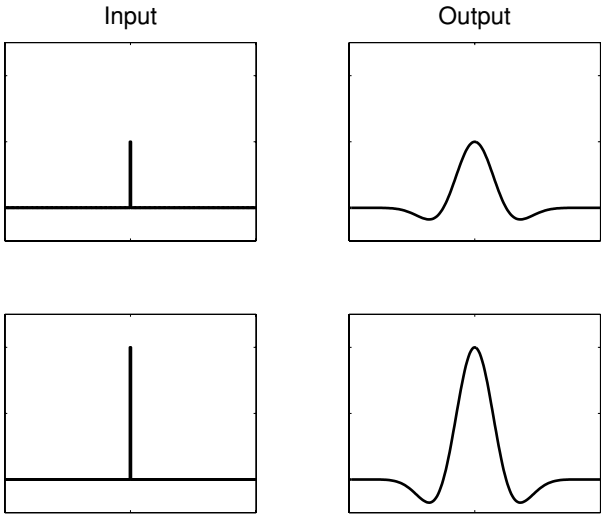


Figure 8.2
Graphical example of the homogeneity principle of a linear system. The signals in the left quadrants represent the system's input, and the signals in the right ones represent its output. An increase in the input signal causes a proportional increase in the output signal.

over *space* rather than *time*. For example, in Fig. 8.2, the input signal is a spatial unit impulse and the output is a spatial *Gabor* function (a Gaussian modulated by a cosine function)[1]

The principle of additivity states that if the input signal is composed of elementary signals, then the system's response is the composition of its responses to each of the elementary signals:

$$F[x_1(t)+x_2(t)+\cdots+x_n(t)] = F[x_1(t)]+F[x_2(t)]+\cdots+F[x_n(t)]. \quad (8.1.3)$$

Figure 8.3 shows a graphical example: If the response of the system to a spatial impulse is a Gabor function, and if the system's input signal is a linear combination of spatial unit impulses, then the system's response will be a linear combination of Gabor functions.

The principles of homogeneity and additivity taken together are commonly referred to as the *principle of superposition*, which state that a system is linear

1 Gabor functions are commonly used to model the (linear) response properties of a particular class of neurons in the visual cortex.

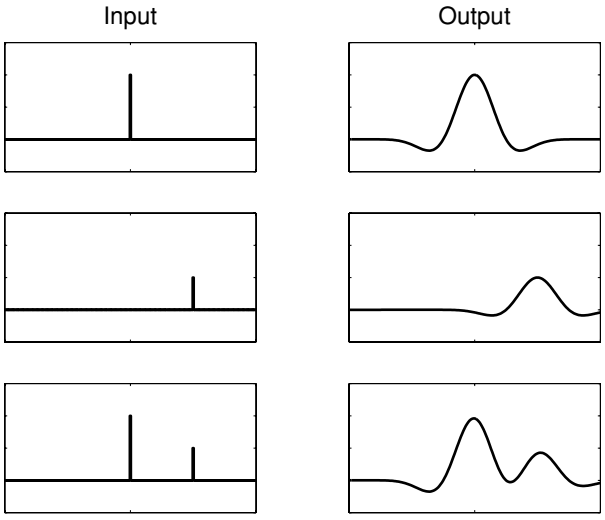


Figure 8.3
Graphical example of the additivity principle of a linear system. The signals in the left quadrants represent the system's input, and the signals in the right ones represent its output.

if

$$y(t) = \sum_k a_k F[x_k(t)] \tag{8.1.4}$$

for input $x(t) = \sum_k a_k x_k(t)$, and a_k constant for all k.

In other words, a system is linear if its response function F is a *linear operator*:

$$F\left[\sum_k a_k x_k(t)\right] = a_k \sum_k F[x_k(t)]. \tag{8.1.5}$$

Shift Invariance A system is said to be shift-invariant if its responses to identical stimuli shifted in time are also identical, except for the corresponding time shift (Fig. 8.4).

If a system is shift-invariant, then its response function is also shift-invariant: Given input signal $x(t)$, its time-shifted variant $x(t-\tau)$ will produce

$$F[x(t-\tau)] = y(t-\tau). \tag{8.1.6}$$

The system's output signal is unchanged, except for a time shift.

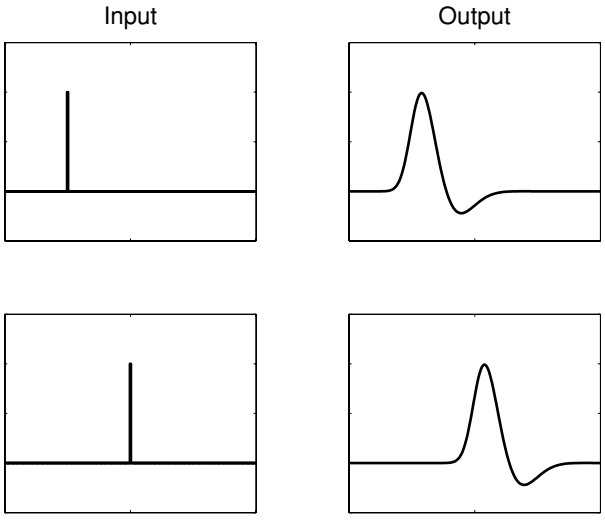


Figure 8.4
Graphical example of a time-invariant system's response.

Time invariance and linearity are two independent characteristics. *Not all linear systems are time-invariant and, similarly, not all time-invariant systems are linear.*

8.2 Convolution

Convolution is an important mathematical operator used in linear systems analysis. The convolution of two time-varying signals, $v(t)$ and $w(t)$, is

$$v(t) * w(t) \equiv \int_{-\infty}^{+\infty} v(\lambda) w(t - \lambda) d\lambda \tag{8.2.1}$$

where λ is the integration variable, and t is the independent variable.

Figure 8.5 shows a graphical representation of the convolution process between signals $v(t)$ (Fig. 8.5(a)) and $w(t)$ (Fig. 8.5(b)) at three different time steps. The result of the convolution is shown in Fig. 8.6. Note how the overlap between the two curves is null for $t < 0$, increases for $0 < t < T$, peaks at $t = T$ and decreases again for $t > T$.

If the independent variable for both input signals is the same, then it can be omitted and we express the convolution between to the two signals $v(t)$ and

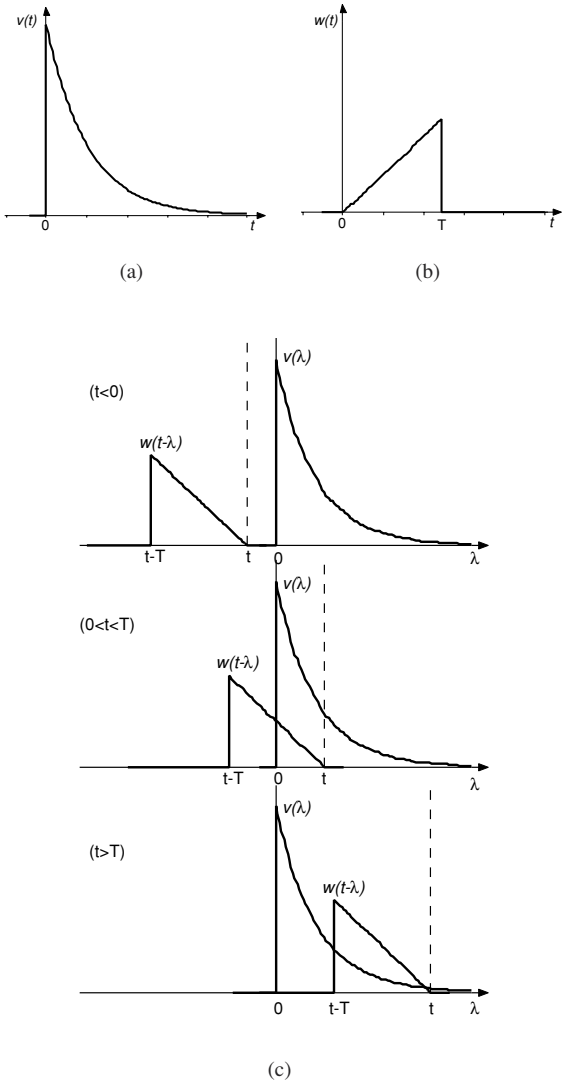


Figure 8.5
Graphical representation of the convolution between $v(t)$ and $w(t)$ for three different values of t. Note that the integration variable λ in (c) is on the abscissae of the plots. Modified from Carlson, A. B. (1986).

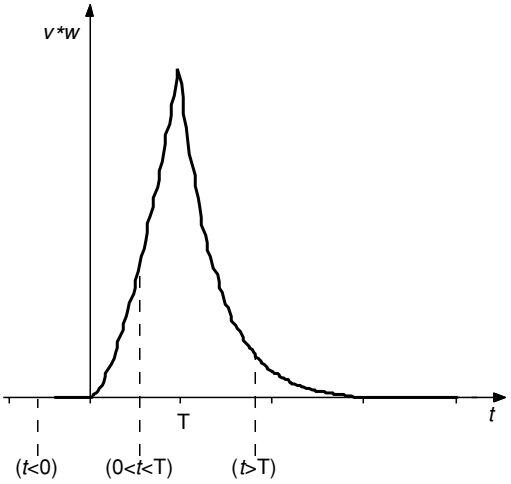


Figure 8.6
Result of the convolution between the two signals $v(t)$ and $w(t)$ of Fig. 8.5. The three dashed lines are at the three values of t used in Fig. 8.5.

$w(t)$ simply as $v * w$. The convolution operator is linear and has the following properties:

commutative:	$v * w$	$=$	$w * v$
associative:	$v * (w * z)$	$=$	$(v * w) * z$
distributive:	$v * (w + z)$	$=$	$(v * w) + (v * z)$

8.3 Impulses

The *unit impulse* or *Dirac delta function* $\delta(t)$ is not a function in the strict mathematical sense. It is defined by a set of assignment rules.

- If $v(t)$ is a continuous function at $t = 0$ then

$$\int_{t_1}^{t_2} v(t)\delta(t)dt \;=\; \begin{cases} v(0) & t_1 < 0 < t_2 \\ 0 & \text{otherwise.} \end{cases} \tag{8.3.1}$$

- If ϵ is an arbitrary small number

$$\int_{-\infty}^{+\infty} \delta(t)dt \;=\; \int_{-\epsilon}^{+\epsilon} \delta(t)dt \;=\; 1. \tag{8.3.2}$$

From these rules we can infer that $\delta(t)$ has unit area at $t = 0$ and that $\delta(t) = 0$, for all $t \neq 0$. We can also note that the Dirac delta function has no mathematical or physical meaning, unless it appears under the integral operator.

Impulse Integration Properties

When used in conjunction with the integral operator, the Dirac delta function has the following properties:

- Replication:

$$v(t) * \delta(t - \tau) = v(t - \tau) \tag{8.3.3}$$

- Sampling:

$$\int_{-\infty}^{+\infty} v(t)\delta(t - \tau)dt = v(\tau) \tag{8.3.4}$$

where $v(t)$ is a continuous time-varying signal.

Impulses in the Limit

There are many (proper mathematical) functions $\delta_\epsilon(t)$ that approach the Dirac delta function $\delta(t)$, in the limit:

$$\lim_{\epsilon \to 0} \delta_\epsilon(t) = \delta(t) \tag{8.3.5}$$

An example of such a function that is commonly used is

$$\delta_\epsilon = \frac{\sin(\frac{t}{\epsilon})}{t}. \tag{8.3.6}$$

Figure 8.7 shows how δ_ϵ of Eq. 8.3.6 approaches the Dirac delta function as ϵ decreases.

8.4 Impulse Response of a System

We can now use the notions of convolution and unit impulse to define the *impulse response* of a linear time-invariant system. If $y(t)$ is the system's response to its input $x(t)$ we can write

$$y(t) = F[x(t)]. \tag{8.4.1}$$

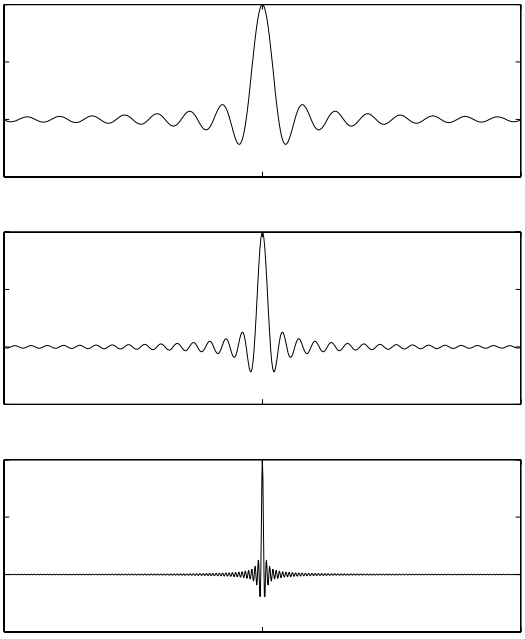


Figure 8.7
Plot of the function $\sin(t/\epsilon)/t$ for three decreasing values of ϵ.

If the input signal is the Dirac delta function ($x(t) = \delta(t)$), then the system's response to the unit impulse is defined as

$$h(t) \equiv F[\delta(t)]. \tag{8.4.2}$$

If $x(t)$ is continuous in time, the replication property of the unit impulse allows us to rewrite $x(t)$ as $x(t) * \delta(t)$. With this reformulation of the system's input signal, Eq. 8.4.1 becomes

$$y(t) = F\left[\int_{-\infty}^{+\infty} x(\lambda)\delta(t-\lambda)d\lambda\right]. \tag{8.4.3}$$

If the system is linear, Eq. 8.4.3 is equivalent to:

$$y(t) = \int_{-\infty}^{+\infty} x(\lambda)F\left[\delta(t-\lambda)\right]d\lambda \tag{8.4.4}$$

If we substitute Eq. 8.4.2 into Eq. 8.4.4, and if the system is time-invariant,

then

$$y(t) = \int_{-\infty}^{+\infty} x(\lambda)h(t-\lambda)d\lambda = \int_{-\infty}^{+\infty} h(\lambda)x(t-\lambda)d\lambda. \tag{8.4.5}$$

This property of linear time-invariant systems is extremely powerful. It states that if a system's impulse response $h(t)$ is known, the response of the system to any arbitrary signal $x(t)$ can be computed simply by performing the convolution of its impulse response with the signal itself:

$$\boxed{y(t) = h(t) * x(t)}. \tag{8.4.6}$$

Step Response

We can define a system's *step response* in the same way we defined its impulse response. If the input signal $x(t)$ is the step function

$$u(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{8.4.7}$$

then we define its step response to be

$$g(t) \equiv F[u(t)]. \tag{8.4.8}$$

A first interesting property can be obtained by exploiting the system's impulse response (see Eq. 8.4.6):

$$g(t) = h(t) * u(t). \tag{8.4.9}$$

By applying the derivative operator to this equation, and noting that the derivative of the step function is the unit impulse, we obtain

$$\frac{d}{dt}g(t) = h(t) * \frac{d}{dt}u(t) = h(t) * \delta(t). \tag{8.4.10}$$

If we apply the unit impulse replication property (Eq. 8.3.3), then we obtain

$$h(t) = \frac{d}{dt}g(t). \tag{8.4.11}$$

Thus, a system's impulse response can be obtained by computing the derivative of its step response. This property is extremely useful in practical situations because unit impulses are impossible to generate with physical instruments but it is easy to generate waveforms that approximate ideal step functions. Consequently, a physical linear time-invariant system is characterized experimentally

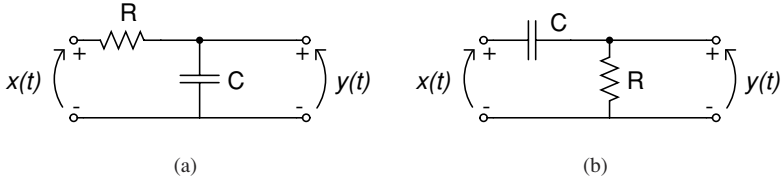


Figure 8.8
Resistor capacitor (RC) circuits. The signals $x(t)$ represent input voltages, and the signals $y(t)$ represent output voltages. (a) Integrator circuit; (b) Differentiator circuit.

by measuring its step response and then deriving its impulse response from Eq. 8.4.11.

8.5 Resistor-Capacitor Circuits

The resistor-capacitor (RC) circuits of Fig. 8.8 represent first order, linear, time-invariant systems. In both circuits, the input signal is $x(t)$ and the output signal is $y(t)$. The circuits of Fig. 8.8(a) and (b) are referred to as *RC integrator* and *RC differentiator* respectively. In this section we focus only on the properties of the RC integrator. The properties of the RC differentiator will be described in Chapter 9.

The integrator circuit of Fig. 8.8(a) is governed by the differential equation:

$$RC\frac{d}{dt}y(t) + y(t) = x(t). \tag{8.5.1}$$

By solving Eq. 8.5.1 for a unit impulse input signal ($x(t) = \delta(t)$), we obtain the circuit's *impulse response*:

$$h(t) = \frac{1}{RC}\, e^{-t/RC} \cdot u(t) \tag{8.5.2}$$

where u(t) is the step function. Similarly, solving Eq. 8.5.1 for a step input signal ($x(t) = u(t)$), we obtain the circuit's *step response*

$$g(t) = (1 - e^{-t/RC}) \cdot u(t). \tag{8.5.3}$$

Figure 8.9 shows the impulse response and the step response. The value RC is defined as the system's *time-constant* and is often labeled τ. As pointed out in Section 8.4, the response of the circuit to an arbitrary input signal can be

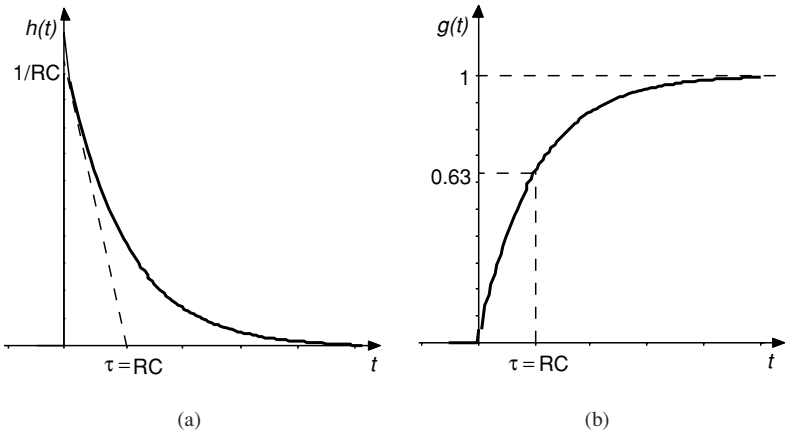


Figure 8.9
Impulse response (a) and step response (b) of an RC circuit.

obtained by the convolution between the input signal and the circuit's impulse response:

$$y(t) = x(t) * h(t) = \int_0^\infty \frac{1}{RC}\, e^{-\lambda/RC} \cdot x(t-\lambda)d\lambda. \tag{8.5.4}$$

8.6 Higher Order Equations

Time-domain analysis becomes increasingly difficult for higher order systems. Fortunately there is a *unified representation* in which any solution to a linear system can be expressed: **Exponentials with complex arguments.** All solutions to linear homogeneous (undriven) equations are of the form $e^{\,st}$ where s is a *complex number* (see Fig. 8.10):

$$s = \sigma + j\omega = M\cos(\phi) + jM\sin(\phi) \tag{8.6.1}$$

where $j = \sqrt{-1}$, σ is the real part of the complex number, ω is the imaginary part, M represents its *magnitude*, and ϕ its *phase*. Magnitude and phase of a

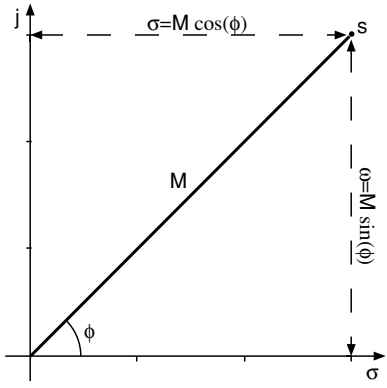


Figure 8.10
Complex number representation. The complex number s has magnitude M and phase ϕ. Its real part is σ and imaginary part is ω.

complex number obey the following relationships:

$$M = \sqrt{\sigma^2 + \omega^2} \tag{8.6.2}$$

$$\phi = \arctan\left(\frac{\omega}{\sigma}\right). \tag{8.6.3}$$

The magnitude of a complex number s is often denoted as $|s|$. Furthermore, applying the properties of complex exponentials, one can observe that

$$e^{j\phi} = \cos(\phi) + j\sin(\phi) \tag{8.6.4}$$

$$e^{-j\phi} = \cos(\phi) - j\sin(\phi). \tag{8.6.5}$$

It follows that s can be also written as

$$s = Me^{j\phi}. \tag{8.6.6}$$

These notations can be used to solve higher order differential equations. As an example, we consider the second order linear homogeneous equation

$$\frac{d^2}{dt^2}V + \alpha\frac{d}{dt}V + \beta V = 0. \tag{8.6.7}$$

Assume that e^{st} is an *eigenfunction*[2] and substitute for V:

$$s^2 e^{st} + \alpha s e^{st} + \beta e^{st} = 0. \tag{8.6.8}$$

Solving for s we obtain

$$s = \frac{-\alpha \pm \sqrt{\alpha^2 - 4\beta}}{2}. \tag{8.6.9}$$

Consequently, if $\alpha^2 - 4\beta \geq 0$, s is real, otherwise s is a complex number. In practice, if Eq. 8.6.7 is a linear system, we could measure its response $V = e^{st}$ with a *real* instrument: but if e^{st} was a complex exponential, we would measure only its real component:

$$Re\{e^{st}\} = Re\left\{e^{(\sigma + j\omega)t}\right\} = e^{\sigma t} Re\{e^{j\omega t}\}. \tag{8.6.10}$$

So the measured response of the system would be

$$V_{meas} = e^{\sigma t} \cos \omega t. \tag{8.6.11}$$

Figure 8.11 shows the possible kinds of response of V_{meas} for different values of ω and σ. If $\sigma < 0$ all the solutions are stable and decay to zero with time. If $\sigma > 0$ all solutions are unstable and diverge with time. If $\sigma = 0$ the solutions are naturally stable (they neither decay, nor diverge). All physical *passive* linear systems will have stable solutions ($\sigma < 0$). The ω axis scales the oscillation frequency f of a solution ($\omega = 2\pi f$).

8.7 The Heaviside-Laplace Transform

By analyzing the example of the previous section (see Eq. 8.6.7) we can make the following observation: Any time we substitute the eigenfunction e^{st} into a linear differential equation of order n, the following property obtains:

$$\frac{d^n}{dt^n} e^{st} = s^n e^{st}. \tag{8.7.1}$$

In other words:

We can consider s as an operator meaning *derivative* with respect to time. Similarly, we can view $\frac{1}{s}$ as the operator for *integration* with respect to time (Heaviside).

2 An eigenfunction is a nonzero solution of a second order linear homogenous differential equation

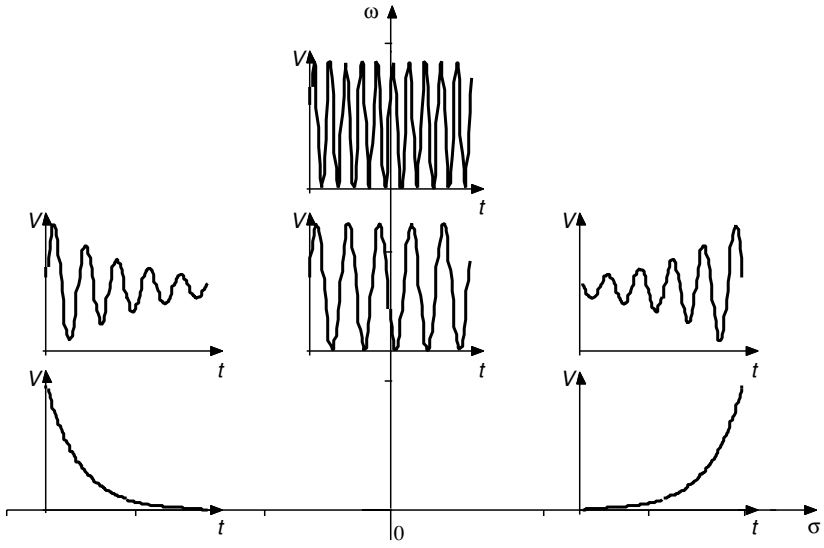


Figure 8.11
The possible kinds of *measured* responses for a first order linear system.

This observation was made by Heaviside, when trying to analyze analog circuits: but was also formalized by Laplace when he introduced the *Laplace Transform*. The Laplace transform is a useful operator that links functions that operate in the time domain with functions of complex variables:

$$\mathcal{L}[y(t)] = Y(s) \equiv \int_{-\infty}^{\infty} y(t)e^{-st}dt. \tag{8.7.2}$$

8.8 Linear System's Transfer Function

Now that we have introduced the concepts of convolution (Section 8.2), impulse response (Section 8.4), and the Laplace transform (Section 8.7), we can define a linear system's *transfer function.* It is a function defined in the complex domain:

$$H(s) \equiv \frac{Y(s)}{X(s)} \tag{8.8.1}$$

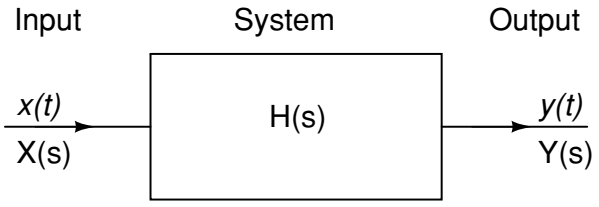


Figure 8.12
Typical representation of a linear system with input and output signals both in the time domain ($x(t)$, $y(t)$) and in the Laplace domain ($X(s)$, $Y(s)$).

where $Y(s)$ is the Laplace transform of the system's output $y(t)$ and $X(s)$ is the Laplace transform of the system's input $x(t)$ (see Fig. 8.12). Conversely, we can say that the output of any linear time-invariant system is determined by *multiplying* the system's transfer function with its input:

$$\boxed{Y(s) = H(s)X(s)}. \tag{8.8.2}$$

Transfer Function and Impulse Response

Consider the special case in which the system's input signal $x(t)$ is the unit impulse $x(t) = \delta(t)$. Its Laplace transform $X(s)$ is

$$X(s) = \int_{-\infty}^{\infty} x(t)e^{-st}dt = \int_{-\infty}^{\infty} \delta(t)e^{-st}dt = 1. \tag{8.8.3}$$

In this case, following the definition of Eq. 8.8.1, the system's response in the complex plane is

$$Y(s) = H(s). \tag{8.8.4}$$

On the other hand, the system's response in the time domain is (by definition) its impulse response:

$$y(t) = h(t). \tag{8.8.5}$$

Because $Y(s)$ is the Laplace transform of $y(t)$, we can substitute Eq. 8.8.5 into Eq. 8.7.2:

$$Y(s) = \int_{-\infty}^{\infty} y(t)e^{-st}dt = \int_{-\infty}^{\infty} h(t)e^{-st}dt \tag{8.8.6}$$

and so

$$\boxed{H(s) = \int_{-\infty}^{\infty} h(\lambda) e^{-\lambda s} d\lambda = \mathcal{L}[h(t)]}. \tag{8.8.7}$$

The transfer function $H(s)$ is the Laplace transform of the impulse response $h(t)$.

Summary Given a linear time-invariant system with input $x(t)$, output $y(t)$, and impulse response $h(t)$:

$$\begin{aligned} y(t) &= x(t) * h(t) \\ Y(s) &= X(s) H(s) \end{aligned}$$

where

$$\begin{aligned} X(s) &= \mathcal{L}[x(t)] \\ Y(s) &= \mathcal{L}[y(t)] \\ H(s) &= \mathcal{L}[h(t)]. \end{aligned}$$

8.9 The Resistor-Capacitor Circuit (A Second Look)

Consider again the RC circuit of Fig. 8.8. As mentioned in Section 8.5, this circuit is governed by

$$\tau \frac{d}{dt} y(t) + y(t) = x(t) \tag{8.9.1}$$

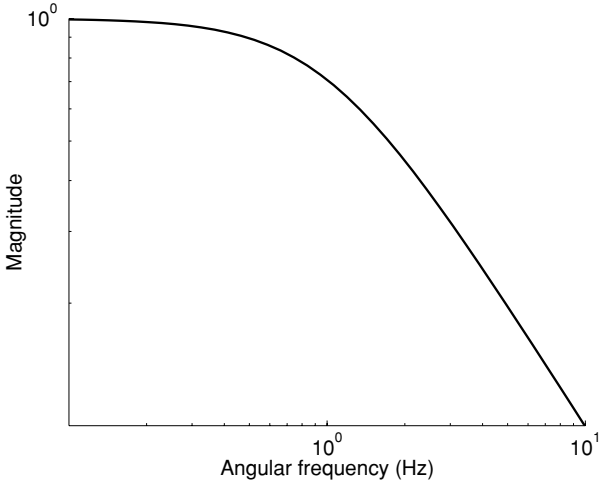
where $\tau = RC$.
In the complex domain, we have

$$Y(s)(\tau s + 1) = X(s). \tag{8.9.2}$$

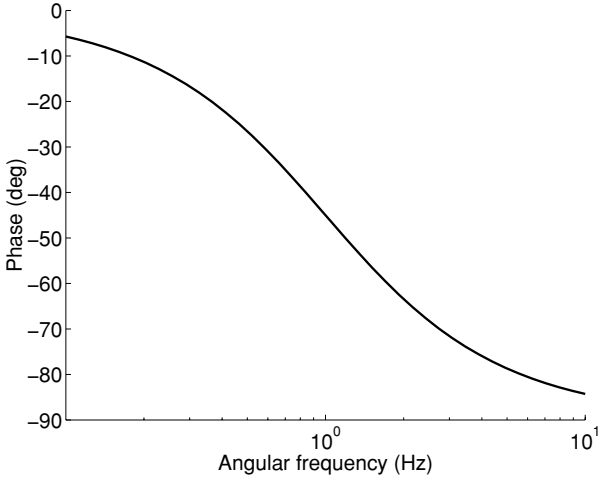
Therefore the circuit's transfer function is

$$H(s) = \frac{1}{1 + \tau s}. \tag{8.9.3}$$

Consider now how this circuit responds to sinusoidal signals of different *frequencies*. Sinusoids have a very special relationship to shift-invariant linear systems, such as the one we are analyzing. When a sinusoidal signal is applied



(a)



(b)

Figure 8.13
Bode plot of a first order linear system, such as the RC circuit of Fig. 8.8. (a) Magnitude, (b) Phase.

as input to a shift-invariant linear system, then its response will be another sinusoidal signal, with possibly a different amplitude and a different phase, but certainly with exactly the same frequency! That is, if the input is $x(t) = \sin(\omega t)$, the output will be $y(t) = A\sin(\omega t + \phi)$, where A and ϕ determine the scaling and shift.

When we analyze a system using sinusoidal signals of different frequencies, we are working in the frequency domain. In this domain $s = j\omega$ and the circuit's transfer function is

$$H(j\omega) = \frac{1}{1 + j\omega\tau}. \tag{8.9.4}$$

From this transfer function, we make two useful observations:

1. If the frequencies of the sinusoidal signals are small with respect to the circuit's time-constant ($\omega\tau \ll 1$), then the circuit's output will resemble its input ($Y(j\omega) \approx X(j\omega)$).
2. On the other hand, if the frequencies are large with respect to the circuit's time-constant ($\omega\tau \gg 1$), then

$$\frac{Y(j\omega)}{X(j\omega)} \approx \frac{1}{j\omega\tau}. \tag{8.9.5}$$

These observations are also reflected in the plots of the transfer function's magnitude and phase (Fig. 8.13). These plots are referred to as *Bode* plots and they are used to analyze the response of a dynamic system in terms of its transfer function. The magnitude of the transfer function is

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\omega\tau)^2}} \tag{8.9.6}$$

and its phase is

$$\phi = \arctan(-\omega\tau). \tag{8.9.7}$$

The frequency $\omega = \frac{1}{\tau}$ is defined as the *cutoff frequency* . In Fig. 8.13 it is set to one.

The RC circuit of Fig. 8.8 is a *low-pass filter*, because it allows sinusoidal signals with frequencies lower than the cutoff frequency to pass virtually unchanged. On the other hand, the frequency components of the input signals that are above the cutoff frequency are attenuated. The phase lag between the input and the output of the system increases with ω (see Fig. 8.13(b)) and saturates at $-90°$. Figure 8.14 shows experimental data measured from an RC

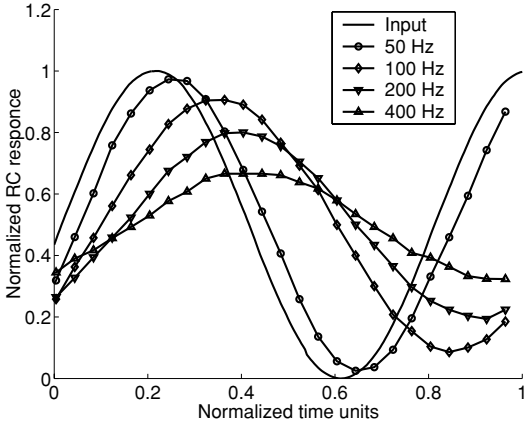


Figure 8.14
Response of an RC low-pass filter ($R = 10M\Omega$, $C = 1nF$) to input sinusoids of different frequencies. The input signals have been normalized to unity, and the outputs have been normalized with respect to the input. The time axis has also been normalized so that the responses to all the frequencies could be presented on the same graph.

lowpass filter with $R = 10M\Omega$ and $C = 1nF$. Sinuosoids of increasing frequency were applied to the circuit and the corresponding responses were measured. To show the effect of a range of input frequencies on the circuit's response, all the data are plotted on a normalized scale. The responses have been normalized with respect to the input and time has been normalized to unity. As expected, the output signal is attenuated as the input frequency increases; and the phase lag between the input and output signals increases with increasing frequency.

8.10 Low-Pass, High-Pass, and Band-Pass Filters

The RC circuit analyzed in the previous sections is the simplest example of a passive *filter*. Filters are typically used to alter the frequency spectrum of their input signals. Specifically, filters allow one or more frequency *bands* to pass unchanged (except for a multiplicative gain factor), whereas others are attenuated. Passive filters do not amplify the input signal, whereas active filters can also amplify the frequency components of the input signal. If a filter transmits low frequency components (from DC to a lower cutoff value ω_l), it is said to be a *low-pass* filter. If it transmits high frequency components (higher

than a cutoff value ω_u), it is said to be a *high-pass* filter. Filters that transmit only frequency components between a lower cutoff and an upper cutoff are said to be *band-pass* filters.

The analysis of linear systems often assumes *ideal filters*. These filters have distortionless signal transmission over one or more frequency bands, and have zeros responses at all other frequencies. For example, the transfer function of an ideal *bandpass* filter is:

$$H(\omega) = \begin{cases} Ke^{-j\omega t_d} & \omega_l \leq |\omega| \leq \omega_u \\ 0 & \text{otherwise.} \end{cases} \tag{8.10.1}$$

Although ideal filters cannot be implemented in practice, their use in theoretical analysis simplifies the study of linear systems.

9 Integrator-Differentiator Circuits

Integrators are a very useful class of low-pass circuits suitable for filtering out the high-frequency components of the signal (often present due to noise). On the other hand, differentiators filter out the low-frequency components of the input signal and respond best to its changes. Integrators and differentiators can also be used to implement *adaptation* in neuromorphic systems. Adaptation is ubiquitous in neural systems and allows a system to optimize its dynamic range against the characteristics of the prevailing input signal.

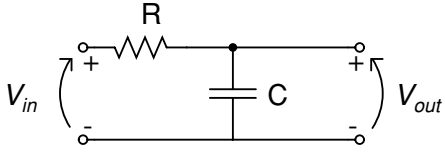


Figure 9.1
Resistor-capacitor (R-C) integrator circuit.

The simplest type of integrator is the RC circuit in Fig. 9.1. This circuit's transfer function (Section 8.9) is

$$H(s) = \frac{1}{1 + \tau s} \tag{9.0.1}$$

where $\tau = RC$ is the circuit's time-constant. This time-constant can be controlled by setting the capacitance or the resistance values of the elements of the RC circuit.

In VLSI technology, the time-constant of RC circuits implemented with passive elements cannot be changed once the chip has been fabricated. Both resistance and capacitance are fixed at design time by the geometries of the layout mask layers (see Section 12.1). By contrast, the transconductance amplifier (Section 5.3) has a transconductance that depends on its bias voltage which can be set on the operating chip. Therefore, this device can be used to design an integrator circuit that has an adjustable time-constant: This *follower-integrator* (Mead, 1989) is shown in Fig. 9.2.

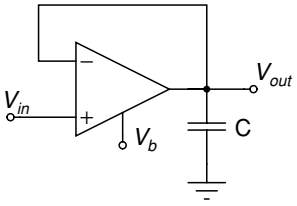


Figure 9.2
Follower-integrator circuit. The bias voltage V_b, which sets the transconductance amplifier's bias current I_b, can be used to modify the integrator's time-constant.

9.1 The Follower-Integrator

This circuit comprises a unity-gain follower (see Section 5.3), and a capacitor connected to the follower's output node. The input voltage is applied to the '+' terminal of the follower. If the circuit operates in subthreshold, we can apply Kirchhoff's current law at the circuit's output node, and write

$$\boxed{C\frac{d}{dt}V_{out} = I_b \tanh\left(\frac{\kappa(V_{in} - V_{out})}{2U_T}\right)} \tag{9.1.1}$$

as derived from Eq. 5.3.6 of Section 5.3.

Small Signal Behavior

In the small signal regime where the transconductance amplifier operates in its linear range[1], Eq. 9.1.1 can be simplified to

$$C\frac{d}{dt}V_{out} = G(V_{in} - V_{out}) \tag{9.1.2}$$

where $G = \frac{\kappa I_b}{2U_T}$ is the amplifier's transconductance (see also Eq. 5.3.8 of Section 5.3). Under these conditions, the follower-integrator has a transfer function that is identical to that of the R-C integrator (Eq. 9.0.1), with a time-constant $\tau = \frac{C}{G}$:

$$\frac{V_{out}}{V_{in}} = \frac{1}{1 + \tau s}. \tag{9.1.3}$$

The amplifier will operate in its linear range provided that V_{in} does not change too rapidly. Specifically, Eq. 9.1.2 is valid as long as $\frac{d}{dt}V_{in} < \frac{4U_T}{\tau}$ (Mead,

1 The DC component of V_{in} is in a range in which the amplifier is well behaved, and the AC component of V_{in} is sufficiently small.

1989). Under these conditions the transfer functions of the integrators of Fig. 9.1 and Fig. 9.2 are identical. However, the two circuits differ in a key property: the *composition* of transfer functions.

Composition Property

We can *compose* (connect) multiple instances of the R-C integrator circuit (see Fig. 9.3), or multiple instances of the follower-integrator circuit (see Fig. 9.4), in sequence to form a delay line.

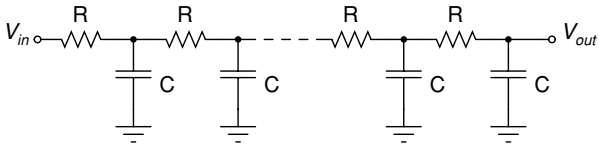


Figure 9.3
Delay line formed by connecting a large number of R-C integrator circuits.

Each section in the delay line of Fig. 9.4 is modular (from a functional point of view) and independent of the other sections. The current out of the transconductance amplifier of one section can charge only the capacitor connected to the output node: It is not affected by the other sections connected to the output node. On the other hand, the sections of the R-C delay line of Fig. 9.3 are tightly coupled to one another. Current of one section flows into both the capacitor of that section and the resistor of the next. Because the characteristics of the single R-C circuit change when connected to other R-C circuits (as in Fig. 9.3), the transfer function of the composition of R-C circuits is not equal to the composition of individual transfer functions.

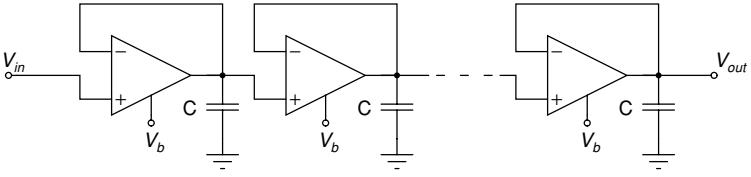


Figure 9.4
Delay line formed by connecting a large number of follower-integrator circuits.

Due to the modularity of the follower-integrator elements, the transfer function of the composition of follower-integrator circuits corresponds to the composition of their individual transfer functions. For example, if the number of elements in the delay line of Fig. 9.4 is n, we can write

$$\frac{V_{out}}{V_{in}} = \left(\frac{1}{1+\tau s}\right)^n . \tag{9.1.4}$$

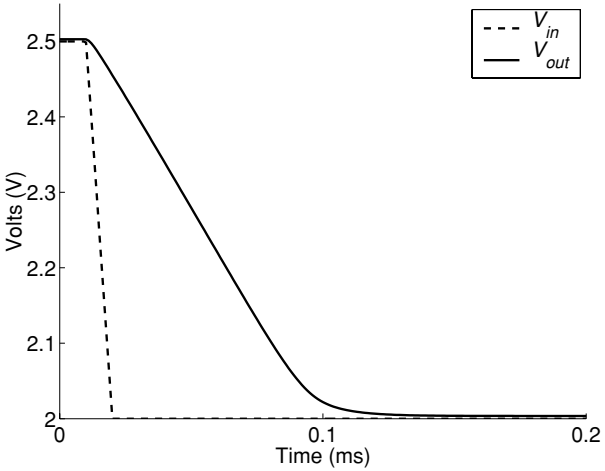


Figure 9.5
Large signal behavior of a follower-integrator. Response of the circuit to a large negative step input (dashed line). The output voltage V_{out} decreases linearly for large difference $V_{out} - V_{in}$ values and asymptotes exponentially for small differences.

This is a very useful property, because it allows the frequency response properties of the delay line to be evaluated analytically. Consider the case where sinusoidal inputs are applied to the circuit of Fig. 9.4 ($s = j\omega$). The delay line's transfer function is

$$\frac{V_{out}}{V_{in}} = \frac{1}{(1+j\omega\tau)^n}. \tag{9.1.5}$$

Exploiting the fact that

$$(1+j\omega\tau) \approx \left(1+\frac{1}{2}(w\tau)^2\right) e^{j\omega\tau} \quad (\text{for } \omega\tau \ll 1) \tag{9.1.6}$$

we can write the transfer function explicitly in terms of *magnitude* and *phase*:

$$\frac{V_{out}}{V_{in}} \approx \frac{1}{1 + \frac{n}{2}(j\omega\tau)^2}\, e^{-jn\omega\tau} \tag{9.1.7}$$

where the pre-exponential ratio is the magnitude and the exponential's argument is the phase. From this equation, we can conclude that the magnitude of the output signal is attenuated by the factor $\frac{1}{2}(\omega\tau)^2$ as it crosses each section of the follower-integrator delay line, and the phase delay introduced by each section corresponds to $\omega\tau$ radians (equivalent to a time delay of τ). This analysis is valid provided that each follower-integrator operates in its linear region.

Large Signal Behavior

When the AC component of V_{in} is large, the transconductance amplifier is no longer linear and Eq. 9.0.1 is invalid. For very large variations in V_{in}, the output current of the transconductance amplifier saturates at $\pm I_b$ (the asymptotes of Eq. 9.1.1). In this condition, the transconductance amplifier acts as a constant current source rather than as a linear conductance. While the difference $|V_{out} - V_{in}|$ is greater than $4U_T$, V_{out} changes linearly with time. As the the difference enters the small signal regime, the amplifier begins to behave as a linear conductance and V_{out} begins increasing (or decreasing) exponentially (see Fig. 9.5).

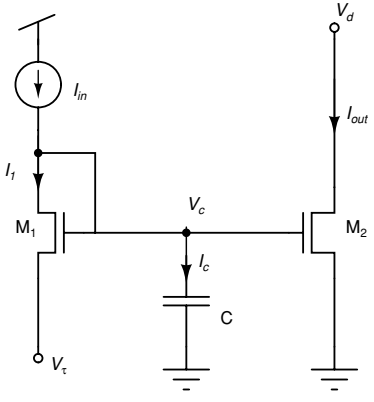


Figure 9.6
Current-mirror integrator. The current I_{in} is the input signal while the output signal is the current I_{out}.

Figure 9.5 shows the typical response of the follower-integrator circuit to a large step voltage input. The rate of change of output voltage $\frac{dV_{out}}{dt}$ in the region of constant slope, is defined as *slew rate* and is usually specified in units of $V/\mu sec$. Slew rate is one measure of the performance limit of an operational amplifier, and is proportional to the maximum output current of the amplifier.

9.2 The Current-Mirror Integrator

This circuit is a non-linear integrator. It does not have the composition property of the follower-integrator circuit but it is the advantage that it is compact, and so various forms have been used to implement dense arrays of synaptic circuits for spiking neural networks (Boahen, 1998; Häfliger and Mahowald, 1998; Indiveri, 2000).

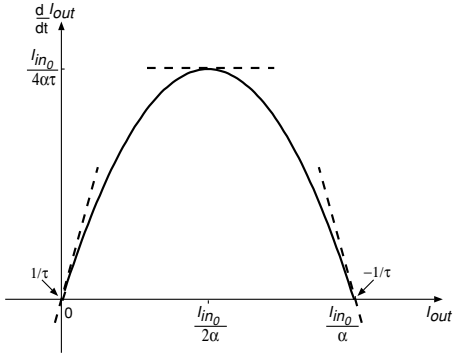


Figure 9.7
Plot of $\frac{d}{dt}I_{out}$ as a function of I_{out}, according to Eq. 9.2.3. Note the slopes of the parabola at ordinate values $I_{out} = 0$, $I_{out} = \frac{I_{in_0}}{2\alpha}$ and $I_{out} = \frac{I_{in_0}}{\alpha}$.

The circuit comprises only two transistors and one capacitor (see Fig. 9.6). Input current I_{in} is applied to M_1 and its low-pass filtered version passes through M_2. The circuit can be thought of as a diode-capacitor filter, rather than a resistor-capacitor one. As its response is not linear we cannot apply the methodology introduced in Chapter 8 to obtain analytical solutions of its response to any arbitrary input. However, we can obtain analytical solutions for responses to some typical input signals by solving the following system of

equations:

$$\begin{aligned} I_{in}(t) &= I_1(t) + I_c(t) \\ I_1(t) &= I_0 e^{\kappa \frac{V_c(t)}{U_T} - \frac{V_\tau}{U_T}} \\ I_c(t) &= C\frac{dV_c}{dt} \\ I_{out}(t) &= I_0 e^{\kappa \frac{V_c(t)}{U_T}}. \end{aligned} \tag{9.2.1}$$

For simplicity, we assume that M_1 and M_2 are identical (I_0 and κ are the same).

Of particular interest is the circuit's response to a pulse input (that could represent incoming action potentials from a silicon neuron). We shall subdivide the pulse function into two step functions: an step $I_{in}(t) = I_{in_0}u(t_0)$; and a step $I_{in}(t) = I_{in_0}(1 - u(t_1))$ (see Eq. 8.4.7 for the definition of $u(t)$).

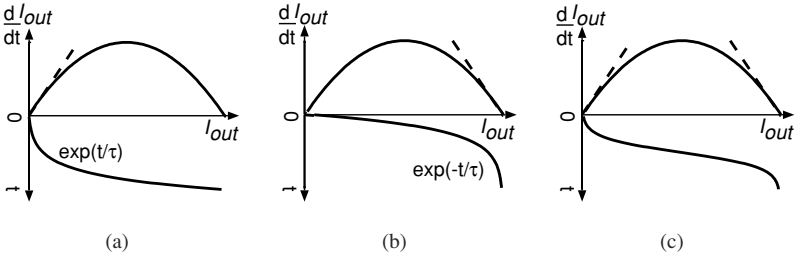


Figure 9.8
Profiles of $\frac{d}{dt}I_{out}$ as a function of I_{out}, and locally estimated profiles of I_{out} as a function of time. (a) Local estimate of $I_{out}(t)$ for values of I_{out} close to zero. (b) Local estimate of $I_{out}(t)$ for values of I_{out} close to $I_{in_0}e^{V_\tau/U_T}$. (c) Estimate of the profile of $I_{out}(t)$ for all values of I_{out}.

Case $I_{in}(t) = I_{in_0}u(t_0)$

This case is solved by analyzing how I_{out} changes with time. By using the chain-rule for differentiation:

$$\frac{d}{dt}I_{out} = \frac{d}{dV_c}I_{out} \cdot \frac{d}{dt}V_c. \tag{9.2.2}$$

Using the relationships of Eq. 9.2.1; expressing $\frac{d}{dt}V_c$ in terms of I_c; and substituting I_c with $(I_{in} - I_1)$ we obtain

$$\frac{d}{dt}I_{out} = \frac{\kappa}{CU_T}I_{out}(I_{in_0} - I_1). \tag{9.2.3}$$

When I_1 is expressed in terms of I_{out}, Eq. 9.2.3 becomes

$$\frac{d}{dt}I_{out} = \frac{\kappa}{CU_T}I_{out}(I_{in_0} - \alpha I_{out}) \tag{9.2.4}$$

where

$$\alpha = e^{-\frac{V_\tau}{U_T}} \tag{9.2.5}$$

is the circuit's gain. Equation 9.2.4 can be re-written in the simplified form

$$\frac{d}{dt}I_{out} = \frac{1}{\tau}I_{out}\left(1 - \alpha\frac{I_{out}}{I_{in_0}}\right) \tag{9.2.6}$$

where $\tau = \frac{CU_T}{\kappa I_{in_0}}$. Plotting $\frac{d}{dt}I_{out}$ as a function of I_{out} (Fig. 9.7), we obtain a parabola that intersects the abscissa at $I_{out} = 0$ and $I_{out} = I_{in_0}/\alpha$.

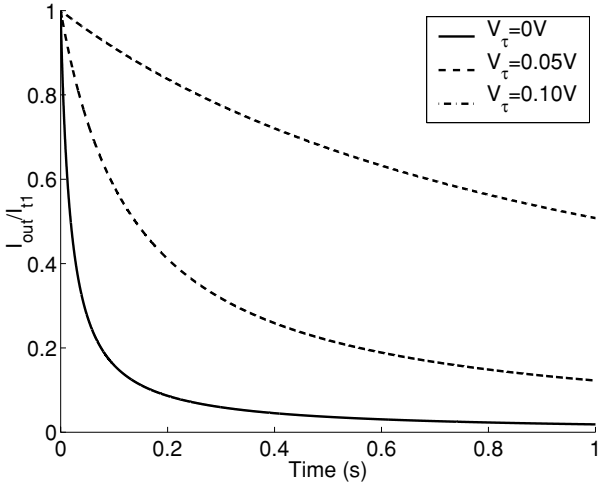


Figure 9.9
Response profiles of the current-mirror integrator to a downward current step (from I_{in_0} to 0A). The three curves show the responses for three values of V_τ (see Figure legend). The curves have been normalized to show the effect of V_τ on the time-course of the response.

The slope of the parabola at these two intersections is $\frac{1}{\tau}$ and $-\frac{1}{\tau}$ respectively. The maximum of the parabola (point of zero slope) is at $I_{out} = \frac{I_{in_0}}{2\alpha}$. These points can be used to derive the equation that describes how I_{out} changes with time. We know that for I_{out} close to zero, $I_{out}(t)$ must resemble an exponential $e^{t/\tau}$ (see Fig. 9.8(a)); for I_{out} close to I_{in_0}/α it must resemble an exponential $e^{-t/\tau}$ (see Fig. 9.8(b)); and for I_{out} close to the center of the parabola, $I_{out}(t)$ must be linear. Furthermore because the parabola is continuous, $I_{out}(t)$ must change smoothly from one profile to the other (see Fig. 9.8(c)). The form of $I_{out}(t)$ in Fig. 9.8(c) resembles a hyperbolic tangent function. We can test whether the solution contains a $\tanh(\cdot)$ function, by substituting the generic function $I_{out} = a\left[\tanh(bt + c) + d\right]$ into Eq. 9.2.6 and testing for parameters a, b, c, and d for which the equality holds.

Correct parameters do exist and they form the solution

$$\boxed{I_{out}(t) = \frac{I_{in_0}}{2\alpha}\left[1 + \tanh\left(\frac{1}{2\tau}(t - t_0)\right)\right]} \tag{9.2.7}$$

which describes the response of the circuit to a step input current from $I_{in} = 0$ to $I_{in} = I_{in_0}$ at time t_0.

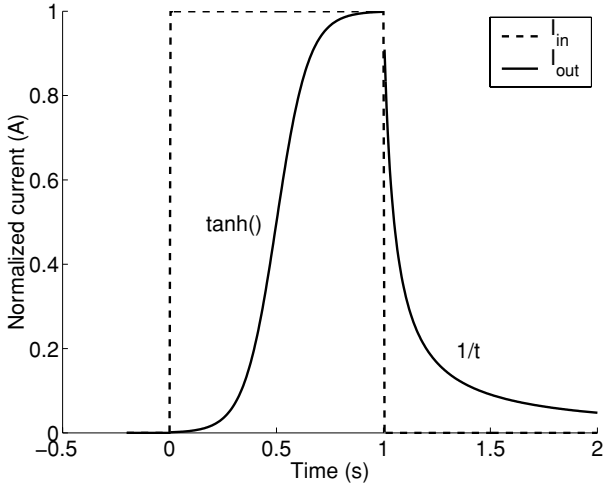


Figure 9.10
The output (solid line) of the current-mirror integrator's in response to a pulse input current (dashed line).

Case $I_{in}(t) = I_{in_0}(1 - u(t_1))$

In this case there is a step change that brings the input current from $I_{in} = I_{in_0}$ to $I_{in} = 0$ at time t_1. For $t > t_1$ we have $I_{in} = 0$. In this case the system of equations 9.2.1 can be simplified to obtain

$$C\frac{dV_c}{dt} = -I_0 e^{\kappa \frac{V_c}{U_T} - \frac{V_\tau}{U_T}}. \tag{9.2.8}$$

Integrating both sides of the equation yields

$$\int_{V_c(t_1)}^{V_c(t)} e^{-\kappa \frac{V_c}{U_T}} dV_c = -\frac{I_0}{C} e^{-\frac{V_\tau}{U_T}} \int_{t_1}^{t} dt \tag{9.2.9}$$

and by expressing $e^{-\kappa V_c}$ in terms of I_{out} (through the last relationship of Eq. 9.2.1), we obtain

$$I_{out}(t) = \frac{I_{t1}}{1 + \frac{I_{t1}\kappa}{CU_T} e^{-\frac{V_\tau}{U_T}} t} \tag{9.2.10}$$

where I_{t1} is the output current at $t = t_1$, $I_{t1} = I_{out}(t_1) = I_0 e^{V_c(t_1)/U_T}$, and $V_c(t_1)$ is the voltage generated by the charge stored on the capacitor at $t = t_1$. If we define

$$\tau = \frac{CU_T}{I_{t1}\kappa} e^{V_\tau / U_T} \tag{9.2.11}$$

then Eq. 9.2.10 can be rewritten as

$$\boxed{I_{out}(t) = \frac{I_{t1}}{1 + \frac{t}{\tau}}}. \tag{9.2.12}$$

When no input current is applied, for $t \gg \tau$ the output current decreases with a $\frac{1}{t}$ profile. The shape of the profile can be modulated (exponentially) by the control parameter V_τ. Figure 9.9 shows the profile of Eq. 9.2.10 for three different values of V_τ.

Figure 9.10 summarizes graphically the current-mirror integrator's response properties when a current pulse input is applied. This circuit is commonly used to implement models of synapses that integrate pulses (spikes). Figure 9.11 shows experimental data obtained from one of these synapses when stimulated with uniformly distributed spikes at three different frequency settings.

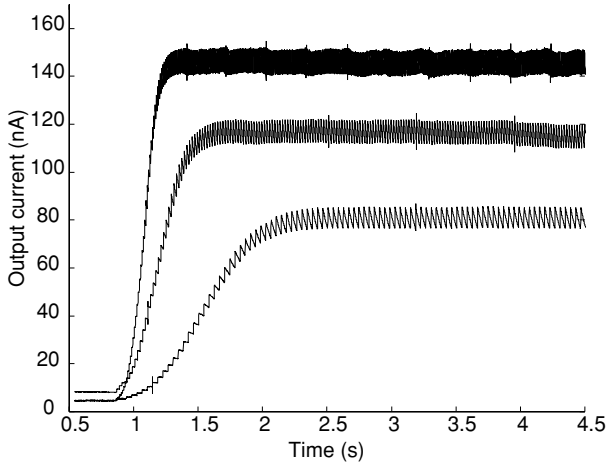


Figure 9.11
Response of a synaptic circuit based on a current-mirror integrator to input spike trains of different frequencies. The bottom trace shows the circuit's response to uniformly distributed spikes of 25Hz. The middle trace and the top trace show the circuit's response to an input spike train of 50Hz and of 100Hz respectively.

9.3 The Capacitor

The simplest differentiator is a capacitor (see Fig. 9.12). The transfer function of a capacitor is

$$I_{out}(t) = C\frac{d}{dt}(V_{in}(t) - V_{out}(t)). \tag{9.3.1}$$

This element converts an input voltage signal into an output current, computing the exact derivative of its input signal. Ideally, the capacitor should provide an infinite output current in response to an input voltage step. This ideal performance cannot be realized in physical capacitors, because they have a finite output impedance that limits I_{out}. Furthermore, if we need the same type of signal (voltage, for example) at both the input and the output of the differentiator circuit, then the output current must be converted back into a voltage. A resistor can be used for this purpose, leading to the capacitor-resistor (C-R) circuit shown in Fig. 9.13, described by

$$V_{out}(t) = RI_{out} = RC\frac{d}{dt}(V_{in}(t) - V_{out}(t)). \tag{9.3.2}$$

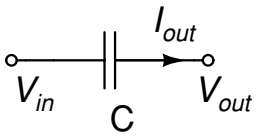


Figure 9.12
The perfect differentiator.

When a step function $V_{in} = V_0\ u(t)$ is applied (see Eq. 8.4.7), the circuit's step response is

$$V_{out} = V_0 e^{-t/\tau} \tag{9.3.3}$$

where τ is the circuit's time-constant ($\tau = RC$). The differentiator's transfer function is

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{\tau s}{1 + \tau s}. \tag{9.3.4}$$

Recalling that the operator s stands for the derivative operator $\frac{d}{dt}$, we observe that at low frequencies ($\tau s \ll 1$) $H(s) \approx \tau s$ which approximates the transfer function of an ideal differentiator. On the other hand, at high frequencies ($\tau s \gg 1$), $H(s) \approx 1$ and the circuit behaves like a unity-gain follower. The circuit is commonly referred to as a "high-pass filter" because it allows high-frequency signals to pass unaltered, while suppressing low-frequency signals. Typically, the resistors fabricated in VLSI technology are restricted to less than a few $k\Omega$. This range provides fairly good differentiator circuits, but small V_{out}. As in the case of integrator circuits with passive

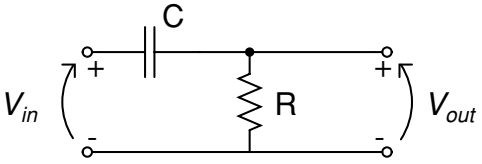


Figure 9.13
The capacitor-resistor (C-R) circuit.

resistors, differentiators of the type of Fig. 9.13 have limited flexibility when implemented using linear resistors. An adjustable time-constant requires that the differentiator circuits use active elements, such as the transconductance

amplifier, to implement the resistor component. The simplest of these circuits is the *follower-differentiator* (see Fig. 9.14).

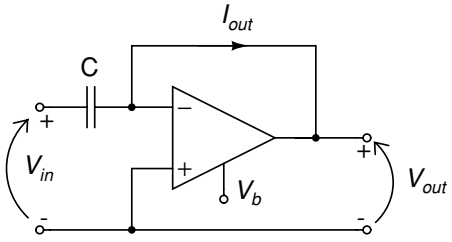


Figure 9.14
The follower-differentiator circuit.

9.4 The Follower-Differentiator Circuit

Like its cousin the follower-integrator (see Section 9.1), the follower-differentiator comprises a unity-gain follower and a capacitor. However in this case, the capacitor is connected to the input node rather than the output. Furthermore, the input signal is applied to the negative input terminal of the amplifier, while the positive terminal is connected to the reference voltage.

Intuitively we can see that the unity-gain follower tries to clamp V_{out} to the reference potential, while changes in the input signal are capacitively coupled to V_{out} and act against the clamp. The output of this circuit in response to a step input is shown in Fig. 9.15.

In the small signal regime, the behavior of the circuit is described by

$$I_{out} = GV_{out} \tag{9.4.1}$$

$$I_{out} = C\frac{d}{dt}\left(V_{in} - V_{out}\right) \tag{9.4.2}$$

where G is the amplifier's transconductance. The Heaviside-Laplace transform (Section 8.7) is used to write the circuit's transfer function:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{\tau s}{1 + \tau s} \tag{9.4.3}$$

where τ is C/G. Like the $C-R$ differentiator circuit, the follower-differentiator acts as a high-pass filter ($H(s) \approx \tau s$ for $\tau s \ll 1$ and $H(s) \approx 1$ for $\tau s \gg 1$).

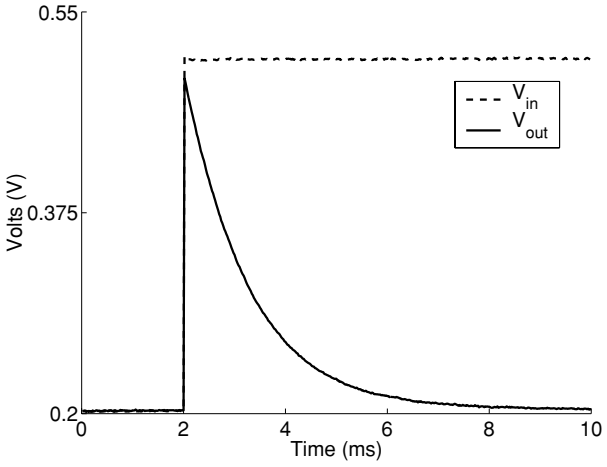


Figure 9.15
Step response measured from a follower-differentiator circuit.

This circuit permits τ to be adjusted over several orders of magnitude, but its output saturates to $\pm I_b$ for large input variations leading to distortion in the circuit's response.

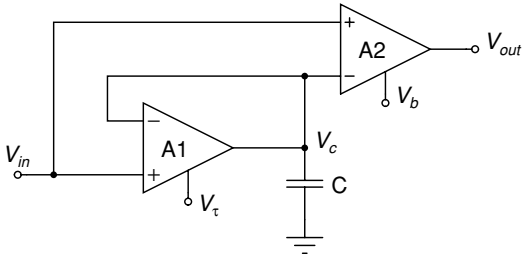


Figure 9.16
The diff1 circuit.

9.5 The diff1 and diff2 Circuits

These two types of circuits, previously described in Mead (1990), exploit the fact that an approximation of a temporal derivative can be obtained by

comparing a signal to its time-averaged version (for example, by subtracting a low-passed version of the signal from its original version). The most direct way to compute a time-averaged of a signal, using VLSI circuits, is to use the follower-integrator circuit of Section 9.1. The simplest way to subtract two voltage signals is to apply them to a transconductance amplifier: $V_{out} = A(V_{in_+} - V_{in_-})$ (where A is the amplifier's voltage gain). The diff1 circuit, shown in Fig. 9.16, does exactly this: It subtracts the low-passed version of the input signal V_c from the input signal V_{in}. The problem with this arrangement is that amplifier A2 is in an open-loop configuration. Consequently, if the open circuit voltage gain is high (as is typically the case - see Section 5.3), then any small input offset will be greatly amplified and the output voltage of the diff1 circuit will be clamped at one of the power-supply rails, even with steady-state inputs.

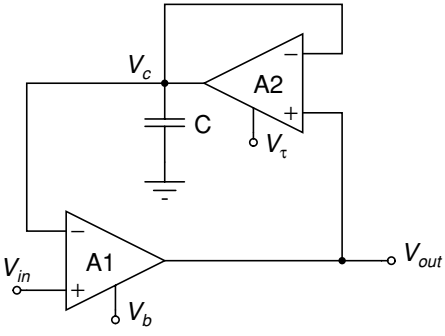
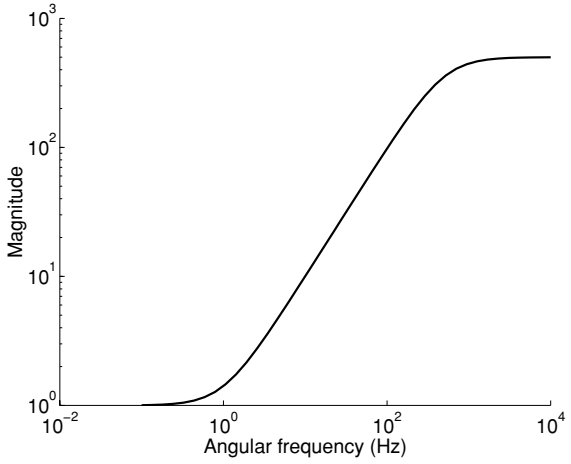


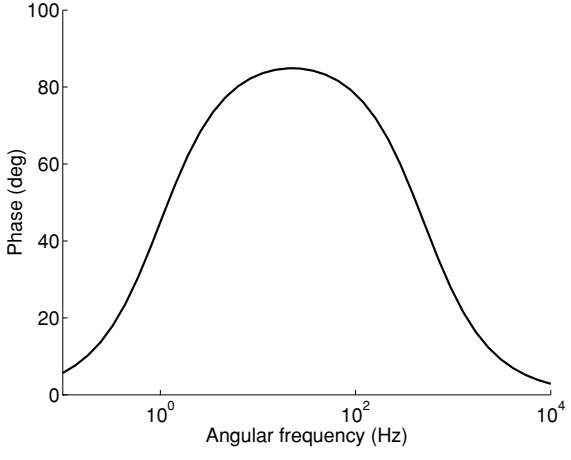
Figure 9.17
The diff2 circuit.

On the other hand, the diff2 circuit has both of its amplifiers arranged in a negative-feedback configuration, and subtracts the time-averaged version of the *output* voltage from the input (see Fig. 9.17). In this case, amplifier A2 is configured as a follower-integrator and its output, which is the low-passed version of V_{out}, is fed back to amplifier A1. For constant and slowly varying signals, this circuit is simply a unity-gain follower (see Section 5.3): The negative feedback on A1 drives V_{out} to values very close to V_{in}. In this case, offset voltages are multiplied by unity gain $(A/(A+1))^2$ rather than by the amplifier's open loop gain A.

2 This is close to unity if V_b is biased properly (*that is*, voltage offsets are not amplified).



(a)



(b)

Figure 9.18
Bode plot of the diff2 circuit's transfer function. (a) Magnitude (b) Phase.

The diff2 circuit's characteristic equations are

$$V_{out} = A(V_{in} - V_c) \tag{9.5.1}$$

$$V_c = \frac{1}{1+\tau s} V_{out} \tag{9.5.2}$$

where A is the open circuit voltage gain of A1, $\tau = C/G$, and G is the transconductance of A2. Substituting Eq. 9.5.2 into Eq. 9.5.1 we obtain

$$\frac{V_{out}}{V_{in}} = \frac{A}{A+1} \quad \frac{1+\tau s}{1+(\tau/(A+1))\, s}. \tag{9.5.3}$$

This transfer function can be simplified for the following three domains:

$$\frac{V_{out}}{V_{in}} \approx \begin{cases} \frac{A}{A+1} & \text{for } \tau s \ll 1 \\ \frac{A}{A+1}\, \tau s & \text{for } \frac{\tau s}{A+1} \ll 1 \ll \tau s \\ A & \text{for } \frac{\tau s}{A+1} \gg 1. \end{cases} \tag{9.5.4}$$

The frequency response of this circuit can be found by setting $s = j\omega$ and evaluating the magnitude and phase of the circuit's transfer function. The

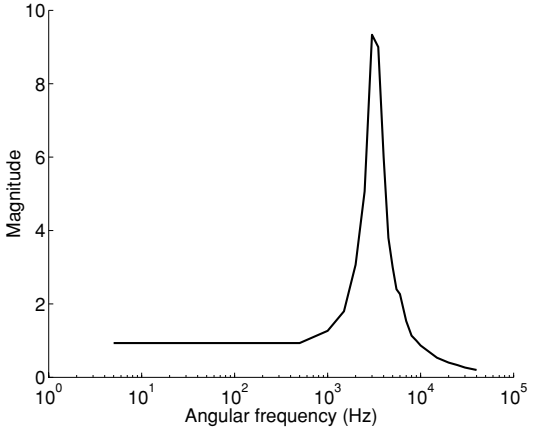


Figure 9.19
Actual frequency response measured from a diff2 circuit.

Bode plot of Fig. 9.18 illustrates the approximations made in Eq. 9.5.4: At low frequencies the diff2 circuit acts as a unity-gain follower; at intermediate frequencies as a differentiator; and at high frequencies as an amplifier.

The transfer function of Eq. 9.5.3 (and the corresponding frequency response of Fig. 9.18) does not take into account the effect of parasitic capacitances that are present in physical implementations of the circuit. The measured frequency response curve from a real diff2 circuit is shown in Fig. 9.19. The main difference between this plot and the one obtained from a first order analysis of the circuit, is the distinct resonant peak at high frequencies. As the input frequency increases, rather than smoothly saturating to the gain A of the amplifier, the response first peaks and then decreases well below A. This behavior is largely due to the parasitic capacitance at V_{out}.

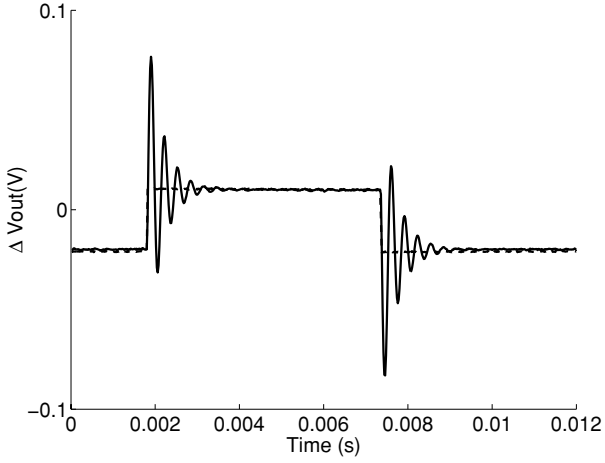


Figure 9.20
Response of a diff2 circuit (solid line) to a small (ΔV_{in}=30 mV) step input signal (dashed line).

Small-Signal Step Response

Figure 9.20 shows the response of the diff2 circuit to a small step input ($\Delta V_{in} = 30$mV). The circuit's response has high transient gain, linear decay and ringing. The transient gain is high because there are frequency components in the input signal high enought that $\frac{\tau s}{A+1} \gg 1$ (see Eq. 9.5.4). On the other hand, the linear decay on the other hand is a consequence of the slew-rate limited of the follower-integrator in the diff2 circuit (see Section 9.1). Finally, the ringing is present due to the same parasitic capacitance on the V_{out} node that caused the resonant peak in Fig. 9.19.

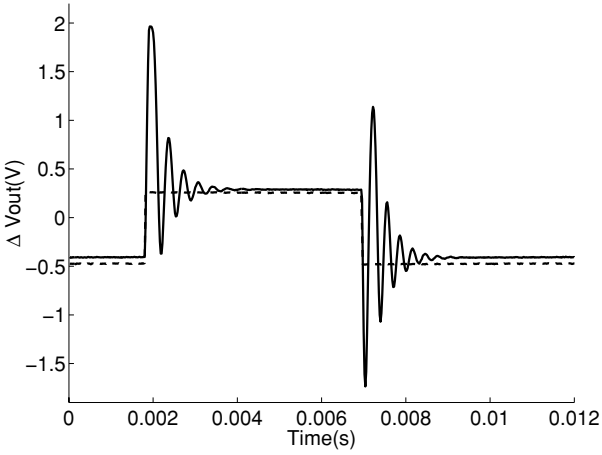


Figure 9.21
Response of a diff2 circuit (solid line) to a large (ΔV_{in}=600 mV) step input signal (dashed line).

Large Signal Step Response

If the gain A of the amplifier is of several hundred, a change of V_{in} by a value greater than a few millivolts brings V_{out} to the power supply rails. The response peaks are clipped by these limits (see Fig. 9.21). The voltage swing would be much higher, if the transient gain was not limited by the power supply rails. Except for the regions where the voltage is clipped, the profile of the circuit's response to large input steps is qualitatively equivalent to the one obtained for small input steps: there is a high transient gain, a linear decay and ringing.

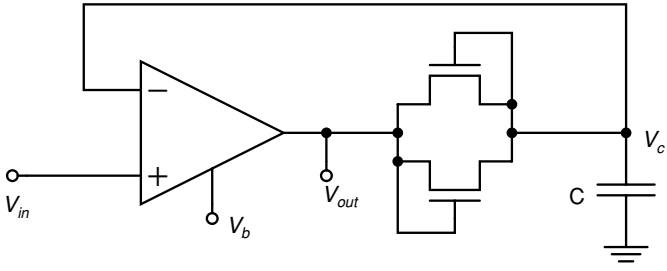


Figure 9.22
Hysteretic differentiator circuit.

9.6 Hysteretic Differentiators

The unity-gain follower in the diff2 circuit can be regarded as a resistive element with a linear region for small signals, and a *compressive non-linearity* for larger signals. An entirely different characteristic is obtained if this element is replaced by an element with an *expansive non-linearity*, in which the resistance is large for small signals and gets smaller for larger ones. Elements of this type, that have exponential current-voltage characteristics in both directions, can be constructed with simple MOSFET circuits. An obvious solution is a configuration of two antiparallel diode-connected MOSFETs of the same type (Fig. 9.22) (Mead, 1989). Another implementation of a bidirectional exponential element, involving a single MOSFET and parasitic bipolar junction transistors, will be discussed in Section 10.4.

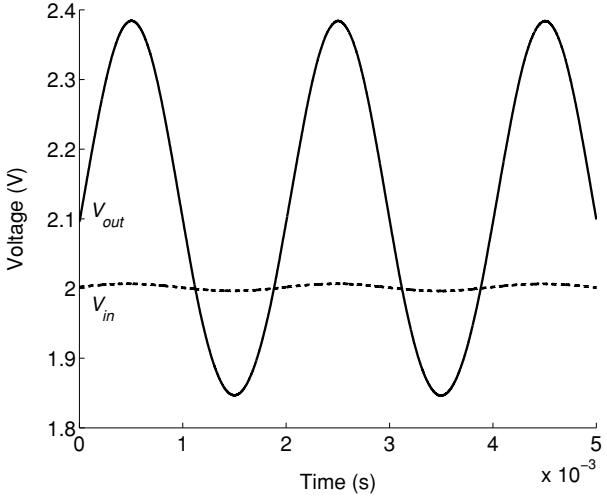


Figure 9.23
Small-signal sine-wave response of the hysteretic differentiator, showing open-loop amplification.

A hysteretic differentiator with an exponential resistive element can exhibit a variety of different characteristics depending on the frequency and amplitude of the input signal. Due to the strong non-linearity of the resistive element, the time-constant of the feedback loop varies over several orders of magnitude as a function of the voltage difference between the output terminal and the

feedback terminal. Small voltage differences result in large time-constants and large voltage differences in small time-constants.

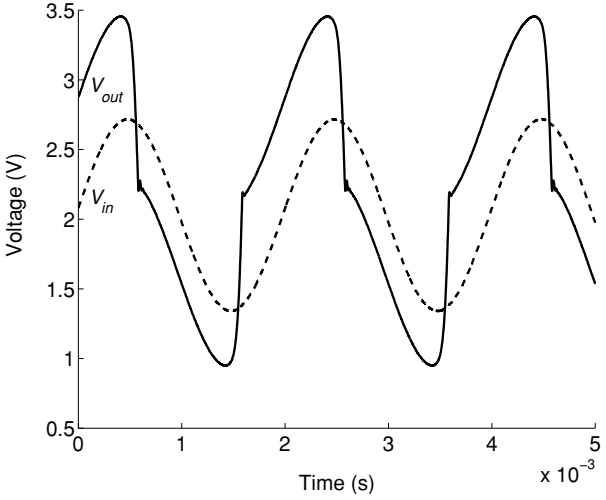


Figure 9.24
Large-signal sine-wave response of the hysteretic differentiator, showing hysteretic following behavior.

We will first examine the circuit characteristics for a sine-wave input. For small signals and high frequencies the time-constant of the feedback loop is much larger than that of the input signal, and the transconductance amplifier acts as an open-loop circuit in which the signal is amplified linearly by the open-loop voltage gain of the amplifier, as plotted in Fig. 9.23. If the frequency is lowered, or the amplitude increased, then the voltage on the feedback node will begin to follow the input voltage. Since the time constant is strongly dependent on the voltage difference between the output and the feedback node, V_c will lag behind V_{in} for small differences and then suddenly catches up when V_{out} has a large enough excursion. These dynamics result in a following-behavior with possible overshoots during the catch-up phase. The following-behavior occurs whenever V_{out} and V_c are sufficiently different that the time-constant of the feedback circuit is small. During following, V_{out} thus has an offset with respect to V_{in}, whose sign depends on the direction of the current flow, which in turn depends on whether V_{in} is increasing or decreasing. This jump of the DC offset in the following-mode as a function

of the direction of the current flow can be regarded as hysteretic behavior and gives the circuit its name. For low frequencies and high amplitudes the overshoots disappear and the circuit remains in the hysteretic following mode with the signal discontinuities appearing whenever the sign of the derivative of the input signal changes. This situation is shown in Fig. 9.24.

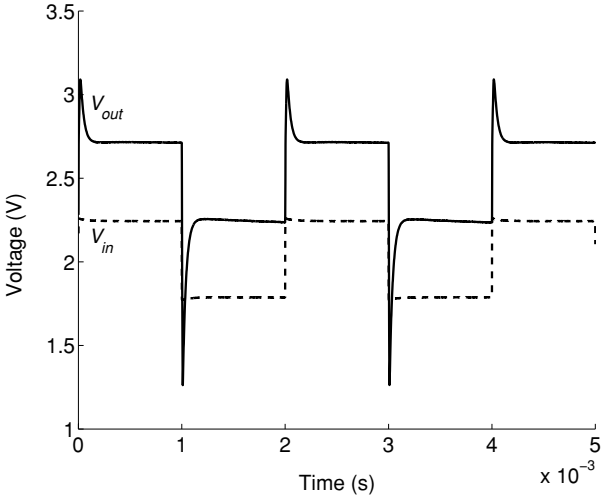


Figure 9.25
Large-signal square-wave response of the hysteretic differentiator, showing differentiating and following behavior. The asymmetric behavior and the DC offset are due to the use of an asymmetric resistive element in the circuit.

For square-wave inputs, small signals are also amplified by the open-loop gain, while larger signals are followed with overshoots at the transitions (Fig. 9.25). These overshoots can be regarded as a differentiating characteristic. The data plotted in Figs. 9.23−9.25 were obtained from a hysteretic differentiator with an asymmetric resistive element (see Fig. 10.11(a) in Section 10.4). This asymmetry explains the DC offset between input and output signals and the asymmetric response to square-wave input signals.

An interesting variant of the circuit is obtained by replacing the resistive element with a bidirectional source-follower circuit (Fig. 9.26) (Mead, 1989). This element has a very high impedance at the V_{out} node: The currents to charge and discharge the capacitance at the feedback node are obtained from the power rails via the source followers, and not from the transconductance

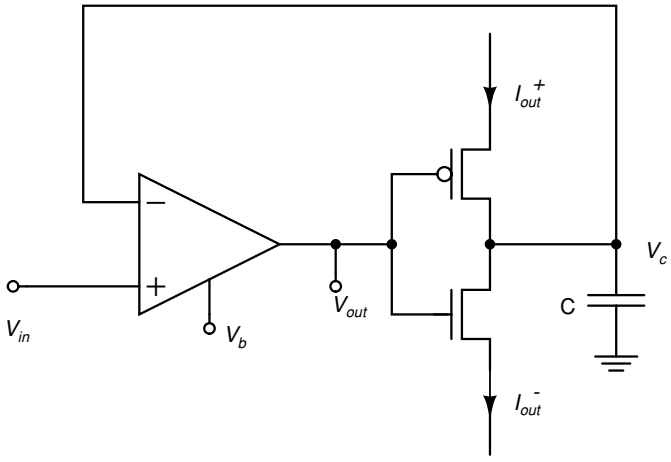


Figure 9.26
Hysteretic differentiator circuit with rectifying current outputs.

amplifier. Thus, the circuit has a faster large-signal response than the circuit of Fig. 9.22. Furthermore, if the currents I_{out}^{+} and I_{out}^{-} are used as output signals the circuit acts a rectifying temporal differentiator (Kramer et al., 1997) with

$$C\frac{d}{dt}V_{in} \approx \begin{cases} I_{out}^{+} & \text{for } \frac{d}{dt}V_{in} > 0 \\ I_{out}^{-} & \text{for } \frac{d}{dt}V_{in} < 0\,. \end{cases} \tag{9.6.1}$$

The output signals for positive and negative temporal transients appear at two different terminals, both of which has very small DC responses. This property reduces DC offsets due to variations in fabrication parameters, which often limit the performance of analog integrated circuits. In addition, the responses are thresholded, because a substantial excursion of V_{out} is required to draw a significant current at either terminal. This property is useful for suppression of low-amplitude temporal noise introduced by thermal effects and statistical fluctuations in the input signal.

This page intentionally left blank

10 Photosensors

Photosensors convert electromagnetic radiation into a different physical form, usually electrical charge. The growth in the market for optical communication and electronic imaging has brought a hugh importance to the photosensor as optoelectronic interfaces. Consequently, a wide range of photosensors designed for different applications are available. Here, we will consider only photosensors that can be fabricated with semiconductor processes used for the implementation of transistor-based electronic devices.

10.1 Photodiode

A *photodiode* is a diode that is used as a photosensor. Commercial photodiodes are specifically tailored to different application domains. However, since the diode forms part of every transistor, photodiodes are also available in standard semiconductor processes, and so they can be monolithically integrated with electronic circuitry.

In a semiconductor, an incident photon and therefore its energy can be absorbed by an electron; a process known as the *inner photo-electric effect*. A photon with an energy larger than or approximately equal to the bandgap energy can excite an electron from the valence band into the conduction band. In the energy-band diagram, this process corresponds to the generation of an electron-hole pair. Illumination of a semiconductor therefore increases the concentration of mobile charge carriers above the thermal equilibrium value in the exposed area. If the motion of the carriers is driven by diffusion, then the generation is balanced by recombination. However, in the presence of an electric field, electrons and holes can separate and some of the separated carriers contribute to an electrical output signal. This phenomenon is called *photoconduction*. The simplest device that implements the phenomenon is the *photoconductor*, which is a slab of semiconductor in an externally applied electric field. Photoconductors exhibit a large *dark current*, which is a background current which is still present in the absence of optical stimulation. This current is due to the relatively large doping concentration used in most modern semiconductor processes. This doping results in high conductivity values and a poor signal-to-noise ratio of photoconductors.

A diode is a much more suitable photosensor, because it has a depletion region with a low conductivity and a built-in electric field. The presence of

a depletion region substantially reduces the dark current, while the built-in electric field in the depletion region performs charge separation even in the absence of an externally applied voltage. Electron-hole pairs generated in the depletion region and within a diffusion length of it, are likely to be separated (Fig. 10.1). The resulting reverse current component is called *photocurrent*. As in the photoconductor, an incident photon cannot contribute more than one electron to the photocurrent.

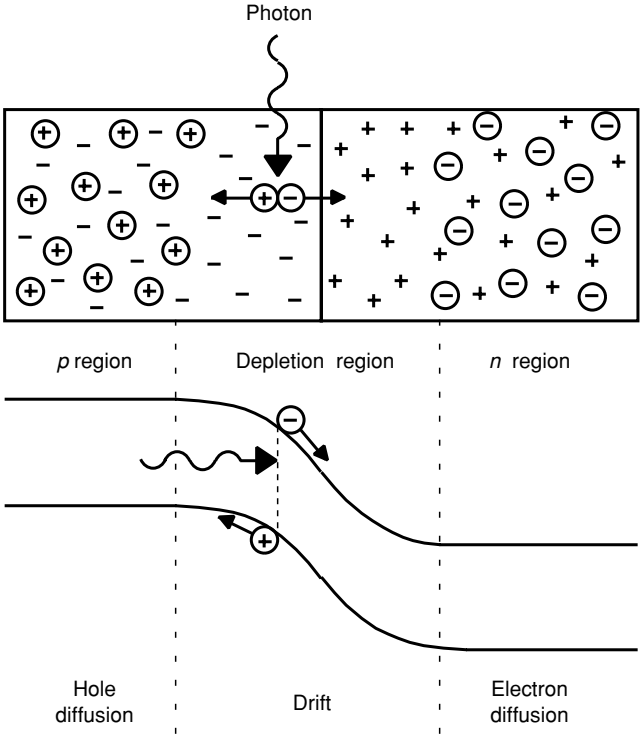


Figure 10.1
Principle of operation of a photodiode. Electron-hole pairs generated by incident photons in or within a diffusion length outside the depletion region become separated and contribute to a reverse generation current.

If the photodiode is open-circuited, that is, no external current is allowed to flow, then generated charge accumulates at the boundaries of the depletion region until a steady state is attained. In steady state, a forward bias appears

across the junction, such that the photocurrent is balanced by a forward diffusion current. If the photodiode terminals are short-circuited the photocurrent can be measured as a reverse diode current. In the presence of an applied external bias the photocurrent is superimposed onto the diode current discussed in Chapter 2. The current-voltage characteristic of a photodiode has the same shape as that of a normal diode, but the curve is displaced along the current axis by the value of the photocurrent I_{ph} (Fig. 10.2).

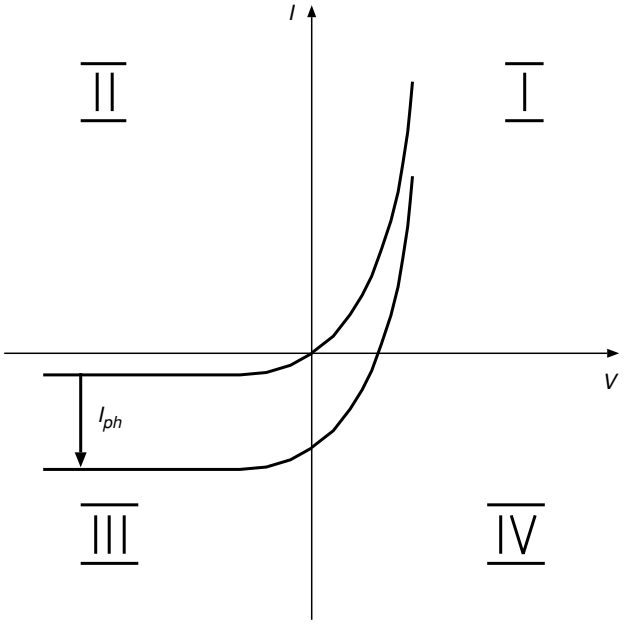


Figure 10.2
Steady-state current-voltage characteristics of a photodiode. The upper curve is the normal diode characteristic (dark characteristic). The lower curve shows the characteristic under illumination. Photodiodes are usually operated either in quadrant III as photosensors, or in the quadrant IV as solar cells.

A photodiode has two principal modes of operation, depending on its application. If the photodiode is used to convert optical power into electrical power it is called a *solar cell* and operated in the *photovoltaic mode*. In this mode, a load is connected between the two terminals, such that a reverse current flows through the diode in the presence of a forward voltage. A solar cell is thus operated in the fourth quadrant of the current-voltage characteristic.

The generated power is given by the product of the reverse current and the forward voltage. Commercial solar cells are complicated devices, which are optimized with respect to optical-to-electrical power-conversion efficiency that is typically between 10% and 20%[1].

In the other mode of operation, the *photosensing mode*, the photodiode is used to estimate the photon flux. In steady state, the diode is typically either open-circuited and the forward voltage is read out or a reverse (or zero) bias is applied to the diode and the reverse current is read out. In the latter case, the photodiode is operated in the third quadrant of the current-voltage characteristic. Here, the photodiode is quite a good current source, because the photocurrent is almost independent of the applied reverse bias.

In electronic imaging applications, the outputs of an array of photodiodes are read out in sequence onto a shared wire. In these applications, the photogenerated charge is usually integrated while the photodiode is not addressed. During readout the charge is either transferred as a current pulse onto the output wire or the voltage change across the diode due to the accumulated charge is sensed and the charge is drained after readout before a new integration cycle is started.

Photodiode Characteristics

In order to compute the photodiode current we have to consider how light is absorbed in the material. At any given position in the semiconductor the number of absorbed photons is proportional to the total number of photons for a given photon energy. This results in an exponential decrease of the photon flux Φ per unit area of the material with penetration depth x, according to

$$\Phi(x) = \Phi_0 e^{-\alpha x} \tag{10.1.1}$$

where Φ_0 denotes the flux of photons per unit area penetrating the semiconductor surface and α is called *optical absorption coefficient*. The optical absorption coefficient is the inverse of the distance over which the photon flux is reduced to a fraction of $1/e$ from its initial value. Figure 10.3 shows the optical absorption coefficients and light penetration depths for different photosensing materials as a function of optical wavelength. Photons with larger wavelengths penetrate deeper, because they have less energy and are thus less likely to generate an electron-hole pair at any given location in the material.

1 High-efficiency solar cells are built with a layering of different semiconductors of different bandgaps.

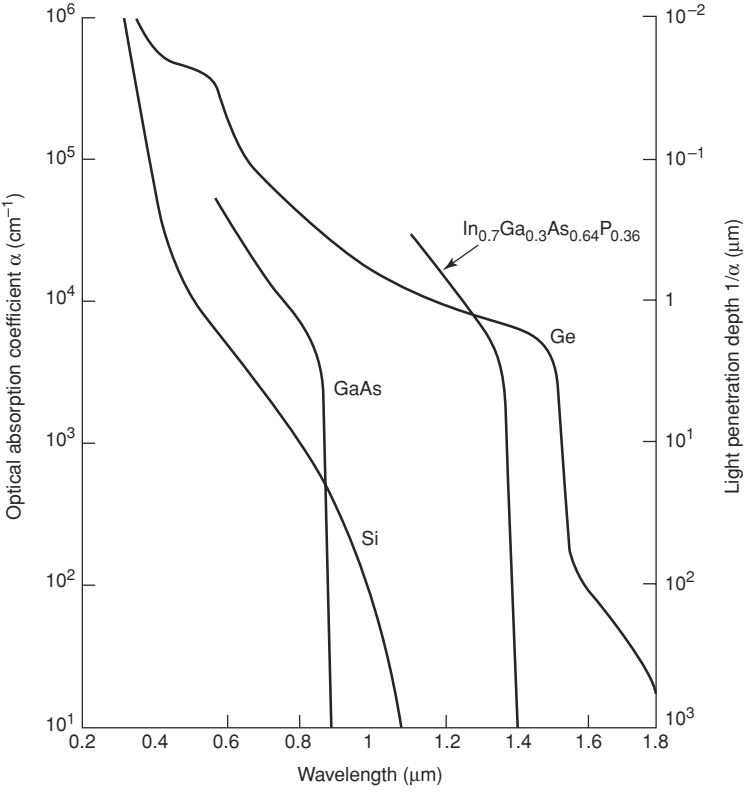


Figure 10.3
Optical absorption coefficients for different semiconductor crystals at room temperature as a function of wavelength of the incident light. Figure adapted from H. Melchior (1972), Demodulation and photodetection techniques, in Laser Handbook, Vol. 1, 725-825. ©1972, with permission from Elsevier Science.

The wavelength at which the absorption coefficient drops to zero is called the *cutoff wavelength* λ_c and corresponds to a photon energy that equals the *bandgap energy*. For silicon, $\lambda_c = 1.1\ \mu$m, which is in the near infrared. The term Φ_0 can be computed from the incident optical power P_{opt} as

$$\Phi_0 = \frac{1-R}{A}\frac{\lambda}{hc}P_{opt} \tag{10.1.2}$$

where R is the reflection at the semiconductor surface, A is the cross-section of the semiconductor perpendicular to the direction of the photon flux, and hc/λ is the photon energy as given by the photon wavelength λ, the Planck constant

h, and the speed of light c. The generation rate G for electron-hole pairs can be computed from the attenuation of the photon flux as

$$G(x) = -\frac{d\Phi(x)}{dx} = \alpha\Phi_0 e^{-\alpha x}. \tag{10.1.3}$$

The photocurrent density J_{ph} consists of a drift component $J_{ph,drift}$ due to carriers generated in the depletion region and a diffusion component $J_{ph,diff}$ due to carriers generated outside the depletion region. We can compute these components under the assumptions that the thermal generation current is much smaller than the photocurrent and that the neutral semiconductor layer the light has to cross to reach the depletion region is much thinner than $1/\alpha$. For the drift component we obtain

$$J_{ph,drift} = -q\int_0^W G(x)dx = q\left(\Phi(W) - \Phi(0)\right) = -q\Phi_0\left(1 - e^{-\alpha W}\right) \tag{10.1.4}$$

where W is the depth of the depletion region in the direction of the photon flux. The diffusion current density can be computed from the continuity equations (Eqs. 2.5.20 and 2.5.21) under the Shockley approximation stated in Chapter 2. It consists of the reverse diffusion current density under dark conditions, which for a sufficiently large reverse bias is computed from the Shockley equation (Eq. 2.6.22) to be $-J_s$, and the photogenerated contribution from the bulk region, which is obtained from the one-dimensional diffusion equation as

$$J_{ph,diff} = -q\Phi_0\frac{\alpha L}{1+\alpha L}e^{-\alpha W} \tag{10.1.5}$$

where L is the minority carrier diffusion length in the bulk. The total photodiode current density is then given by

$$J = J_{ph} - J_s = J_{ph,drift} + J_{ph,diff} - J_s = -q\Phi_0\left(1 - \frac{e^{-\alpha W}}{1+\alpha L}\right) - J_s\,. \tag{10.1.6}$$

Note that J depends only on the applied reverse bias via the width of the depletion region W. For an ideal photodiode J_s, as given by Eq. 2.6.23, is the only contribution to the dark current density. In a real photodiode, the dark current density is significantly larger than J_s and depends somewhat on the applied reverse bias. However, the operating conditions of photodiodes are usually chosen such that the photocurrent is much larger than the dark current. The reverse current density is then proportional to the incident photon flux Φ_0 per unit area. This linear relationship is expressed by the definition of the

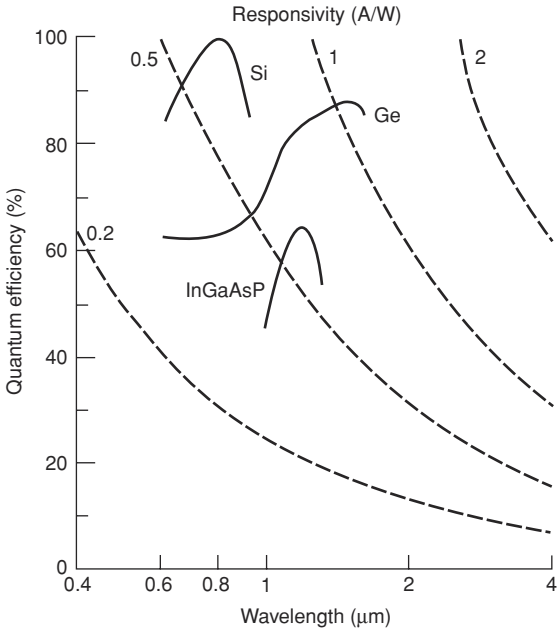


Figure 10.4
Quantum efficiencies and responsivities of photosensors fabricated from different semiconductors. Silicon exhibits a very good quantum efficiency peaking in the near infrared. Also shown are lines of equal responsivity. For a given responsivity the relationship between quantum efficiency and wavelength is inverse, as can be seen from Eq. 10.1.7. Figure adapted from S. M. Sze (1981), Physics of Semiconductor Devices, 2nd Edition. ©1981 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.

quantum efficiency η as the number of electron-hole pairs contributing to the photocurrent generated per incident photon. The quantum efficiency can be computed from Eqs. 10.1.2 and 10.1.6 as the ratio of the generated charge carrier flow density J_{ph}/q to the incident photon flux density $\lambda P_{opt}/Ahc$ as

$$\eta = \frac{Ahc}{q\lambda}\frac{J_{ph}}{P_{opt}} = \frac{hc}{q\lambda}\frac{I_{ph}}{P_{opt}} = (1-R)\left(1-\frac{e^{-\alpha W}}{1+\alpha L}\right) \tag{10.1.7}$$

where $I_{ph} = AJ_{ph}$ is the photocurrent. The ratio of photocurrent to incident optical power I_{ph}/P_{opt} is called *responsivity*. Typical quantum efficiencies and responsivities of photosensors fabricated with different semiconductor materials are shown in Fig. 10.4. Silicon has a very good quantum efficiency in the visible and near infrared, which may approach 100% in a certain spectral

range. The corresponding responsivity is on the order of 0.5 A/W. The above analysis does not take into account that charge carriers generated near the semiconductor surface are likely to recombine due to surface effects. Photocurrent, quantum efficiency, and responsivity for blue light are therefore much lower in practice than expected from the above formulas.

Types of Photodiodes

Photodiodes designed to be operated in the continuous-current photosensing mode are usually optimized with respect to their quantum efficiency and their response time, which are two partly conflicting requirements. The quantum efficiency can be optimized by applying an anti-reflection coating to the semiconductor surface to reduce R and by generating a thick depletion region close to the surface. The response time can be kept small by minimizing the junction capacitance, the carrier transit time through the depletion region, and the carrier diffusion time to the depletion region. A small junction capacitance requires a thick depletion region, while a short transit time favors a thin depletion region and a large drift velocity. Diffusion times to the depletion region can be minimized if the depletion region extends close to the surface. It is thus advantageous to operate a photodiode at a large reverse bias in order to increase the depletion region width and the drift velocity. The depletion region width can also be increased by having a low impurity-doping concentration in the junction region, as is done in most commercially-available photodiodes. Such photodiodes are known as *p-i-n photodiodes*, because they have a (nearly) intrinsic region between the *n*-type and *p*-type region.

In the photodiode operation range described so far, the quantum efficiency is smaller than unity: Each photon cannot produce more than one electron-hole pair. However, if a diode is operated in the avalanche multiplication regime in the vicinity of reverse junction breakdown, the photogenerated carriers multiply in the depletion region due to impact ionization and the quantum efficiency can be significantly larger than unity. *Avalanche photodiodes* are photodiodes designed to be operated in this domain. They have small response times and better signal-to-noise ratios than normal photodiodes. A normal semiconductor process provides very poor avalanche photodiodes with instability and matching problems. A photosensor with an internal gain mechanism that may be implemented more efficiently with standard processes is the *phototransistor*, described in the following section.

10.2 Phototransistor

A phototransistor is usually fabricated as a *bipolar junction transistor (BJT)* structure. One type of phototransistor that is available in standard CMOS processes is the parasitic BJT formed by a source/drain implant as the emitter, a well as the base, and the substrate (or another source/drain implant) as the collector.

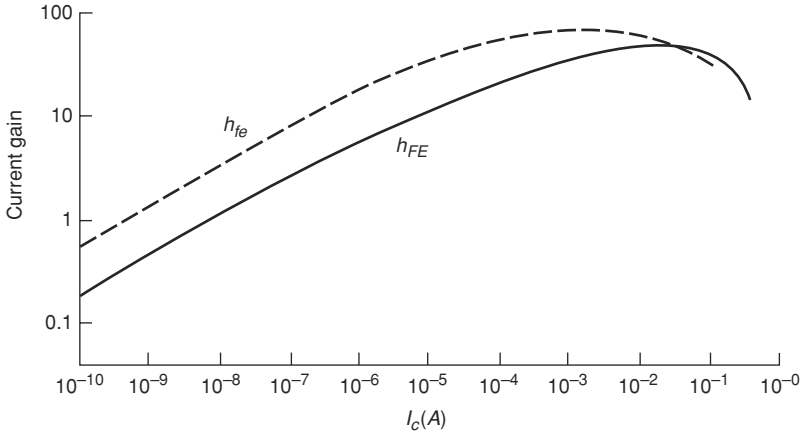


Figure 10.5
Static common-emitter current gain h_{FE} and small-signal common-emitter current gain h_{fe} versus collector current I_C of a phototransistor. Figure adapted from A. S. Grove (1967), Physics and Technology of Semiconductor Devices.

In such a bipolar phototransistor, the photon-generated electron-hole pairs are separated in the depletion region of the reverse-biased base-collector junction, resulting in a collector current that is the photodiode current I_{pd} of the junction. In a phototransistor, the base is floating and its potential is determined by the concentration of photogenerated majority carriers in the base, and therefore by the photon flux. This irradiance dependence of the base voltage causes a light-induced modulation of the collector current. An additional contribution to the collector current is thus obtained by carrier injection from the emitter through the base. This component is an amplification of the photo-

diode current by the common-emitter current gain h_{FE}[2], which is typically on the order of 100. The resulting total photocurrent is

$$I_{ph} = (1 + h_{FE})I_{pd}. \tag{10.2.1}$$

Phototransistors are useful in applications where significant leakage currents generated by the connected circuitry have to be overcome or large capacitive loads have to be driven directly by the photocurrent, because they amplify a typical photodiode current in the pA range into the nA range. However, the phototransistor has several disadvantages with respect to the photodiode. It uses more area and its response is not linear with the photon flux, because h_{FE} depends on the current level, as shown in Fig. 10.5. In addition, it has an inherently slow response because the base potential changes with illumination. This necessitates the charging and discharging of the large base-collector capacitance, while in a photodiode the voltage at both terminals can be clamped, making it typically about two orders of magnitude faster than a phototransistor, as we will see in the next section. Furthermore, the amplification mechanism degrades the signal-to-noise ratio of the phototransistor with respect to that of the photodiode, and it worsens with increasing gain.

10.3 Photogate

The Metal-Insulator-Semiconductor (MIS) structure discussed in Chapter 2 can be used as a photosensor, where one type of photogenerated charge carriers is collected in a depletion region underneath the conductor plate, while the other type leaves the semiconductor via the substrate. In this application, the MIS structure is referred to as a *photogate*.

Since in most cases the light impinges onto the structure through the conductor plate, the quantum efficiency, particularly for blue light, is reduced by absorption in and reflection at the conductor plate. This effect is mimimized by using thin polysilicon as conductor plates instead of metal.

As opposed to the photodiode, the photogate is by nature an integrating photosensor, which accumulates the charge until it is read out by a change in the applied voltages. It is therefore not surprising that it is mainly used in image sensors. As a matter of fact, it is the basis of the most common image sensor, the Charge-Coupled Device (CCD), where the MIS structure is not only used for charge integration but also for charge transportation by appropriate

2 The common-emitter gain is often called β.

clocking of the bias voltages applied to neighboring MIS gates, as we will see in Section 10.5.

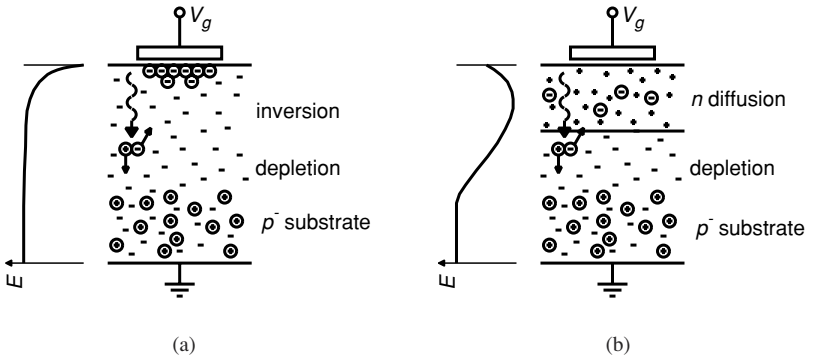


Figure 10.6
Cross-sections of photogates with (a) surface charge storage and (b) bulk charge storage. The energy E of the conduction band edge in the semiconductor as a function of depth is shown on the left of each cross- section.

The simplest photogate collects the charge carriers in a uniformly doped semiconductor substrate. The gate voltage is chosen such that the semiconductor surface is depleted of majority carriers. In the depletion region, and within a diffusion length from it further in the the bulk the photogenerated electron-hole pairs are likely to get separated such that the minority carriers drift to the surface and the majority carriers to the bulk, as depicted in Fig. 10.6(a) for a p-type substrate. The minority carriers stay at the surface as an inversion layer, but due to structural imperfections at the semiconductor-insulator interface caused by lattice mismatches and impurities, some of them get trapped in interface states with different time constants, which causes smearing effects when the collected charge is removed from underneath the gate electrode. In applications where high bandwidth is required and where the charge needs to be transported over large distances, as in the case for CCDs, it is therefore desirable to collect the charge in a region removed from the interface, because the density of trapping centers is much smaller in the bulk than at the surface.

Such a charge distribution is accomplished by collecting the charge carriers in a depleted diffusion layer between the interface and the bulk with a doping

opposite from the bulk doping. The collected charge carriers are majority carriers in this diffusion layer and replenish the depleted states, as shown in Fig. 10.6(b) for an *n*-type diffusion layer. Biasing of such a bulk-storage device is a little bit trickier than for a surface storage device. In the case of an *n*-type diffusion layer, for example, the diffusion layer has to be biased with a much more positive potential than the gate and the bulk to achieve depletion at the semiconductor surface and at the junction between diffusion layer and bulk. The charge carriers then collect within the diffusion layer between the two depletion regions. Furthermore, the collected charge has to be laterally confined to prevent it from leaking directly to the diffusion contacts, as in a photodiode. This can be done with additional gate electrodes at nearby locations that are held at a lower potential than the charge-collecting electrode, or with lateral *p*-type diffusions. While surface storage has a large charge-retaining capacity, the maximum charge-storing density of a bulk storage device cannot exceed the doping density of the diffusion layer. Additional photo-generated charge spills over to neighboring locations and the diffusion contacts. Despite this disadvantage, bulk storage is much more common in commercial CCDs than surface storage due to the reduction of the above-mentioned trapping effects.

Surface-storage photogates can be fabricated with the same silicon-processing technology as standard MOSFETs, but for commercial devices special processes with a cleaner silicon-insulator interface are used. Bulk-storage requires a moderately doped diffusion layer underneath the gates, which is not available in standard MOSFET technology, but is provided as an additional implant in some processes.

As we will see in Section 10.5, CCDs using surface-storage photogates are called *surface-channel CCDs* and those using bulk-storage photogates are referred to as *buried-channel CCDs*.

10.4 Logarithmic Photosensors

In this section we will make use of the fact that photodiodes can be readily integrated with transistors in standard MOS technologies to show different ways to convert the photocurrent logarithmically into a voltage signal.

The logarithmic characteristic turns out to be quite for environments in natural ambient lighting. Such lighting conditions are characterized by a time-varying intensity, due the continuously shifting angle of incidence of the

Table 10.1
Illuminance and irradiance values under typical lighting conditions

Light source	Illuminance [lx]	Irradiance [W/m^2]
Full moon	1	0.01
Street lighting	10	1
Office lighting	100-1000	10-100
Direct sun	100000	1000

sunlight and changing optical transmission of the atmosphere (obscuration by clouds, etc.). Such temporal variations simultaneously affect large parts of the environment, but do not carry any information about it. In order to robustly characterize an environment in this situation, the characterization has to rely on parameters that are insensitive to global illumination changes.

Object surfaces are optically characterized by their spectral reflectivity distribution. Since the ambient light level is usually unknown, but can be assumed to be locally approximately constant (except at edges of shadows), the *relative reflectivity* of nearby areas in object space becomes an important parameter for the description of the objects. A measure for this relative reflectivity as mapped by an optical system into an image space is called *contrast*. There are different ways of defining contrast, but they all have in common that they are based on ratios of *irradiances* or *illuminances*[3] in the image and do thus not depend on the ambient scene illumination. A logarithmic mapping has the property that a ratio is mapped onto a difference. Consequently, in a circuit where the photocurrent is logarithmically converted into a voltage, contrasts are encoded as voltage differences, while the ambient light level sets the absolute voltage value of the operating point. Illuminances and irradiances for typical lighting conditions are shown in Table 10.1.

To simplify the equations in this section, we will assume that the width-to-length ratios W/L of the MOSFETs are taken into account by their respective current-scaling constants I_0 and we will use a first-order approximation of the subthreshold MOSFET characteristics, neglecting the Early effect, unless stated otherwise.

Basic Implementations

As we saw in Section 10.1, the current through an ideal photodiode is approximately linear with the incident power for a given spectral distribution of the

3 Irradiance denotes the incident radiant flux (power) per unit area, while illuminance denotes the incident luminous flux (obtained by weighting the radiant flux withthe human photopic sensitivity curve) per unit area.

incoming radiation. This approximation turns out to be quite good over several orders of magnitude in the photocurrent, corresponding to a substantial part of the illuminance range encountered in natural environments. It is thus possible to build electronic image sensors that operate over a large dynamic range without the need for any additional range-reducing devices, such as aperture stops, typically used in video cameras or light amplifiers. Depending on the particular application the photocurrent may have to be converted into a voltage. For certain applications, a linear conversion may be desirable, but given a power supply voltage of a few volts and typical noise levels of integrated circuits of the order of a few millivolts, the dynamic range in the voltage domain spans only about three orders of magnitude. In order to retain the large dynamic range of the photodiodes it is thus appropriate to use a *compressive* mapping, which is easily available via the voltage-current characteristics of transistors (Chamberlain and Lee, 1983). The most straightforward implementation of such a mapping is by connecting one or more transistors in series with the photodiode, such that the photodiode acts as a current source and the transistors as current sensors. Several possible MOSFET-based configurations of such a circuit are shown in Fig. 10.7. A diode-connected pFET is used as a current sensor in Fig. 10.7(a). A stack of two diode-connected pFETs, as shown in Fig. 10.7(b), increases the voltage signal's amplitude and moves its operating point further away from the power supply rail. The source-follower configuration of Fig. 10.7(c) sets the operating point with a bias voltage V_b, but further compresses the signal with respect to the other implementations due to the subthreshold slope factor of the MOSFET (see Chapter 3), This compression can be avoided if a unity-gain source follower (see Section 5.2) with the source connected to the bulk is used. Of course, all these configurations can also be implemented by simultaneously exchanging the type of MOSFET, the direction of the photodiode, and the power supply rails, or by using other types of transistors.

Typical irradiances under ambient lighting conditions (see Table 10.1) elicit photocurrents in the pA or nA range in photodiodes with areas of the order of $(10\,\mu\mathrm{m})^2$ as employed in imagers. At these current levels typical MOSFETs operate in their subthreshold domain, such that the current-to-voltage conversions performed by the circuits of Fig. 10.7 show a logarithmic characteristic. We then obtain

$$V_{out} = V_{dd} - \frac{U_T}{\kappa} \log\left(\frac{I_{ph}}{I_0}\right) \tag{10.4.1}$$

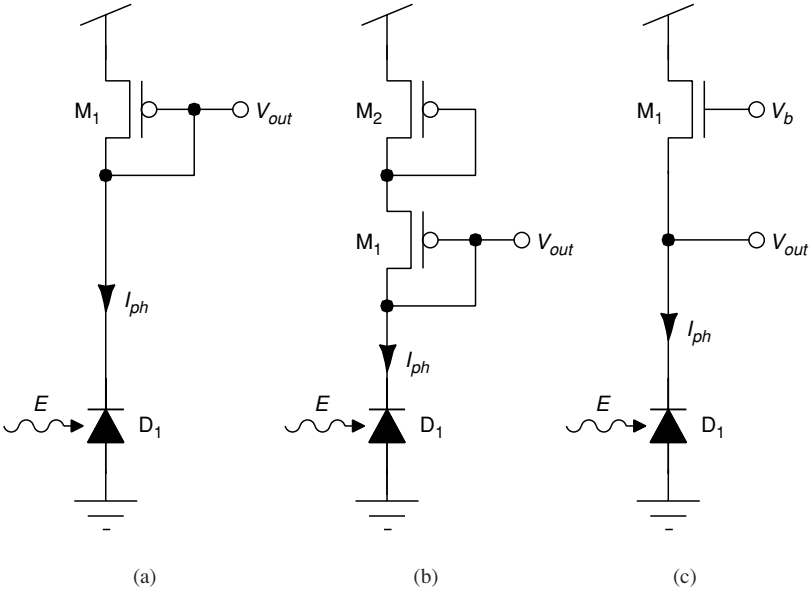


Figure 10.7
Photosensors with logarithmic irradiance-to-voltage conversion for subthreshold photocurrents, consisting of a photodiode and a current-to-voltage conversion stage implemented as (a) a diode-connected MOSFET, (b) two diode-connected MOSFETs in series, (c) a MOSFET in source-follower configuration.

for the circuit of Fig. 10.7(a), where U_T denotes the thermal voltage, κ the subthreshold slope factor of M_1 and I_0 the current-scaling constant of the current-sensing MOSFET. We define I_{ph} to be positive if it is a reverse photodiode current, as indicated in Fig. 10.7. We obtain

$$V_{out} = V_{dd} - U_T \frac{\kappa + 1}{\kappa^2} \log\left(\frac{I_{ph}}{I_0}\right) \tag{10.4.2}$$

for the circuit of Fig. 10.7(b) under the assumptions that both MOSFETs are identical and that the voltage dependence of κ can be neglected; and we obtain

$$V_{out} = \kappa V_g - U_T \log\left(\frac{I_{ph}}{I_0}\right) \tag{10.4.3}$$

for the circuit of Fig. 10.7(c). The contrast-encoding property can be readily

seen when the current-to-voltage conversion characteristics are written in differential form. For example, differentiating Eq. 10.4.3 gives

$$dV_{out} = -U_T \frac{dI_{ph}}{I_{ph}} . \tag{10.4.4}$$

The photodiodes in the circuits of Fig. 10.7 can also be replaced by phototransistors. However, the logarithmic characteristic is lost in this case because of the nonlinear current-irradiance dependence of the phototransistor. In addition, if MOSFETs are used as current sensors they may be forced into their above threshold operating domain by the amplified photocurrents provided by the phototransistor, which further distorts the logarithmic characteristic.

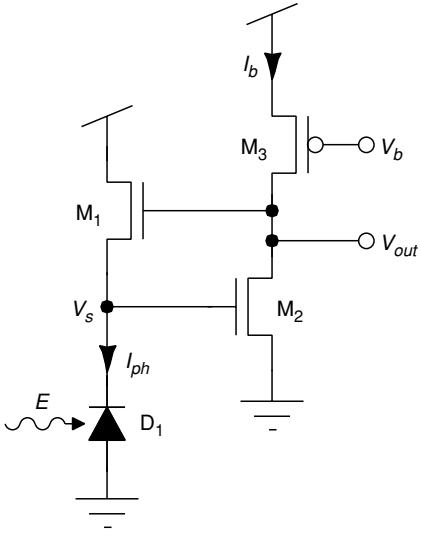


Figure 10.8
Logarithmic photosensor with feedback loop increasing the bandwidth by clamping the voltage V_s across the photodiode.

Feedback Implementation

A limitation of the photosensors depicted in Fig. 10.7 is their small bandwidth for time-varying signals. This is due to the photodiode junction capacitance and the parasitic capacitances of the output node onto the different MOSFET terminals. These capacitances have to be charged or discharged by the typi-

cally very small photocurrent. If a phototransistor is employed, the amplified collector current can easily handle the MOSFET capacitances, but the base-emitter junction capacitance still has to be driven by the non-amplified photocurrent. Since the base voltage of a phototransistor has to be modulated to cause transistor action the problem is inherent. In the case of the photodiode, however, the response speed can be enhanced by adding a high-gain negative feedback loop from the source to the gate of the current-sensing MOSFET in the source-follower configuration of Fig. 10.7(c). The voltage output signal then appears at the gate of the MOSFET, while the source and therefore the voltage across the photodiode is practically clamped. Such an arrangement is depicted in Fig. 10.8, where the source node drives a two-transistor amplifier with a bias voltage V_b and a voltage gain A whose output is connected to the gate node. The output voltage is then given by

$$V_{out} = \kappa^{-1} \left(V_s + U_T \log \left(\frac{I_{ph}}{I_0} \right) \right) \tag{10.4.5}$$

with the source V_s being held nearly clamped to the constant value

$$V_s = \kappa_n^{-1} \left(\kappa_p (V_{dd} - V_b) + U_T \log \left(\frac{I_{0p}}{I_{0n}} \right) \right) \tag{10.4.6}$$

where κ_n and κ_p are the subthreshold slope factors of M_2 and M_3 respectively and I_{0n} and I_{0p} are the corresponding current-scaling constants.

A differential photocurrent change dI_{ph} results in a differential output voltage change

$$dV_{out} = U_T \frac{A}{\kappa A - 1} \frac{dI_{ph}}{I_{ph}} \approx \frac{U_T}{\kappa} \frac{dI_{ph}}{I_{ph}} \,. \tag{10.4.7}$$

From Eqs. 10.4.4 and 10.4.7 we see that the feedback circuit amplifies the response of the simple source-follower configuration by a factor $1/\kappa$ and inverts it, such that an increase in I_{ph} results in an increase in V_{out}. The measured steady-state irradiance-voltage characteristic of such a photosensor is shown in Fig. 10.9. The differential source voltage change is

$$dV_s = \frac{U_T}{\kappa A - 1} \frac{dI_{ph}}{I_{ph}} \approx \frac{U_T}{\kappa A} \frac{dI_{ph}}{I_{ph}} \,. \tag{10.4.8}$$

Hence, the feedback circuit reduces the voltage variations on the source node by a factor of κA. If the bias current of the amplifier is chosen such that it is much larger than the photocurrent then the time constant of the circuit is determined by the dynamics of the source node of the current-sensing

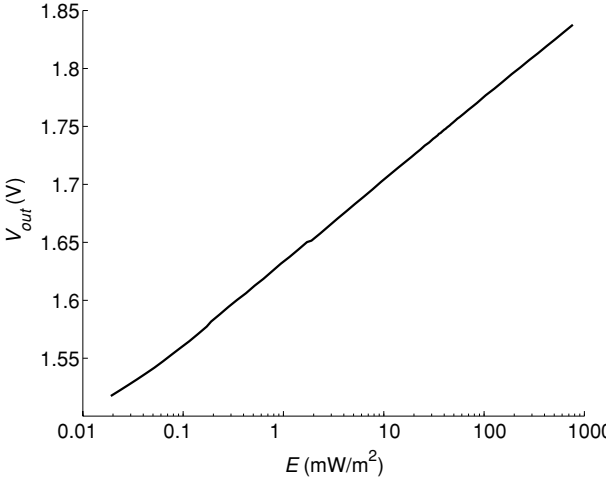


Figure 10.9
Steady-state voltage response V_{out} of an implementation of the logarithmic photosensor of Fig. 10.8 for different irradiance levels. The response function is to a good approximation logarithmic over the measured irradiance range spanning 4.5 orders of magnitude.

MOSFET as in the case of the simple source-follower version. If the additional parasitic capacitance introduced by the feedback circuit can be neglected, the feedback circuit reduces the response time by a factor of κA. Note that this increase in bandwidth decreases the time over which the signal is integrated and correspondingly increases the noise at the output.

Adaptive Photosensor

In the following, we will illustrate how *adaptation*, an ubiquitous phenomenon in biological nervous systems, can enhance the performance of a photosensor. This example shows some benefits of adaptation that can be applied to other data processing systems, as well.

A major concern in the design of photosensor arrays for imaging applications is mismatches between the characteristics of identically-designed photosensing elements. These mismatches are due to inaccuracies of the semiconductor processing. The effect of these mismatches on the output signal is known as *fixed-pattern noise (FPN)*. In the logarithmic photosensors described in this section the current-sensing MOSFETs that are operated in their subthreshold domain are the main source of FPN. An e-fold increase in the photocurrent corresponding to the effect of a medium optical contrast elicits

a voltage change of the order of U_T, which can be of the same order of magnitude as the gate-to-source voltage mismatch between the MOSFETs for a given current. The photocurrents are proportional to the areas of the corresponding photodiode junctions, whose spatial variations are the main source of photocurrent mismatches and are typically limited to a few percent.

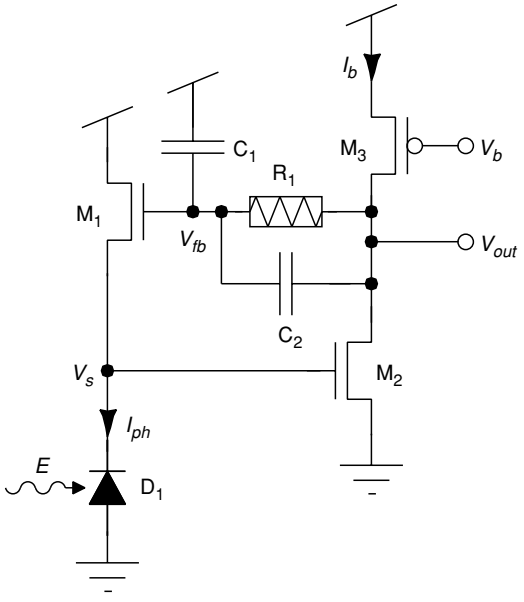


Figure 10.10
Adaptive logarithmic photosensor with amplified transient response. The amplification stage consists of a capacitive divider and a resistive element, which typically shows a non-linear current-voltage characteristic.

The FPN can be reduced by individual tuning of each photosensor. However, in addition to a calibration procedure, such a solution requires either individual biasing of each sensor or local memory at each element to store the calibration parameter values. A more economic solution is the enhancement of transient signals with respect to the mismatched steady-state signals using a well-matched amplifier stage. This strategy increases the signal-to-noise ratio and makes better use of the voltage range provided by the power supply. Such a well-matched transient gain stage can be fabricated with a capacitive divider. Adaptation to the DC value can be provided by a resistive element. The gain

stage can be integrated into the feedback loop of the photosensor of Fig. 10.8, as shown in Fig. 10.10 (Delbrück, 1993; Delbrück and Mead, 1994). The transient output voltage change dV_{out} is amplified with respect to the DC output voltage change dV_{fb} (Eq. 10.4.7) by the capacitive divider ratio

$$A_C \equiv \frac{C_1 + C_2}{C_2} \tag{10.4.9}$$

as long as the adaptation effect can be neglected. The transient output voltage change is thus

$$dV_{out} = A_C U_T \frac{A}{\kappa A - 1} \frac{dI_{ph}}{I_{ph}} \approx A_C \frac{U_T}{\kappa} \frac{dI_{ph}}{I_{ph}} . \tag{10.4.10}$$

The photosensor adapts to variations in the photocurrent on a long time scale, which usually reflect slow changes in the background illumination that are typical of natural lighting conditions. The adaptation state is represented by the charge Q_{fb} stored on the capacitor plates of the feedback node. The output voltage V_{out} depends on this adaptation state and on the input signal represented by V_{fb}. It can be expressed as

$$V_{out} = A_C V_{fb} - \frac{Q_{fb} + C_1 V_{dd}}{C_2} \tag{10.4.11}$$

where V_{fb} is given by

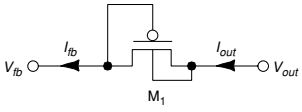
$$V_{fb} = \kappa^{-1} \left(V_s + U_T \log \left(\frac{I_{ph}}{I_0} \right) \right) . \tag{10.4.12}$$

The adaptation dynamics are determined by the characteristics of the resistive element. A large resistance results in slow adaptation: a small resistance in fast adaptation. A linear resistor with a resistance R makes V_{out} adapt exponentially after an irradiance step. This can be seen from the differential equation that is obtained when V_{fb} remains constant following an initial step change ΔV_{fb}: Differentiating Eq. 10.4.11 and using Ohm's law yields

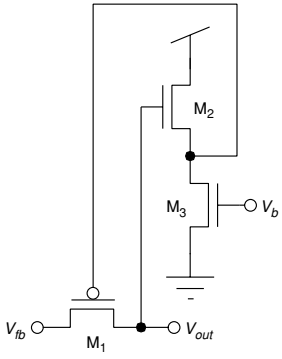
$$\frac{dV_{out}}{dt} = -C_2^{-1} \frac{dQ_{fb}}{dt} = -(RC_2)^{-1} (V_{out} - V_{fb}) . \tag{10.4.13}$$

If the sensor was fully adapted before the irradiance step at time $t = 0$ we obtain the initial condition that immediately after the step

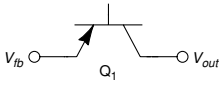
$$V_{out}(0^+) = V_{fb} + A_C \Delta V_{fb} . \tag{10.4.14}$$



(a)



(b)



(c)

Figure 10.11
Resistive elements suitable for implementation in the adaptive photosensor of Fig. 10.10. (a) Expansive element with low common-mode sensitivity. (b) Expansive element with more symmetric characteristics. (c) Compressive element.

Using this condition, Eq. 10.4.13 is easily solved:

$$V_{out}(t) = V_{fb} + A_C \Delta V_{fb} e^{-t/RC_2} \,. \tag{10.4.15}$$

With the specific capacitance and sheet resistance values provided by typical semiconductor technology the decay time constants achieved with linear resistors would be much too small for practical applications. The resistive elements are therefore more conveniently built from transistors, which can be operated at low currents that are matched to typical capacitance values. The

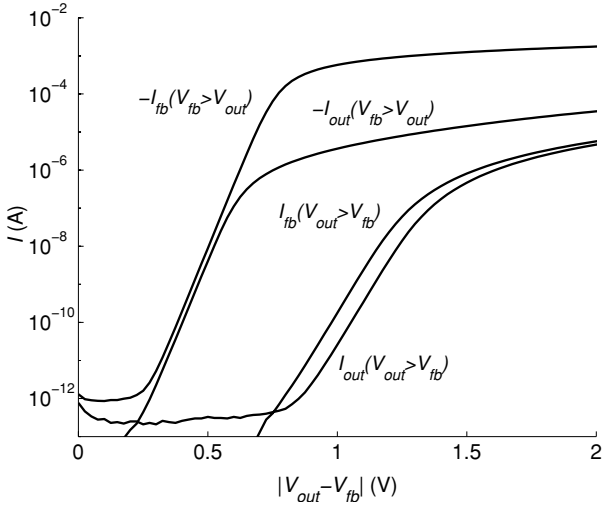


Figure 10.12
Current-voltage characteristics of the resistive element of Fig. 10.11(a). The currents at both terminals are shown for both directions of current flow. For $V_{out} > V_{fb}$ the element operates as a diode-connected MOSFET. For $V_{fb} > V_{out}$ the MOSFET is shut down, but current flows through a lateral BJT between the two terminals and through a vertical BJT from the feedback node to the substrate.

current-voltage characteristics of these elements are *non-linear*, with the resistance in most cases decreasing or increasing monotonically with the absolute value of the voltage difference across the element. If it decreases we speak of an *expansive resistive element*; if it increases we call it a *compressive resistive element*. In the case of an expansive element the adaptation is faster than exponential, and for a compressive adaptive element it is slower than exponential. Expansive elements provide rapid adaptation to new lighting conditions, but do not adapt out low-contrast signals, which usually correspond to significant image details rather than the effects of changes in the scene illumination. Compressive elements may take a long time to adapt to large changes in illumination and can therefore support a larger voltage swing of the amplified signal, which may be due to either a larger amplifier gain or to a larger input signal. The price to be paid for this enhanced response is slow adaptation to new lighting conditions. During this adaptation, the response of the output node may be saturated until a new equilibrium is reached.

Three different implementations of non-linear resistive elements are shown in Fig. 10.11. The element of Fig. 10.11(a) (Delbrück, 1993; Delbrück and

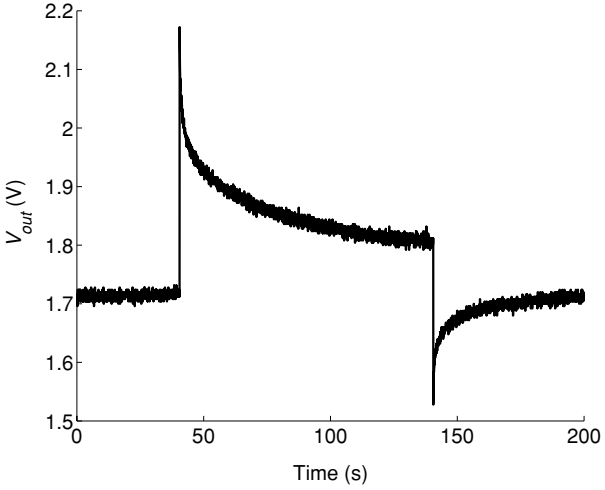


Figure 10.13
Voltage response V_{out} to an irradiance pulse of an implementation of the adaptive photosensor of Fig. 10.10 with the resistive element of Fig. 10.11(a). The response shows large transients followed by gradual adaptation to the DC values.

Mead, 1994) is expansive. For current flow from the output node to the feedback node it has the exponential characteristics of a diode-connected pFET. In the reverse direction the pFET is turned off, but the diffusion at the feedback node acts as the emitter and the well as the base of a BJT with two collectors, one being the diffusion at the output node, the other being the semiconductor substrate. The current-voltage characteristic, shown in Fig. 10.12, is thus also exponential in this direction, but without the subthreshold slope factor κ in the exponent and with a much larger linear current-scaling parameter. The response of an adaptive photoreceptor with such a resistive element to an irradiance pulse is shown in Fig. 10.13. The asymmetry between upward and downward adaptation can clearly be seen. A more symmetric expansive arrangement, where the resistance can be modulated by a bias voltage V_b, is shown in Fig. 10.11(b) (Liu, 1999). The behavior in both directions is governed by the pFET characteristics with the remaining asymmetry that the gate voltage is set by V_{out}, via a source follower. A further difference between the two elements is the fact that the element of Fig. 10.11(a) is a good approximation of a two-terminal device and its resistance thus only depends on $V_{out} - V_{fb}$, the only common-mode effect being the well-to-substrate voltage,

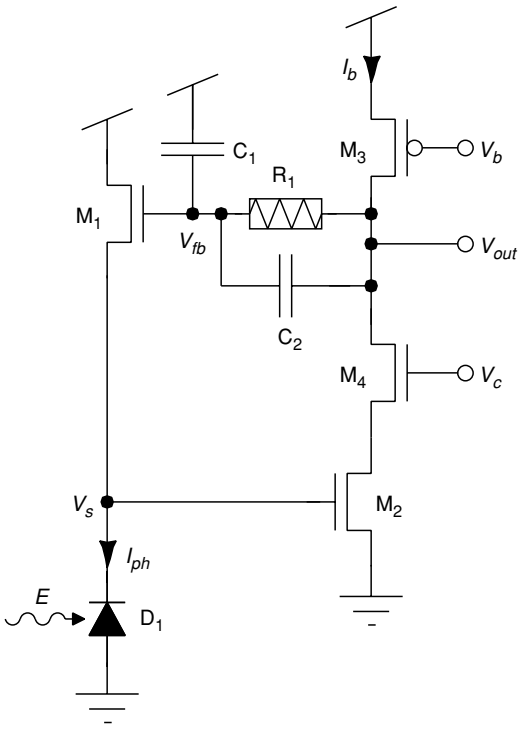


Figure 10.14
Adaptive logarithmic photosensor with cascode transistor increasing the bandwidth for small photocurrents.

which is the emitter-to-collector voltage of one of the BJTs. The other element, however, is subject to the body effects of the source-follower nFET and the resistive pFET and therefore changes its characteristics with irradiance: an effect that may be beneficial if faster adaptation at higher light levels is desired. Furthermore, it is more sensitive to leakage currents to the power supply rails induced by photon-generated minority carriers. The resistive element of Fig. 10.11(c) (Delbrück, 1993) is a bipolar transistor with a floating base, which can be regarded as two diodes with reversed polarity connected in series. In either direction, the current is limited by the dark current of the reverse-biased diode. The current-voltage characteristic has a compressive, sigmoidal shape, which for ideal diodes saturates at the corresponding reverse diffusion currents J_s, as described by the Shockley equation (Eq. 2.6.22). However, as

we saw in Chapter 2, the reverse currents of real diodes are substantially larger than predicted by the Shockley equation and do not saturate completely.

The transient voltage variations at the output node of the adaptive photosensor can be quite large, depending on the chosen capacitive divider ratio and resistive element. The expansive resistive elements of Fig. 10.11(a) and (b) support variations of the order of 1 V, while with the compressive resistive element of Fig. 10.11(c) and a large enough capacitive divider ratio almost the entire range between the power supply voltages can be used. The parasitic capacitance C_p from the source of M_1 onto the output node via the gate-to-drain capacitance of nFET M_2 gives rise to the *Miller effect* (Gregorian and Temes, 1986). Hereby, the apparent capacitance, as seen from the source of M_1, is increased by C_p multiplied by the voltage gain A from gate to source. For small photocurrents this effect leads to a slow response. The introduction of a *cascode* nFET M_4 with a fixed gate voltage V_c (Fig. 10.14) clamps the voltage on the drain of M_2, because the current through the amplifier is approximately constant, and largely nullifies the Miller effect. However, for certain biasing conditions the presence of the cascode can make the circuit unstable.

10.5 Imaging Arrays

There is a large market for electronic imaging devices, which consequently have reached a high degree of sophistication. In the following, we will present a brief overview of the basic principles of the most commonly used types of imagers, most of which are typically designed to yield a linear measure of the irradiance distribution of the image projected onto them. An individual photosensing element, called *pixel* (for **pict**ure **el**ement), is typically small, so that a high image resolution is achieved. Because of its small size, a pixel can usually only contain transistors of one type.

Large photosensor arrays integrated on a single semiconductor substrate, as used in electronic imaging applications, cannot usually be read out in parallel, due to the large number of pixels and the limited number of output terminals. The number of pixels scales with the square of the linear imager size, while the number of output terminals scales only linearly. The outputs of the individual pixels have to be multiplexed onto few signal lines: in most cases a single one, and in some cases one per row. The readout duty cycle for each photosensor is typically quite small. Given this data sampling requirement, there are two basic strategies for pixel operation: In the *continuous-time mode* the

photogenerated charge is converted into a steady state current, which is sensed whenever the pixel is addressed. In the *integration mode* the photogenerated charge is collected on a capacitor until it is read out, whereafter the capacitor charge is reset to a given value and a new integration cycle is started. In continuous-time operation the photogenerated charge is not sensed when the pixel is not addressed. By contrast, in the integration mode all the collected charge contributes to the output signal. This mode offers better sensitivity and signal-to-noise ratio provided that the integration time is larger than the response time of the sensor.

Integration of low-level parallel image-processing circuitry with photosensor arrays (*focal-plane processing*) is a method of efficiently implementing a host of functions that are also found in biological retinal structures, such as adaptation to background illumination or color, edge enhancement and motion extraction. Such functions often rely on local spatial coupling among the pixels, where the instantaneous value of one pixel influences the value of its neighbors to implement a spatial filtering function. A simple integrate-and-reset scheme for the photosensors distorts the computation of such functions [4], unless all pixels are reset simultaneously or the readout period is made much smaller than the integration period. Modern imaging devices often have an electronic shutter function, in which the image is synchronously transferred to an array of shielded storage nodes at the end of the integration cycle. However, in the interest of simplicity we will only present the basic version of each type of pixel.

Passive Pixel Sensors

A passive pixel typically consists of a photodiode and a binary transfer gate (Weckler, 1967), as shown in Fig. 10.15(a). The transfer gate M_1 leads to a shared readout line, which is held at a fixed potential V_{ref}. When the transfer gate M_1 is open, the photodiode is reverse-biased to V_{ref}. Closing the transfer gate open-circuits the photodiode and the photogenerated charge accumulates across the capacitance of the photodiode's depletion region, decreasing the voltage V_{ph} across the diode. After a certain integration time the transfer gate is opened again and the charge is sensed by an amplifier connected to the readout line, while the reverse bias is restored to V_{ref}. The readout is *destructive*, which means that the signal charge is lost during the readout.

4 It also induces a temporal distortion to the raw image, which becomes apparent for long integration periods.

If the integration time is fixed, the sensed charge is proportional to the irradiance, provided that the collected charge does not saturate the photodiode. Saturation sets in when V_{ph} approaches zero and charge carriers start to recombine due to forward diffusion. The maximum charge that can be collected without significant saturation effects is given by

$$Q_{max} = CV_{ref} \tag{10.5.1}$$

where C is the depletion capacitance across the diode at zero external bias.

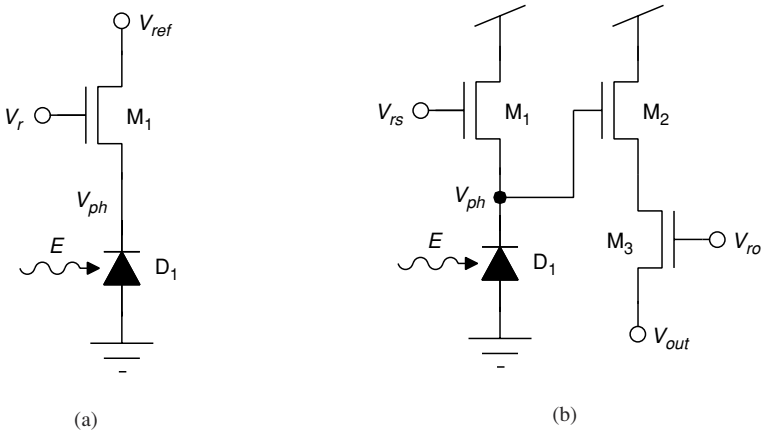


Figure 10.15
Integrating photodiode pixels for imaging applications. (a) Passive pixel for charge readout and reset by a shared line biased at V_{ref} and connected to the pixel via a transfer gate M_1. (b) Active pixel with a reset gate M_1 for voltage readout via a source-follower transistor M_2 and a transfer gate M_3.

In a two-dimensional array of the passive pixels described above, each column has its own readout line and an entire row is read onto the column lines simultaneously. The column lines are then multiplexed onto a common line with an additional transfer gate per column (Fig. 10.16). If random access to each pixel is required the pixel has to include a second transfer gate addressed by additional column lines to decouple the readout along the rows.

If spatially coupled pixels are not read out simultaneously, which is usually the case if they share a readout line, their response is distorted with respect to

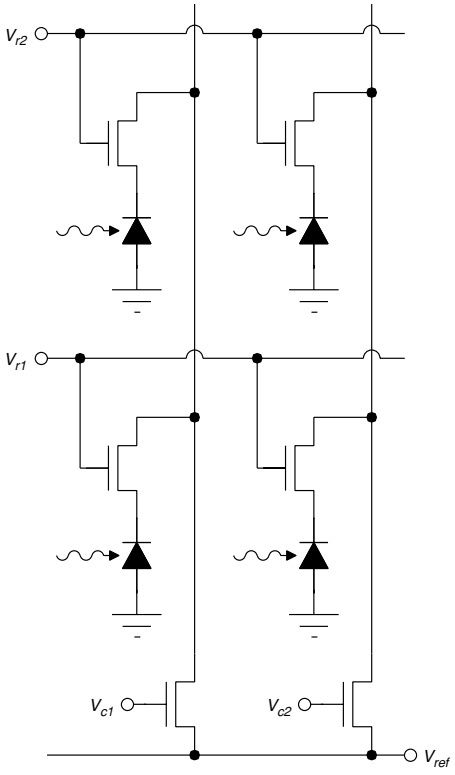


Figure 10.16
Architecture of a two-dimensional passive pixel array. The signals of each row are transferred in parallel to column readout lines via the transfer gates controlled by voltage signals V_{r1} and V_{r2}, respectively. The signals from the different columns are read out serially via the transfer gates controlled by voltage signals V_{c1} and V_{c2}, respectively. The readout lines are biased at a fixed voltage V_{ref}.

the response for true parallel coupling. Sequential addressing of adjacent pixels, which is normally used for reading entire images, minimizes the distortion.

Passive pixels have quite a large *fill factor* (the ratio of the photosensitive area and the total area) because they use only one or two small transistors and two or three global wires per pixel. Furthermore, their FPN is rather small, because the photodiodes can be reasonably well matched and the transistors are only used as switches. Disadvantages of passive pixels are the large parasitic

capacitance of the readout lines resulting in a relatively large readout noise [5] and the low sensitivity and slow response of the charge amplifier. The parasitic capacitance is proportional to the length of the readout line and thus increases with array size.

Active Pixel Sensor

Active Pixel Sensor (APS) (Fossum, 1993, 1997) is the name given to a class of image sensors, where each pixel contains at least a buffer or an amplifier (Noble, 1968). An implementation of an integrating APS pixel with voltage readout is depicted in Fig 10.15(b). The voltage signal V_{ph} generated by the collected charge on the diode capacitance is buffered by a source follower consisting of transistor M_2 and a current source (not shown) common to all pixels connected to the same readout line. A pixel is read while pass transistor M_3 is open and reset when pass transistor M_1 is opened. In this configuration, the signal voltage is roughly linear with irradiance for fixed integration intervals. Non-linearities are introduced by the voltage-dependence of the photodiode capacitance and by the body effect of M_2.

Standard readout strategies for active pixel arrays are similar to those for passive pixel arrays. However, readout and reset signals are now separated. This separation complicates the addressing of the binary gates but it allows for additional modes of operation. In particular, data readout is not destructive. That is, the pixel is not automatically reset upon readout.

The photodiode APS pixel has a smaller fill factor than the passive pixel, but much better signal-to-noise ratio and larger bandwidth due to the buffering. The mismatch of the transistors M_2 that are operated in their analog domain causes a significantly increased FPN with respect to the passive pixel sensor. Hence, active pixel sensors are only practical if they include offset correction circuits, such that the signal of each pixel is measured with respect to a value that represents the pixel's response to a given reference input signal (Nixon et al., 1995).

Integrating active or passive pixels can also be implemented with a photogate instead of a photodiode (Mendis et al., 1997). In this case the photogenerated charge is collected underneath the photogate. At the end of the integration period the charge transferred to a reverse-biased diode by changing the voltage on the photogate. From the diode, it is read out in the same way as in a photodiode pixel. Photogate pixels may have lower noise (due to smaller readout

5 The charge noise on a capacitance C is proportional to $\sqrt{kTC}$.

capacitances) but they also have lower quantum efficiencies than photodiode pixels.

The logarithmic photosensors (Section 10.4) can also be used in active pixels. The simple versions of Fig 10.7 should be buffered, for example with the source-follower circuit of Fig 10.15(b), while the feedback versions may not need a buffer, because they do not use the photocurrent to drive the output node. Instead they use an independent current source. The feedback versions include both types of transistors, at least one of which requires a well (see Chapter 13), and therefore cannot be implemented compactly.

Charge-Coupled Device

A charge-coupled device (CCD) (Boyle and Smith, 1970; Theuwissen, 1995) consists of an array of photogates that integrate the photogenerated charge. At the end of the image acquisition period, which is the common charge-integration period of all pixels in the array, the signal charge packets of the different pixels are simultaneously moved underneath the semiconductor surface towards the readout circuitry. Each charge packet stays together in a locally confined volume until it is read out and does not spread out on a long signal line. This method of readout achieves high bandwidth, high sensitivity, and low noise. Most CCDs are *buried-channel CCDs* (Walden et al., 1972) that consist of bulk-storage photogates and thus integrate and transport the charge in the bulk of the semiconductor. They are usually preferred to *surface-channel CCDs*, where storage and transport occurs at the semiconductor-insulator interface, because the interface has a much larger trapping density than the bulk, due to lattice mismatch and surface effects during fabrication. The main disadvantage of buried-channel CCDs is their small charge-storing capacity. If a charge packet gets overfilled the excess charge may spill into adjacent packets, an effect known as *blooming*. Buried-channel CCDs cannot be fabricated with standard CMOS technology, because no moderately-doped buried channel diffusion is available. Surface-channel CCDs are in principle CMOS-compatible, but in practice, the silicon-oxide interface of a standard CMOS process does not have a good enough quality to guarantee an efficient charge transport across hundreds of stages, required for CCD imagers. Charge-coupled devices for imaging applications are therefore fabricated with specialized processes, which are more expensive than standard CMOS processes.

The charge packets are either transported underneath the photogates themselves, as in the case of the *frame-transfer CCD* (Fig. 10.17(a)), or first transferred to parallel running lines of MIS gates via a transfer gate and transported

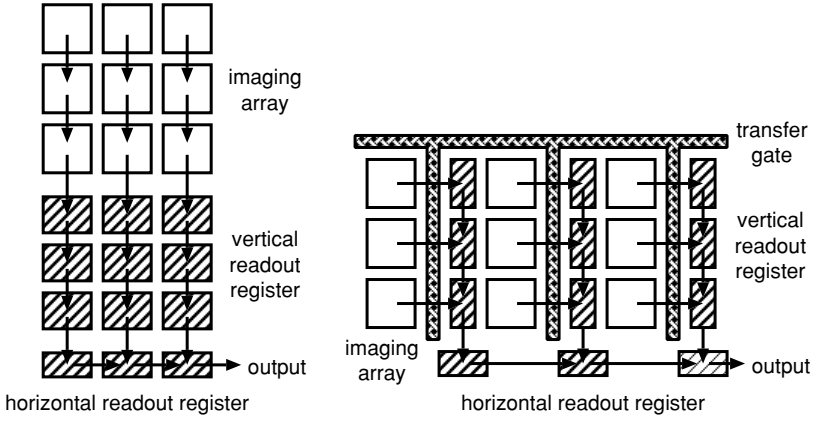


Figure 10.17
Architecture of (a) frame-transfer CCD and (b) interline-transfer CCD. The hatched areas indicate MIS gates that are shielded from light. In the frame-transfer CCD, the photogenerated charge distribution is rapidly shifted into a shielded readout register adjacent to the imaging array at the end of the image acquisition period. The readout register is then scanned out at the speed required by the video standard. In the interline-transfer CCD, the shielded readout register is interlaced with the imaging array. The image is latched from the imaging array into the readout register by a single transfer gate. Interline-transfer CCDs have a smaller fill factor of the imaging area than frame-transfer CCDs, but require less complicated timing and suffer less from image smearing effects.

underneath those (Fig. 10.17(b)) in which case the device is called an *interline-transfer CCD*. In either scheme, the transport is carried out by shifting the local energy minima for the collected charge type in the semiconductor from underneath one gate to an adjacent one in a given direction, such that the charge packets stay separated, each traveling with a different local minimum. The shifting is accomplished by appropriately clocking the voltages of the different gates between two or more voltages. Two different clocking schemes are illustrated in Fig. 10.18 (Amelio et al., 1971). After the integration period, the frame-transfer technique shifts the charge packets rapidly along the columns from the imaging array into a readout register that is shielded from light. The interline-transfer technique uses a single transfer gate to shift the charge into a readout register, whose columns are interlaced with those of the imaging array. The frame-transfer CCD offers a better fill factor than the interline-transfer

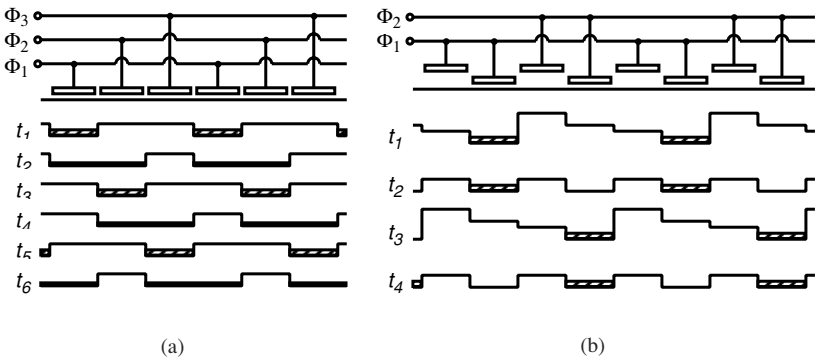


Figure 10.18
Charge-packet transport strategies in CCDs with (a) three-phase clock and (b) two-phase clock with stepped oxide. Cross-sections of the CCDs with indicated connections to the clock phases (Φ_1, Φ_2, Φ_3) are shown together with snapshots at different times (t_1, t_2, t_3, t_4, t_5, t_6) of the potential distributions underneath the photogates. The charge packets, symbolized by the hatched areas, are shifted from left to right as time progresses. By using a stepped oxide with different thicknesses for adjacent photogates connected to the same clock phase, the directionality of charge transfer can be built into the device and a two-phase clock is sufficient. Two-phase operation can even be accomplished with a single clock by keeping one clock phase at a fixed reference potential, and clocking the other one to potentials lower and higher than the reference.

CCD. However, the interline-transfer CCD has simpler clocking and does not suffer as much from light-induced smearing effects.

In a two-dimensional CCD image sensor the charge is transported in parallel, synchronously clocked column CCD shift registers. In some applications requiring a large readout bandwidth, each column has its own charge-sensing amplifier, but more typically a horizontal CCD along the edge of the array transports the charge towards a common charge-sensing amplifier. The single-amplifier scheme limits the FPN to the matching of the photogates and smearing effects due to incomplete charge transfer.

Charge-coupled devices are suitable for the implementation of certain local image processing operations (Fossum, 1989), which can be programmed with appropriate clocking schemes. Since signal processing occurs in the charge domain, operations such as addition, subtraction and averaging are easily implemented. For example, acquisition of image pyramids, that is, the same image at different resolutions, in successive frames can readily be achieved by switching adjacent pixels together to decrease the resolution by a step (Seitz et al., 1993).

The CCD has dominated the solid-state image sensor market since the 1970's (Boyle and Smith, 1970), because of its low FPN and readout noise and its high sensitivity. However, the CCD's large power consumption and the inconvenience of integrating peripheral circuitry on the same substrate with the imaging array, due to the large capacitive load of the gates and the high processing costs, make it more expensive and less suitable for miniaturization than the APS. It is generally predicted that the consumer market will eventually be taken over by the APS (Chute, 2000).

10.6 Limitations Imposed by Dark Current on Photosensing

Junction leakage current in the dark (*dark current*) is the main limitation to photosensitivity at low light levels. In typical CMOS processes, which are not optimized for low junction leakage, large area junctions leak about 1 nA/cm^2. This number seems to be fairly constant over processes with which we have experience, varying over perhaps a factor of 5. Some fabrication houses have specialized processes developed for DRAM or imagers that have much lower claimed leakage currents, down to perhaps 25-50 pA/cm^2. However these processes are presently not available to multiproject wafer customers. Dark current is generally not a very strong function of junction reverse bias voltage; We have seen at most a doubling of current as the reverse voltage is swept from 0 to the power supply.

In addition, dark currents from different pixels can vary by factors of 10 or 100: It is common to see a subpopulation of pixels that form outliers in a histogram. These hot spots are thought to arise from as little as a single generation/recombination state at a level that happens to be right in the middle of the band gap. In an imager, these *hot pixels* cause isolated white spots that are very noticeable in an image and must be corrected for by interpolation of surrounding pixel values[6].

The dark current in a junction is dominated by leakage around the edges of the junction, where the junction meets the interface between Si and SiO_2. It is thought the increased leakage here in the *sidewall* region arises from the additional stress induced in the Si crystal and from interface states at the interface itself (Hawkins, 1985). Measurements have shown wide variation in the relative size of the sidewall leakage relative to the area leakage, but a factor

6 Around 1990, some CCD video camcorders were already equipped with nonvolatile storage of the location of these hot pixels and interpolation from surrounding pixels to correct for them.

anywhere between 5 and 100 is possible. A factor of 10 would mean that 1 μm along the edge leaks as much as 10 μm^2 of area. For a small photodiode, the sidewall leakage completely dominates the total, so claims of low junction leakage sometimes given by vendors can be quite misleading, since because claims are based on measurements from large area junctions. The replacement of LOCOS isolation by STI[7], if the STI is done correctly, is now generally believed to lead to lower sidewall leakage.

Since dark current is a thermal process, it is a strong function of temperature, doubling about every 8 °C (Theuwissen, 1995)[8]. For an imager or photosensor that must operate at ambient temperatures of 60 °C, the dark current will be about 20 times larger than at room temperature. CCD manufacturers usually quote their dark current figures at a temperature of 50–60 °C to keep their customers from being dissatisfied when the camera is operated after lying in a hot car all afternoon.

This junction leakage acts like a "glow in the dark" that degrades image contrast and contaminates outputs from a storage pixel array during readout. We can estimate the equivalent illuminance of a scene that corresponds to a given dark current. First we will compute the chip illumination, then we will calculate the scene illumination. Take a photodiode with area 10 μm^2. If the junction leakage is 1 nA/cm^2, corresponding to about 50 electrons/μm^2/s, and the sidewall leakage is 10 times larger, 500 electrons/μm/s, then this photodiode will leak about 8000 electrons/s. Only 500 of these electrons come from the area leakage, the other 7500 come from the sidewall leakage. The illumination of the chip corresponding to this leakage is computed as follows. We assume the photodiode has a quantum efficiency (QE) of 50%. That QE means that the flux is doubling the resulting current density. So, the photon flux is 8000*2 photons in an area of 100 μm^2 (160 photons/μm^2/s). "White" light of illuminance 1 lux is about 10^4 photons/μm^2/s (Rose, 1973). This conversion is approximate, because it depends the definition of "white", but it is good enough for our estimate. Using this conversion, we compute that the equivalent illuminance of the chip is about 20 mlux. Now we must estimate the illuminance of a scene that would result in this illumination of the chip. We can do this by using a very useful formula that can by derived from spherical geometry and some understanding of the units of photometry. A perfectly transparent lens with a given f number imaging a perfectly white diffusively

7 See Section 13.1.

8 Generation processes depend on the dominant (most likely) activation energy, usually between E_g and $E_g/2$. Lower leakage generally corresponds to higher activation energy.

reflecting surface produces an image illuminance I_{image} related to the scene illumination I_{scene} by

$$I_{\text{image}} = \frac{I_{\text{scene}}}{4f^2} \tag{10.6.1}$$

If we are using a fairly fast $f/2$ lens, the image illumination is only 1/16 that of the scene. We can now compute that the dark-current-equivalent scene illumination is about 0.25 lux. Under full-moon conditions, the illumination is about 0.1 lux (Rose, 1973). We can see that the dark current will limit our photosensor performance! Even with full moon, white objects out in the world will only generate about one third the photodiode's leakage current. It is clear that to make commercially viable devices that can operate over a truly wide range requires access to low-leakage processes.

Since low-leakage is a strong Darwinian survival trait for DRAM and commercial imagers, there has been a huge effort to reduce the junction leakage. Fastidious wafer cleaning before fabrication, gettering of contaminants, and other proprietary tricks have been developed to lower the junction leakage. Plummer et al. (2000) have written a very clear description of some of these tricks. Research users have little access to these specialized process flows.

One trick that is particularly interesting was discovered almost accidently by Teranishi et al. (1984) in pursuit of an interline transfer CCD with lower image lag. The trick is to make a *buried photodiode*: A photodiode which is covered with a thin implant of the substrate doping type. This covering implant, which is shorted to the photodiode substrate, acts to nearly eliminate the sidewall leakage in the photodiode and also reduces the leakage due to interface states at the Si/SiO_2 interface. Unfortunately for ordinary users, this trick requires a thin surface implant which can overlap the edge of the buried junction and connect to the bulk, and it does no particular good unless it is also combined with all the other contamination and defect reduction techniques. This technology is called a *Hole accumulation diode*, HAD for short, because the buried implant is n-type and it is covered with a p-type accumulated layer that fills the interface states, so that they cannot contribute to dark current. This technology is also called *pinned photodiode*, because the surface potential is pinned to the bulk potential. HAD has its drawbacks: It requires additional processing steps, and the blue response is poor and badly controlled, because the blue photons make minority carriers near the surface where they are very likely to recombine with the abundant holes. Since blue photons are always in short supply, this is a big problem for good color imaging.

For imagers designed for still pictures, dark current can be measured in one frame with the mechanical shutter closed, and then later subtracted from the pixel output. This still leaves the shot noise from the dark current that cannot be calibrated away, but substantial improvements in image quality can still be realized by removal of transistor mismatch effects. The problem remaining is the noise in the dark current: If the dark leakage averages N electrons per frame, the variation will be $\sqrt{N}$. This noise due to dark current cannot be calibrated away. In any case, subtracting uncorrelated noise sources, whatever the source of the noise (thermal or photon shot) only increases the noise.

Finally, for APS imagers with storage pixels and electronic shutter, dark current can considerably degrade the image during the storage and readout phase, particularly for a large array. Here it is the leakage on the drains of the transistor storage switch that matters, and not the photodiode leakage. A HAD photodiode does not help the readout problem.

IV SPECIAL TOPICS

This page intentionally left blank

11 Noise in MOS Transistors and Resistors

Many people find the subject of noise mysterious. Although the fundamental physical concepts behind noise are simple, much of this simplicity is often obscured by the mathematics invoked to compute expressions for the noise. In this chapter, we cover the basics of noise in electronic devices and circuits and we also discuss theoretical and experimental results for white noise in the low-power subthreshold region of operation of a MOS transistor (see Chapter 3) (Godfrey, 1992; Sarpeshkar et al., 1993; Sarpeshkar, 1997)

We solve the mystery of how a shot-noise answer derives from a thermal-noise viewpoint by taking a fresh look at noise in subthreshold MOS transistors. We then rederive the expression for thermal noise in a resistor from our viewpoint. We believe that our derivation is simpler and more transparent than the one originally offered in 1928 by Nyquist who counted modes on a transmission line to evaluate the noise in a resistor (Nyquist, 1928). The derivation here leads to a unifying view of the processes of shot noise and thermal noise in electronic devices. Specifically we show that white noise, whether it is labeled as shot or thermal, is completely accounted for as shot noise due to internal diffusion currents in electronic devices. These internal diffusion currents are thermally generated and scale linearly with temperature. Internal diffusion currents are present in all devices independent of whether the dominant current flow mechanism is by drift (which causes almost no noise) or by diffusion (which causes noise)[1].

11.1 Noise Definition

Noise is considered to be any unwanted excitation of a circuit. It comes from both external sources and internal sources. Noise from external sources appears because of unintended coupling of the circuit with other parts of the physical world; noise from internal sources appears because of unpredictable microscopic events that occur in the devices that constitute the circuit. In principle, noise from the former can be eliminated through careful design whereas noise from the latter can be reduced but never eliminated. It is important to consider noise in the design of low-power systems because the signal levels (voltages

1 Some parts of this chapter were taken from Sarpeshkar et al. (1993), White noise in MOS transistors and resistors, IEEE Circuits and Devices Magazine.

or currents) are small. The noise level sets the size of the smallest signal that can be processed meaningfully by a physical system.

The amount of noise in a signal $x(t)$ is characterized by its root-mean-square (RMS) value. It is computed by calculating the mean-square variation of the signal about its mean value:

$$\overline{\Delta x^2} = \lim_{T->\infty} \frac{1}{T} \int_0^T (x(t) - \overline{x})^2 \, dt \tag{11.1.1}$$

where $\overline{x}$ is the mean of the signal. The RMS value of the signal is then $\sqrt{\overline{\Delta x^2}}$. The total noise in a system caused by independent noise generators can be computed by adding up the mean-square deviations from these noise generators:

$$\overline{\Delta x^2} = \overline{\Delta x_1^2} + \overline{\Delta x_2^2}. \tag{11.1.2}$$

In the equation above, we assumed two independent noise sources in the system.

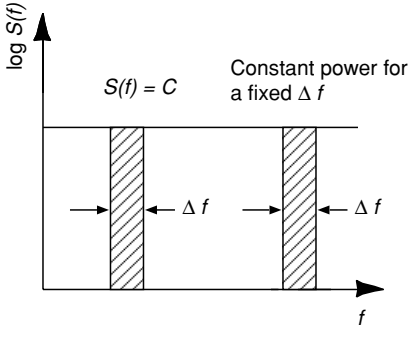


Figure 11.1
White noise spectrum. The spectrum is flat over all frequencies.

Power spectral density One way of characterizing the amount of noise in a system is by computing the noise power over the frequency spectrum. This function is also called the *power spectral density (PSD)* of the noise. It describes how much power the noise waveform carries in each frequency interval. The PSD ($S_x(f)$) of the noise is defined as the average power carried by $x(t)$ in a one-Hertz bandwidth around f, or the power per unit frequency.

This spectrum is obtained by applying the noise waveform to a bandpass filter with center frequency f_1 and a 1-Hz bandwidth; squaring the output; and calculating the average power over a long time to obtain $S_x(f_1)$. By repeating the measurements for filters with differing center frequencies, we obtain the PSD $S_x(f)$. Because each value of $S_x(f)$ is measured for a 1-Hz bandwidth, we express $S_x(f)$ in units of V^2/Hz. It also common to show the noise spectrum of a system in terms of the root of $S_x(f)$, hence units of $V/\sqrt{Hz}$. The total noise power P_x which has units of V^2 is obtained by integrating $S_x(f)$ over the whole frequency range[2]:

$$P_x = \int_0^\infty S_x(f)df. \tag{11.1.3}$$

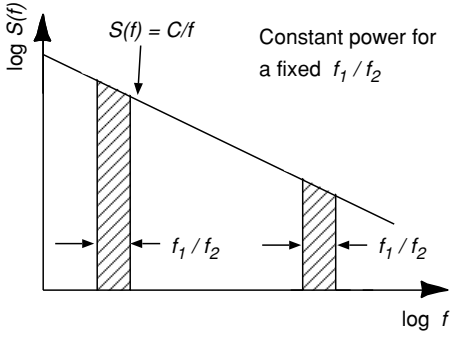


Figure 11.2
Flicker noise spectrum. The noise power is constant for a fixed ratio of frequencies.

There are two distinct classifications of noise spectra: *White noise* and *pink noise*. We call noise "white" if its power is spread uniformly across the spectrum (a flat spectrum), (see Fig. 11.1) and "pink" if the noise power is concentrated at lower frequencies. The most studied pink noise is *flicker noise* which is also known as $1/f$ *noise*[3]. The spectrum for flicker noise is shown in Fig. 11.2.

2 We call P_x the noise power even though the units are in V^2 because we have disregarded the resistor. Of course, the signal can also be, for example, current, light intensity, or sound pressure.

3 Other forms of pink noise include avalanche noise and burst noise.

White Noise

Although the spectrum of white noise is defined to be flat over the entire frequency range, in reality, such a spectrum does not exist. We consider any noise spectrum that is flat in the region of interest as white noise. In conventional literature, white noise in electronic devices is assumed to be composed of two possible types: **Thermal Noise** and **Shot Noise**.

Thermal Noise Thermal noise is believed to be due to random thermal motions of carriers. In a resistor, the thermal noise current $\overline{\Delta I_R{}^2}$ can be modeled by a noise current source in parallel with the noiseless resistor (Fig. 11.3) and described by

$$\overline{\Delta I_R{}^2} = 4k\,T\,G\,\Delta f \tag{11.1.4}$$

where k is the Boltzmann constant, G is the conductance, and Δf is the bandwidth. The units of the PSD $S(f) = \overline{\Delta I_R{}^2}/\Delta f$ are A^2/Hz. The noise spectrum of the resistor is shaped by the circuit of which the resistor is a part. An example of this calculation is shown in Section 11.5.

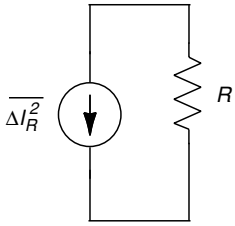


Figure 11.3
Model of thermal noise in a resistor.

Shot Noise Shot noise is believed to be due to discrete random arrivals of electrons traversing an energy barrier. Shot noise is believed to require D.C. current flow whereas thermal noise is believed to require no D.C. current flow. The shot noise current (derived later in the text) in a device is given by

$$\overline{\Delta I^2} = q\overline{I}\Delta f \tag{11.1.5}$$

where q is the charge on the electron, $\overline{I}$ is the mean current flowing through the device, and Δf is the system bandwidth.

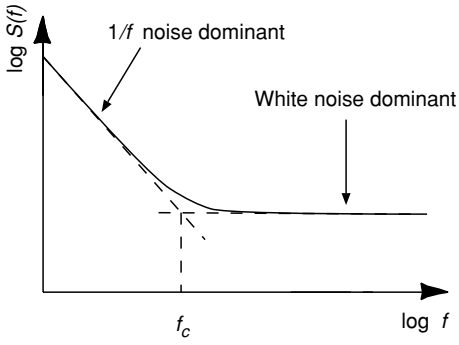


Figure 11.4
Noise spectrum of a device is a mixture of flicker noise and thermal noise. The $1/f$ noise corner f_c is process and bias dependent.

Flicker Noise

Flicker noise is also called $1/f$ **noise** because its power spectrum varies inversely with the frequency. It is found in all active devices and some discrete passive devices. Flicker noise has different origins. In MOSFETs, it is widely believed that it arises from electrons in the channel moving into and out of surface states, and into and out of impurities or defect traps in the gate oxide. It is always associated with a DC current and has the form

$$\overline{\Delta I^2} = K\frac{I^m}{f^n}\Delta f \tag{11.1.6}$$

where K is a constant for a particular device, I is the current in the device, m is between 0.5 to 2 and n is approximately 1. Further details can be found in Gray et al. (2001). This noise is dominant at low frequencies. However, it can also dominate at higher frequencies if the current in the device is large.

Usually, the noise in a device is a mixture of white noise and $1/f$ noise. The spectrum of the noise is as shown in Fig. 11.4. The $1/f$ noise corner or f_c is process and bias dependent.

11.2 Noise in Subthreshold MOSFETs

Noise in MOS transistors is composed of both white noise and flicker noise. We first discuss the formulation for flicker noise in a subthreshold MOSFET, and then the formulation for shot noise.

Flicker Noise

We compute the flicker noise in a MOSFET by looking at the probability of a charge carrier in the channel tunneling in and out of defect traps in the gate oxide. If we assume that there is a uniform density ρ of traps, then a charge carrier can enter or leave a trap at a time scale set by the tunneling probability

$$f \propto e^{-\alpha x} \tag{11.2.1}$$

where α is a constant, x is the distance from the bulk Si–SiO_2 interface, and f is the fluctuation frequency. By taking the natural logarithm of Eq. 11.2.1, we get

$$\log f \propto -\alpha x \Rightarrow \frac{df}{f} \propto -\alpha dx. \tag{11.2.2}$$

For a fixed frequency interval df, the total amount of trapped charge Q will fluctuate as $\sqrt{A\,\rho\,dx}$ where A is the area of the gate oxide, so the mean-square deviation in the charge is

$$\overline{\Delta Q^2} \propto A\,\rho\,dx \propto A\,\frac{df}{f}. \tag{11.2.3}$$

We can think of the trapped charge as modulating the surface potential and hence, the threshold voltage of the transistor V_T:[4]

$$\overline{\Delta {V_T}^2} \propto \frac{\overline{\Delta Q^2}}{C^2} \tag{11.2.4}$$

where C is the capacitance of the gate oxide. A larger area of the transistor leads to a larger oxide capacitance C and a smaller effect of any one fluctuating charge on the transistor's threshold voltage. However, the larger transistor area also leads to more flu ctuating charges because of the constant trap and defect densities. So the increased-capacitance effect reduces the noise power like $1/A^2$, and the increased total-charge effect increases the noise power like A, thus

$$\overline{\Delta {V_T}^2} \propto \frac{A}{A^2}\,\frac{df}{f} \propto \frac{1}{A}\,\frac{df}{f}. \tag{11.2.5}$$

4 We can view $1/f$ noise as noise due to a dynamically changing threshold voltage.

The mean-square deviation in the current due to the deviation in the threshold voltage can then be computed as

$$\overline{\Delta I^2} \propto g_m^2 \, \overline{\Delta {V_T}^2} \propto \frac{g_m^2}{A} \frac{df}{f} \tag{11.2.6}$$

where g_m is the transconductance of the transistor. Refer to Chapter 3 for the definition of transconductance. *Note that this equation also applies to the above threshold MOSFET.* In a subthreshold MOSFET where $g_m = \kappa \, I / U_T$,

$$\boxed{\overline{\Delta I^2} = K \, I^2 \frac{df}{f}} \tag{11.2.7}$$

where the constant K includes U_T; κ; the gate oxide capacitance per unit area; and also the effects of typical offsets in MOS technology which lead to offsets between transistors. These mismatches scale inversely with the area of the transistor. The equivalent voltage noise is

$$\boxed{\overline{\Delta {V_n}^2} = \frac{K'}{C_{ox}^2 A} \frac{df}{f}} \tag{11.2.8}$$

where K' is a constant and C_{ox} is the oxide capacitance per unit area [5].

Shot Noise

A formula for subthreshold noise in MOS transistors has been derived by Enz (1989) and Vittoz (1990) from considerations that model the channel of a transistor as a series of resistors. The integrated thermal noise of all these resistors yields the net thermal noise in the transistor, after some fairly detailed mathematical manipulations. The expression obtained for the noise, however, strongly suggests that the noise is really "shot noise", conventionally believed to be a different kind of white noise from thermal noise.

In this section, we show how one generates a shot-noise answer from a thermal-noise derivation by taking a fresh look at noise in subthreshold MOS transistors. We then rederive the expression for thermal noise in a resistor from our viewpoint. The derivation here leads to a unifying view of the processes of shot noise (noise in vacuum tubes, photo diodes and bipolar transistors) and thermal noise (noise in resistors and MOS devices).

We also show noise measurements in a subthreshold transistor. The mea-

5 Depending on the theory for the origin of flicker noise, the C_{ox} variable in the equation for flicker noise is often equal to 1 or 2 (Tsividis, 1998).

surements were taken at current levels in the 100 fA–100 pA range. White noise was the only noise observable even at frequencies as low as 1 Hz. (Reimbold, 1984) and (Schutte and Rademeyer, 1992) have measured noise for higher subthreshold currents (> 4 nA), but have reported results from flicker-noise measurements only.

We will show that measurements of white noise in subthreshold transistor operation are consistent with theoretical predictions. We also show measurements of noise in photoreceptors (a circuit containing a photodiode and an MOS transistor) that are consistent with theory. The photoreceptor noise measurements illustrate the intimate connection of the equipartition theorem of statistical mechanics with noise calculations.

The measurements of noise corresponding to miniscule subthreshold transistor currents were obtained by conveniently performing them on a transistor with $W/L \approx 10^4$. The photoreceptor noise measurements were obtained by amplifying small voltage changes with a low-noise high-gain on-chip amplifier.

Imagine that you are an electron in the source of an nFET. You shoot out of the source, and if you have enough energy to climb the energy barrier between the source and the channel, you enter it. If you are unlucky, you might collide with a lattice vibration, surface state, or impurity and fall right back into the source. If you do make it into the channel you will suffer a number of randomizing collisions. Eventually, you will actually diffuse your way into the drain. Each arrival of such an electron at the drain contributes an impulse of charge.

Similarly, electrons that originate in the drain may find their way into the source. Thus, there are two *independent* random processes occuring simultaneously that yield a forward current I_f from drain to source, and a reverse current I_r from source to drain. A detailed discussion of these currents can be found in Chapter 3. Since the barrier height at the source is smaller than the barrier height at the drain, more electrons flow from the source to drain than vice-versa and $I_f > I_r$. The channel current I is given by

$$I = I_f - I_r.$$

By using the subthreshold current equation $I_r = I_f\, e^{-\frac{V_{ds}}{U_T}}$ (V_{ds} is the drain-

to-source voltage), we obtain

$$\begin{aligned} I &= I_f \left(1 - e^{-\frac{V_{ds}}{U_T}}\right) \\ &= I_{sat} \left(1 - e^{-\frac{V_{ds}}{U_T}}\right) \end{aligned} \tag{11.2.9}$$

where $I_f = I_{sat}$ is the saturation current of the transistor, and $U_T = kT/q$ is the thermal voltage.

Because the forward and reverse processes are independent, we can compute the noise contributed by each component of the current separately and then add the results. Thus, we first assume that I_r is zero, or equivalently that N_d, the concentration of electrons per unit width at the drain end of the channel, is zero. The arrival of electrons at the drain can be modelled by a Poisson process with an arrival rate λ. A small channel length L, a large channel width W, a large diffusion constant for electrons D_n, and a large concentration of electrons at the source N_s, all lead to a large arrival rate. Because the current in a subthreshold MOS transistor flows by diffusion, the electron concentration is a linear function of distance along the channel. The forward current I_f and arrival rate λ are given by

$$I_f = qD_nW\frac{N_s}{L} \tag{11.2.10}$$

$$\lambda = I_f/q. \tag{11.2.11}$$

Powerful theorems due to Carson and Campbell, as described in standard noise textbooks such as van der Ziel (1970), allow us to compute the power spectrum of the noise. Suppose that each arrival event in this Poisson process causes a response $F(t)$ in a detector sensitive to the event. Let $s(t)$ be the macroscopic variable of the detector that corresponds to the sum of all the $F(t)$'s generated by these events. Then, the mean value of $s(t)$ and the power

spectrum $P(f)$ of the fluctuations in $s(t)$ are given by

$$\overline{s(t)} = \lambda \int_{-\infty}^{\infty} F(t)dt \tag{11.2.12}$$

$$\overline{\left(s(t) - \overline{s(t)}\right)^2} = \lambda \int_{-\infty}^{\infty} F^2(t)dt \tag{11.2.13}$$

$$= 2\lambda \int_{0}^{\infty} |\psi(f)|^2 df \tag{11.2.14}$$

$$= \int_{0}^{\infty} P(f)df \tag{11.2.15}$$

where $\psi(f) = \int_{-\infty}^{+\infty} F(t)e^{-j2\pi ft}dt$ is the Fourier transform of $F(t)$. We have used Parseval's theorem and the symmetry of the Fourier transform for positive and negative frequencies in obtaining the last equation. Each electron arrival event at the drain generates an impulse of charge q that corresponds to $F(t)$: that is, $F(t) = q\,\delta(t)$. Thus, we obtain

$$\overline{I} = q\lambda \tag{11.2.16}$$

$$\overline{\left(I - \overline{I}\right)^2} = 2q^2\lambda \int_{0}^{\Delta f} df \tag{11.2.17}$$

$$= 2q\overline{I}\Delta f \tag{11.2.18}$$

where Δf is the bandwidth of the system. Equation 11.2.18 is the well-known result for the shot-noise power spectrum. Thus, the noise power[6] that corresponds to our forward current is simply given by $2qI_f\Delta f$. Similarly, the noise power that corresponds to the reverse current is given by $2qI_r\Delta f$. The total noise in a given bandwidth Δf is given by

$$\begin{aligned}\overline{\Delta I^2} &= 2q(I_f + I_r)\Delta f \\ &= 2qI_f(1 + e^{-\frac{V_{ds}}{U_T}})\Delta f \\ &= 2qI_{sat}(1 + e^{-\frac{V_{ds}}{U_T}})\Delta f\end{aligned} \tag{11.2.19}$$

where $I_{sat} = I_f = I_0 e^{\frac{\kappa V_g - V_s}{U_T}}$ corresponds to the saturation current at the given gate voltage. Note that as we transition from the linear region of the transistor ($V_{ds} < 4U_T$) to the saturation region, the noise is gradually reduced from $4qI_{sat}\Delta f$ to $2qI_{sat}\Delta f$. This factor of two reduction occurs because

6 Again, we have disregarded the influence of the resistor for convenience's sake and we call this measure "noise power".

the shot-noise component from the drain disappears in saturation. A similar phenomenon occurs in junction diodes where both the forward and reverse components contribute when there is no voltage across the diode; as the diode gets deeply forward or reverse biased, the noise is determined primarily by either the forward or reverse component, respectively (Robinson, 1974).

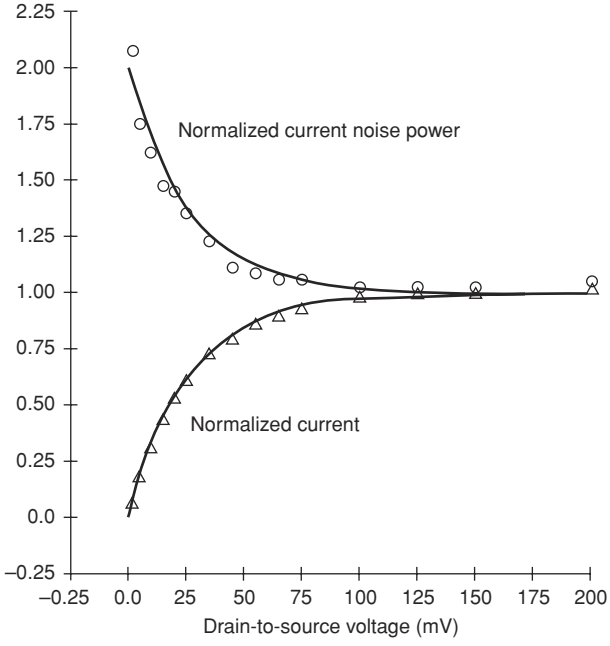


Figure 11.5
Measured current and noise characteristics of a subthreshold MOS transistor. The lower curve is the current, normalized by its saturation value I_{sat}, so that it is 1.0 in saturation and zero when V_{ds} is 0. The upper curve is the noise power $\overline{\Delta I^2}$ normalized by dividing it by the quantity $2qI_{sat}\Delta f$, where Δf is the bandwidth and q is the charge on the electron. We see that as the transistor moves from the linear region to saturation, the noise power decreases by a factor of two. The lines are fits to theory using the measured value of the saturation current and the value for the charge on the electron $q = 1.6 \times 10^{-19}$ C.

The flatness of the noise spectrum arises from the impulsive nature of the microscopic events. We might expect that the flat Fourier transform of the microscopic events that make up the net macroscopic current would be reflected in its noise spectrum. Carson's and Campbell's theorems express formally that this is indeed the case. The variance of a Poisson process is proportional to the

rate, so it is not surprising that the variance in the current is just proportional to the current. Further, the derivation illustrates that the diffusion constant and channel length simply alter the arrival rate by Eq. 11.2.11. Even if some of the electrons recombined in the channel (corresponding to the case of a bipolar transistor or junction diode) the expression for noise in Eq. 11.2.19 is unchanged. The arrival rate is reduced because of recombination. A reduction in arrival rate reduces the current and the noise in the same proportion. The same process that determines the current also determines the noise.

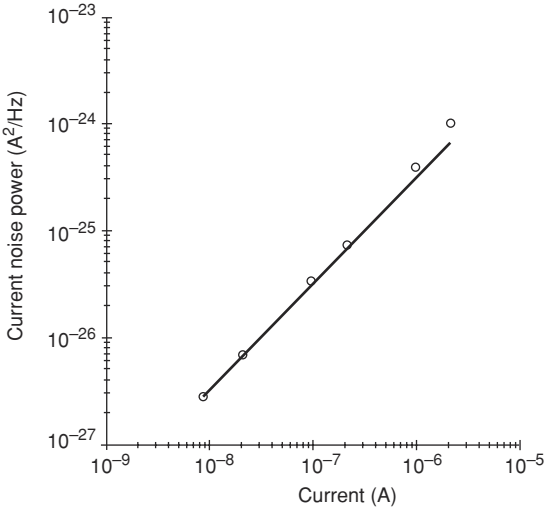


Figure 11.6

The noise power per unit bandwidth $\overline{\Delta I^2}/\Delta f$ plotted against the saturation current I_{sat}.The MOS transistor is operated in saturation. Theory predicts a straight line with a slope of $2q = 3.2 \times 10^{-19}$ C, which is the line drawn through the data points. The small but easily discernible deviations from the line increase with higher levels of I_{sat} due to the increasing levels of $1/f$ noise at these current values.

Experimental measurements were conducted on a transistor with $W/L \approx 10^4$ at a saturation current of 40 nA (Fig. 11.5). A gigantic transistor was used to scale up the tiny subthreshold currents to 10 nA–1 μA levels and make them easily measurable by a low-noise off-chip sense amplifier with commercially available resistor values. The shot noise scales with the current level, so long as the transistor remains in subthreshold. The noise measurements were conducted with a HP3582A spectrum analyzer. The data were taken

over a frequency range of 0–500 Hz. The normalized current noise power $\overline{\Delta I^2}/(2qI_{sat}\Delta f)$ and the normalized current I/I_{sat} are plotted. The lines show the theoretical predictions of Eqs. 11.2.9 and 11.2.19. Using the measured value of the saturation current, the value for the charge on the electron, and the value for the thermal voltage we were able to fit our data with no free parameters whatsoever. Notice that as the normalized current goes from 0 in the linear region to 1 in the saturation region, the normalized noise power goes from 2 to 1 as expected. Figure 11.6 shows measurements of the noise power per unit bandwidth $\overline{\Delta I^2}/\Delta f$ in the saturation region for various saturation currents I_{sat}. Since we expect this noise power to be $2qI_{sat}$, we expect a straight line with slope $2q$ which is the theoretical line drawn through the data points. As the currents start to exceed 1 μA–10 μA for our huge transistor, the presence of $1/f$ noise at the frequencies over which the data were taken begins to be felt. The noise is thus higher than what we would expect purely from white noise considerations.

11.3 Shot Noise versus Thermal Noise

We have taken the trouble to derive the noise from first principles even though we could have simply asserted that the noise was just the sum of shot-noise components from the forward and reverse currents. We have done so to clarify answers to certain questions that naturally arise:

- Is the noise just due to fluctuations in electrons moving across the barrier or does scattering in the channel contribute as well?
- Do electrons in the channel exhibit thermal noise?
- Do we have to add another term for thermal noise?

Our derivation illustrates that the computed noise is the total noise and that we need not add extra terms for thermal noise. Our experiments confirm that this is indeed the case. The scattering events in the channel and the fluctuations in barrier crossings at the source and drain ends of the channel all result in a Poisson process with some electron arrival rate. Both processes occur simultaneously, and are caused by thermal fluctuations, resulting in white noise. Conventionally, the former process is labelled “thermal noise” and the latter process is labelled “shot noise”. In some of the literature, the two kinds of noise are often distinguished by the fact that shot noise requires the presence of a DC current while thermal noise occurs even when there is no

DC current (Gray and Meyer, 1993). However, we notice in our subthreshold MOS transistor that when $I_f = I_r$, there is no net current but the noise is at its maximum value of $4qI_f\Delta f$. Thus a two-sided shot noise process exhibits noise that is reminiscent of thermal noise. We will now show that thermal noise is shot noise due to internal diffusion currents in electronic devices.

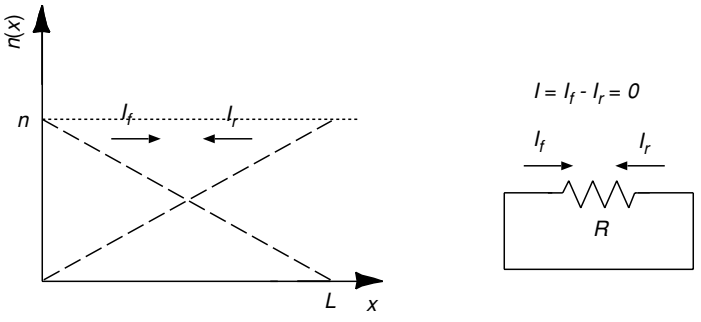


Figure 11.7
Model for thermal noise. The figure on the left shows the concentration per unit volume of electrons diffusing from both ends of the resistor. We can approximate this scenario by a resistor whose total current is zero in the presence of diffusion of carriers.

New Derivation of Thermal Noise in a Resistor

Let us compute the noise current in a resistor shorted across its ends as shown in Fig. 11.7. Since there is no electric field, the fluctuations in current must be due to the random diffusive motions of the electrons. The average concentration of electrons is constant all along the length of the resistor. This situation corresponds to the case of a subthreshold transistor with $V_{ds} = 0$, where the average concentrations of electrons at the source edge of the channel, drain edge of the channel, and all along the channel, are at the same value and therefore the current in the resistor is $I = 0$.

In a transistor, the barrier height and the gate voltage are responsible for setting the concentrations at the source and drain edges of the channel. In a resistor, the concentration is set by the concentration of electrons in its conduction band. The arrival rate of the Poisson process is, however, still determined by the concentration level, diffusion constant, and length of travel. This is so because, in the absence of an electric field, the physical process of diffusion is responsible for the motions of the electrons. Thus, the power

spectrum of the noise is again given by $2q(I_f + I_r)$. The currents I_f and I_r are both equal to $qDnA/L$ (see Fig. 11.7) where D is the diffusion constant of electrons in the resistor, n is the concentration per unit volume, A is the area of cross section and L is the length. Einstein relation yields $D/\mu = kT/q$, where μ is the mobility.

Thus, the relation between the transistor noise power and its conductance is given by

$$\begin{aligned}\overline{\Delta I^2} &= 4\,q\,I_f\,\Delta f = 4\,q\,\frac{qDnA}{L}\,\Delta f \\ &= 4\,q\,\mu kTn\,\frac{A}{L}\,\Delta f = 4\,k\,T\,(q\mu n)\,\frac{A}{L}\,\Delta f \\ &= 4\,k\,T\,(\sigma)\,\frac{A}{L}\,\Delta f\end{aligned}$$

where σ is the conductivity of the material. This equation further reduces to

$$\boxed{\overline{\Delta I^2} = 4\,k\,T\,G\,\Delta f} \tag{11.3.1}$$

where $G = \sigma\,\frac{A}{L}$ is the conductance of the resistor. Thus, we have re-derived Johnson and Nyquist's well-known result for the short circuit noise current in a resistor! The key step in the derivation is the use of the Einstein relation $D/\mu = kT/q$. This relation expresses the connection between the diffusion constant D, which determines the forward and reverse currents, the mobility constant μ, which determines the conductance of the resistor, and the thermal voltage kT/q.

The Einstein relation shows that at zero temperature, $D = 0$, there are no internal diffusion currents, and there is no noise. This prediction is in accord with that obtained from the conventional view of thermal noise. It is because of the internal consistency between thermal noise and shot noise that formulas derived from purely shot noise considerations agree with those derived from purely thermal noise considerations (Enz, 1989).

Our derivation suggests that in a resistor with DC current flowing through it, the noise is still described by internal diffusion currents in the resistor, which are barely changed by the small systematic drift velocities superposed upon the huge thermal velocities of the electrons. Thus, even if there is a DC electric field in the resistor, so long as this field is not large enough to elicit hot-electron behavior, the thermal noise of the resistor is unchanged. Our derivation also explains why it is unwise to add a shot noise term $2qV_{DC}/R$ to the thermal noise term $4kTG$ to get the resistor's PSD at a particular DC voltage V_{DC}.

Thus the current noise PSD in a resistor with DC voltage across it remains at $4kTG$ as is experimentally observed. Nyquist's derivation of thermal noise is valid only in thermal equilibrium and would be unable to predict the noise in a resistor with a DC voltage across it unlike our derivation. Furthermore, our derivation implies that, contrary to what is believed, a DC current flow is not necessary to observe shot noise. Devices that have internal diffusion currents such as resistors exhibit shot noise with no DC current flow, but this noise gets relabeled as "thermal noise".

In summary, white noise in electronic devices is caused by shot noise due to thermally generated diffusion currents in these devices, independent of whether the noise is labeled as "thermal noise" or "shot noise". There is no need to double count various forms of white noise once the shot noise due to diffusion currents has been accounted for.

Our discussion has focused on white noise in semiclassical electronic devices. It is possible to have shot noise statistics without any thermal process: For example, perfectly coherent light generates photons with Poisson statistics due to the fundamental randomness inherent in discrete quantum phenomena.

11.4 The Equipartition Theorem and Noise Calculations

No discussion of thermal noise would be complete without a discussion of the equipartition theorem of statistical mechanics, which lies at the heart of all calculations of thermal noise: Every state variable in a system that is not constrained to have a fixed value is free to fluctuate. The thermal fluctuations in the current through an inductor or the voltage on a capacitor represent the state-variable fluctuations in an electrical system. If the energy stored in the system corresponding to state variable x is proportional to x^2, then x is said to be a degree of freedom of the system. Thus, the voltage on a capacitor constitutes a degree of freedom, since the energy stored on it is $CV^2/2$. Statistical mechanics requires that if a system is in thermal equilibrium with a reservoir at temperature T, then each degree of freedom of the system will have a fluctuation energy of $kT/2$. Thus, the mean square fluctuation $\overline{\Delta V^2}$, in the voltage of a system with a single capacitor must be such that

$$\begin{aligned} \frac{C\overline{\Delta V^2}}{2} &= \frac{kT}{2} \\ \Rightarrow \overline{\Delta V^2} &= \frac{kT}{C}. \end{aligned} \tag{11.4.1}$$

This simple and elegant result shows that if all noise is of thermal origin, and the system is in thermal equilibrium, then the total noise over the entire bandwidth of the system is determined just by the temperature and capacitance (Rose, 1973). If we have a large resistance coupling noise to the capacitor, the noise per unit bandwidth is large but the entire bandwidth of the system is small. If we have a small resistance coupling noise to the capacitor, the noise per unit bandwidth is small but the entire bandwidth of the system is large. Thus, the total noise which is the product of the noise per unit bandwidth $4kTR$ and the bandwidth of the circuit $\frac{1}{RC}$ is constant and independent of R. We will illustrate for the particular circuit configuration of Fig. 11.8 how the noise from various devices interact to yield a total voltage noise of kT/C.

Figure 11.8 shows a network of four transistors all connected to a capacitor, C at the node V_s. We use the sign convention that the forward currents in each transistor flow away from the common source node, and the reverse currents flow toward the common source node[7]. The gate and drain voltages, V_{gi} and V_{di}, respectively are all held at constant values. Thus, V_s is the only degree of freedom in the system. Kirchhoff's current law at the source node requires that in steady state

$$\sum_{i=1}^{n} I_f^i = \sum_{i=1}^{n} I_r^i. \tag{11.4.2}$$

The conductance of the source node is given by

$$g_s = \sum_{i=1}^{n} \frac{I_f^i}{U_T}. \tag{11.4.3}$$

The bandwidth Δf of the system is then

$$\Delta f = \frac{1}{2\pi} \times \frac{\pi}{2} \times \frac{g_s}{C} = \frac{g_s}{4C} \tag{11.4.4}$$

where the factor $1/2\pi$ converts from angular frequency to frequency, and factor $\pi/2$ corrects for the rolloff of the first-order filter not being abrupt[8]. Thus, the

7 Our sign convention is for carrier current, not conventional current. Thus, in an nFET or a pFET, the forward current is the one that flows away from the source irrespective of whether the carriers are electrons or holes. This convention results in a symmetric treatment of nFETs and pFETs.

8 Use

$$\int_0^\infty \frac{df}{1 + \left(\frac{f}{f_c}\right)^2} = \frac{\pi}{2} f_c.$$

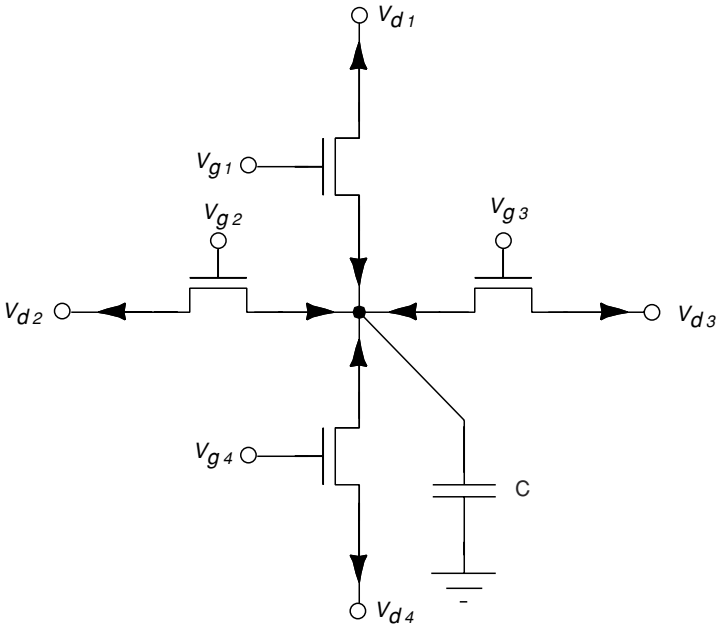


Figure 11.8
A circuit with four transistors connected to a common node with some capacitance C. By convention, the common node is denoted as the source of all transistors, and the forward currents of all transistors are indicated as flowing away from the node while the reverse currents of all transistors are indicated as flowing towards the node. Only the voltage V_s is free to fluctuate, and all other voltages are held at fixed values, so that the system has only one degree of freedom. The equipartition theorem of statistical mechanics predicts that if we add the noise from all transistors over all frequencies to compute the fluctuation in voltage $\overline{\Delta V_s^2}$ the answer will equal kT/C no matter how many transistors are connected to the node, or what the other parameters are, so long as all the noise is of thermal origin. We show in the text and in the data reported in Fig. 11.9 that our expressions for noise yield results that are consistent with this prediction.

total noise is

$$\begin{aligned}\overline{\Delta I^2} &= \sum_{i=1}^{n} 2q\left(I_f^i + I_r^i\right)\frac{g_s}{4C} \\ &= \sum_{i=1}^{n} 4qI_f^i\frac{g_s}{4C}\end{aligned} \tag{11.4.5}$$

where we have used Eq. 11.4.2 to eliminate I_r. The voltage noise is just

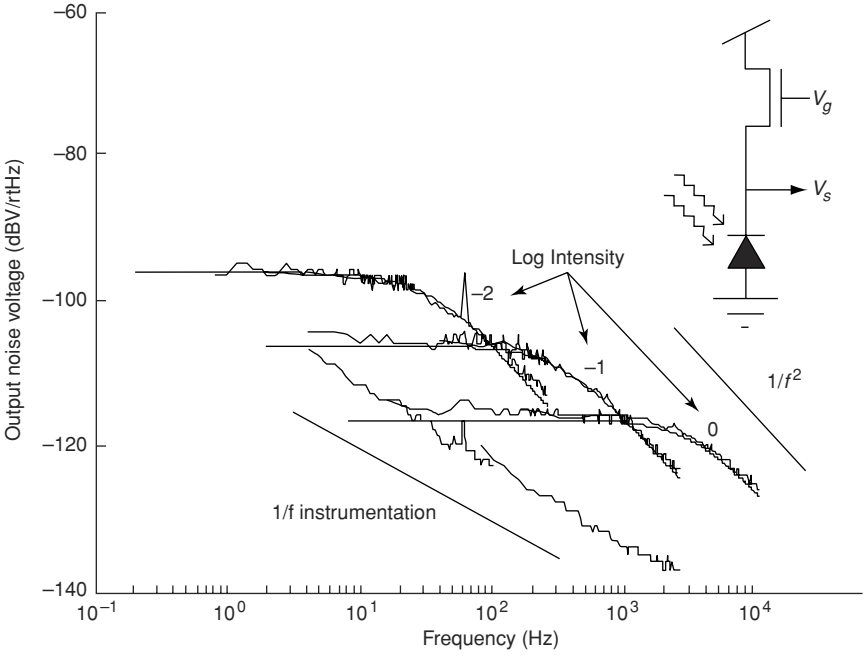


Figure 11.9
Measured noise spectral density in units of dBV/rtHz (0 dBV = 1V, -20dB = 0.1V) for the voltage V_s in the circuit above. The current source is light-dependent and the curves marked 0, -1 and -2 correspond to bright light (high current), moderate light, and dim light (low current) respectively. The intensity levels were changed by interposing neutral density filters between the source of the light and the chip to yield intensities corresponding to 1.7 W/m^2, 0.17 W/m^2, and 0.017 W/m^2 respectively. The $1/f$ instrumentation noise is also shown: Its effects were negligible over most of the range of experimental data. We observe that the noise levels and bandwidth of the circuit change so as to keep the total noise constant: That is, at low current levels, the voltage noise is high and bandwidth low, and the converse is true for high current levels. Thus, the area under the curves marked 0, -1 and -2 are the same. The theoretical fits to the lowpass filter transfer functions are for a temperature of 300° K and a capacitance of 310 fF, estimated from the layout. These results show that the kT/C concept, derived from the equipartition theorem is correct.

$\overline{\Delta I^2}/g_s^2$ or

$$\overline{\Delta {V_s}^2} = \frac{\frac{q}{C}\sum_{i=1}^{n} I_f^i}{\sum_{i=1}^{n} \frac{I_f^i}{U_T}} = \frac{kT}{C}. \tag{11.4.6}$$

The fact that the total noise is kT/C implies that this circuit configuration is

a system in thermal equilibrium. Typically, most circuit configurations yield answers for total voltage noise that are proportional to kT/C.

Noise Measurements in a Photoreceptor Circuit

We obtained direct experimental confirmation of the kT/C result from our measurements of noise in photoreceptors. A detailed theoretical and experimental analysis of noise in different photoreceptor circuits can be found in Delbrück and Mead (1995). Figure 11.9 shows a source-follower configuration which is analogous to the case discussed previously with two transistors connected to a common node. The lower current source is a photodiode which has current levels that are proportional to light intensity. (This photoreceptor circuit was previously described in Section 10.4.) The voltage noise is measured at the output node, V_s. The voltage V_g is such that the MOS transistor shown in the figure is in saturation and its reverse current is zero. The photodiode contributes a shot-noise component of $2qI\Delta f$. Thus, we obtain equal shot-noise components of $2qI\Delta f$ from the transistor and light-dependent current source, respectively. The theory described in Section 11.4 predicts that independently of the current, the total integrated noise over the entire spectrum must remain constant. We observe that as the current levels are scaled down (by decreasing the light intensity), the noise levels rise but the bandwidth falls by exactly the right amount to keep the total noise constant. The system is a lowpass filter with a time constant set by the light level. Thus, the noise spectra show a lowpass filter characteristic. The voltage noise level per unit bandwidth $\overline{\Delta {V_s}^2}$ proportional to $\overline{\Delta I^2}/{g_s}^2$ or to $1/I$, and the bandwidth is proportional to g_s and therefore to the photocurrent I. The product of the noise per unit bandwidth and the bandwidth is proportional to the total noise over the entire spectrum and is independent of I. Consequently, the area under all three curves in Fig. 11.9 is the same. The smooth lines in Fig. 11.9 are theoretical fits using a temperature of 300°K and a value of capacitance estimated from the layout.

It is possible to extend our way of thinking about noise to the above-threshold region of MOS operation as well. However, the mathematics is more difficult because the presence of a longitudinal electric field causes non-independence between the noise resulting from forward and reverse currents. Further, the modulation of the surface potential by the charge carriers results in a feedback process that attenuates fluctuations in the mobile charge concentration; an effect that is referred to as *space-charge smoothing*. A simpler approach to the problem is described in Tsividis (1998) and yields what we

believe to be a mathematically correct answer.

$$\boxed{\overline{\Delta I^2} = 4\,k\,T\,\mu\,\frac{W}{L}\,\overline{Q}\,\Delta f} \tag{11.4.7}$$

where $\overline{Q}$ is the average charge per unit area.
In a subthreshold MOSFET, Eq. 11.4.7 becomes

$$\boxed{\overline{\Delta I^2} = 4\,k\,T\,\mu\,\frac{W}{L}\,\left(\frac{Q_s + Q_d}{2}\right)\,\Delta f}. \tag{11.4.8}$$

In the above threshold MOSFET, Eq. 11.4.7 becomes

$$\boxed{\overline{\Delta I^2} = 4\,k\,T\,\mu\,\frac{W}{L}\,\frac{2}{3}\,\left(\frac{Q_s^2 + Q_s Q_d + Q_d^2}{Q_s + Q_d}\right)\,\Delta f} \tag{11.4.9}$$

Eq. 11.4.7 behaves like the noise in a sheet resistor with an average charge per unit area of $\overline{Q}$, and $G = \mu\frac{W}{L}\overline{Q}$

$$\boxed{\overline{\Delta I^2} = 4\,k\,T\,G\,\Delta f.} \tag{11.4.10}$$

11.5 Noise Examples

Here we show some examples of noise calculation in simple circuits. We look at four cases: An RC circuit, a MOSFET noise model, a MOSFET inverter, and a transconductance amplifier.

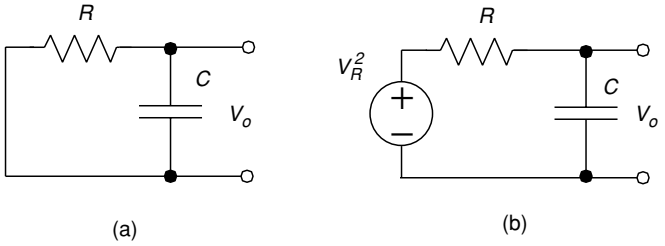


Figure 11.10
Noise in an RC circuit.

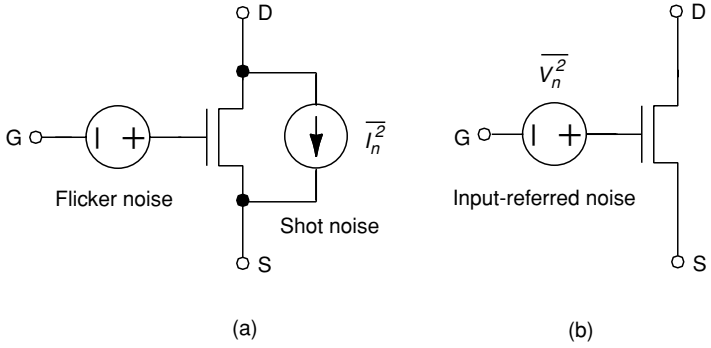


Figure 11.11
Two possible noise models for a MOSFET at low frequency. (a) A noise current source is included in parallel with the MOSFET together with a noise source in series with the gate. (b) The output noise current is converted to a input-referred noise source in series with the gate.

Noise in an RC circuit

We can compute the noise spectrum in an RC circuit as shown in Fig. 11.10. The transfer function of the RC circuit in the Laplace domain is

$$\frac{V_o}{V_R}(s) = \frac{1}{1 + RC\,s} \tag{11.5.1}$$

The noise spectrum of the resistor, $S_R(f)$ is shaped by the transfer function of the circuit. The noise power at the output, $S_o(f)$ is

$$\begin{aligned} S_o(f) &= S_R(f) \left| \frac{V_o}{V_R}(j\omega) \right|^2 \\ &= 4\,k\,T\,R \frac{1}{4\,\pi^2\,R^2 C^2 f^2 + 1} \end{aligned} \tag{11.5.2}$$

and the total noise power at the output is[9]

$$\begin{aligned} P_o &= \int_0^\infty \frac{4\,k\,T\,R}{4\,\pi^2 R^2 C^2 f^2 + 1} df \\ &= \frac{k\,T}{C}. \end{aligned} \tag{11.5.3}$$

The total RMS noise voltage at the output is given by $\sqrt{P_o}$.

9 Use $\int_0^\infty \frac{dx}{x^2+1} = \arctan(\infty) - \arctan(0) = \pi/2$ to solve this equation.

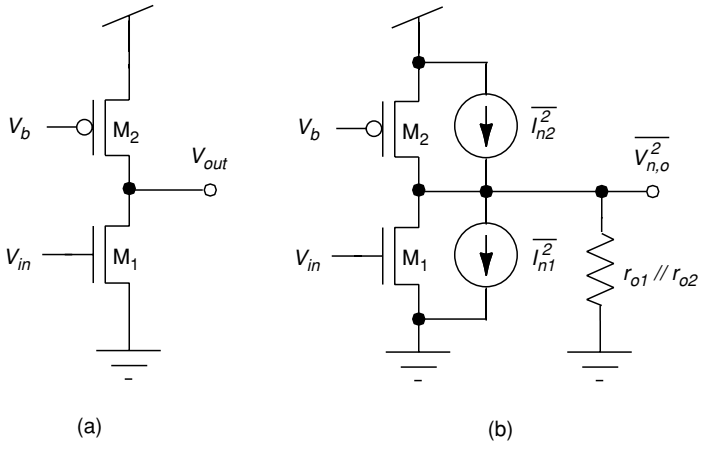


Figure 11.12
Noise model for an inverter. (a) Circuit for an inverter with the input going to the nFET, M_1. The bias voltage to the pFET, M_2 is constant. (b) Noise current sources added to the transistors. The output impedance at the output is the parallel combination of the output conductances of the transistors.

Noise Model of MOSFET

The noise calculations so far refer to the output-referred noise, $V_{n,out}$. Unfortunately, this figure depends on the gain of the circuit, A_v. A better figure of merit would be to compute the "input-referred" noise, $V_{n,in}$. Hence, one can take the total output-referred noise which is caused all the noise sources in the circuit and by dividing this output noise by the gain, we get only one noise voltage source at the input, that is, $\overline{V_{n,in}^2} = \overline{V_{n,out}^2}/A_v^2$. Notice that this input noise source does not exist in the physical circuit.

The noise in a MOSFET at low frequencies comes from two different sources: white noise and flicker noise. The noise in a MOSFET can be modeled by a white noise current source in parallel with the transistor (Fig. 11.11(a)) and a flicker noise source in series with the gate. The current noise can be converted to an input-referred voltage noise V_{sn} by dividing the current noise I_n by the transconductance of the transistor g_m: That is, $V_{sn}^2 = I_n^2/g_m^2$. Assuming that the transistor is in saturation and using Eqs. 11.2.7 and 11.2.19, we can replace these two noise sources by an input-referred noise source as

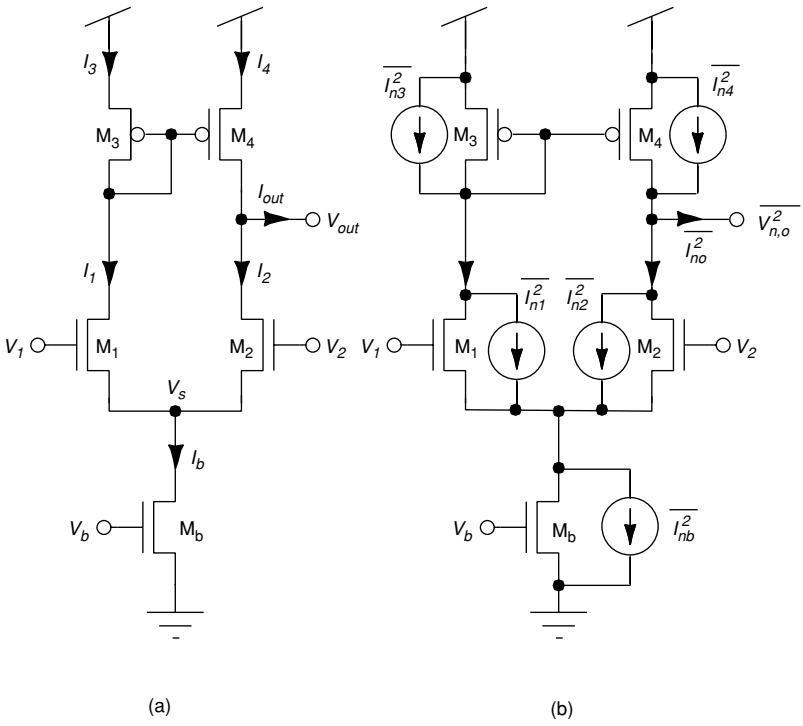


Figure 11.13
Noise model for a transconductance amplifier. (a) Circuit for the amplifier. (b) Noise current sources added to the transistors.

shown in Fig. 11.11(b) where

$$\boxed{\overline{V_n^2} = \frac{2kT}{\kappa g_m} + \frac{K}{WLf}}. \tag{11.5.4}$$

So by increasing the transistor's transconductance and the area of the transistor's gate, we can decrease the input-referred noise voltage.

Noise in an Inverter

We now show the noise calculation in an inverter comprising of an nFET M_1 and a pFET M_2 (Fig. 11.12). The input V_{in} goes to the gate of M_1 and the transistor M_2 acts as a current source. The voltage, V_b, determines the magnitude of this current source. In Fig. 11.12(b), the noise current sources associated

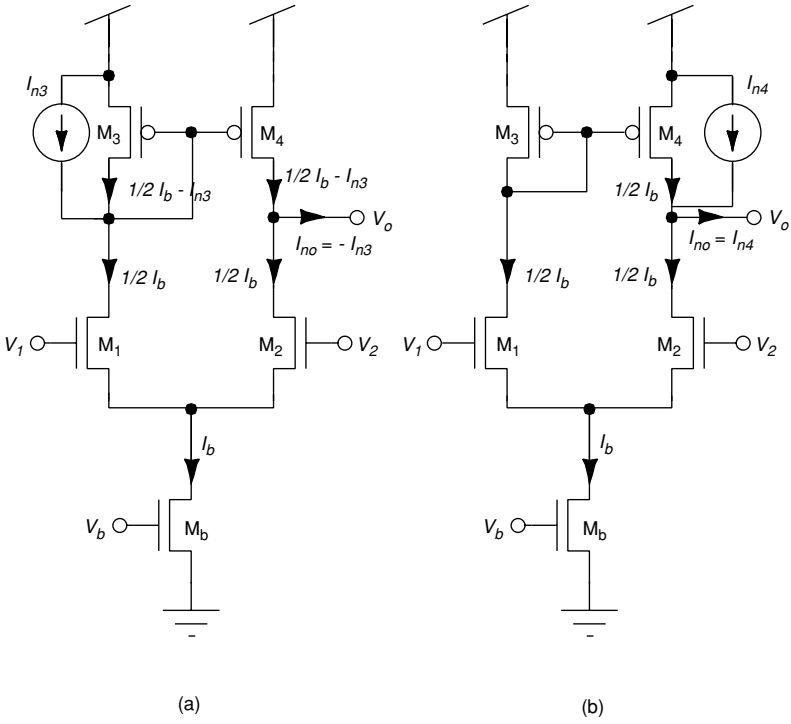


Figure 11.14
Noise current analysis. (a) Noise current analysis for M_3. (b) Noise current analysis for M_4.

with the MOSFETs are added to the circuit. For each noise source, we compute the ac transfer function between its current source and the differential output current. The total output current noise is the sum of the squares of the current noise for each source. The input-referred voltage noise per unit bandwidth is given by the total output current noise divided by g_m. The total voltage noise is then computed by integrating the input-referred voltage noise over the bandwidth of the circuit.

Assuming that these noise sources are uncorrelated and that the subthreshold transistors are operating in saturation, we get

$$\overline{V_{n,o}^2} = 2\,q\,(I_{sat1} + I_{sat2})\,(r_{o1}//r_{o2})^2 \tag{11.5.5}$$

where I_{sat1} and I_{sat2} are the saturation currents of M_1 and M_2 respectively.

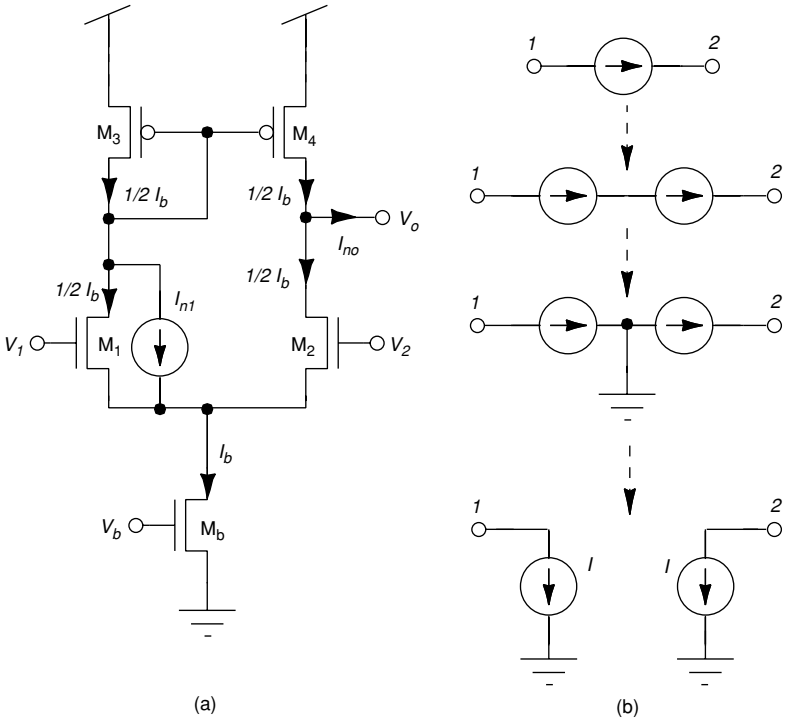


Figure 11.15
Noise current analysis. (a) Noise current analysis for M_1. (b) Circuit transformation for a floating current source.

Substituting the relationship $I_{sat} = g_m U_T/\kappa$ into Eq. 11.5.5:

$$\overline{V_{n,o}^2} = 2\,k\,T\left(\frac{g_{m1}+g_{m2}}{\kappa}\right)(r_{o1}//r_{o2})^2 \tag{11.5.6}$$

The voltage gain of the inverter is $g_{m1}\,(r_{o1}//r_{o2})$. By dividing the output-referred noise by the gain, we get the noise voltage referred to the gate of the input:

$$\overline{V_{n,i}^2} = 2\frac{k\,T}{\kappa}\left(\frac{1}{g_{m1}}+\frac{g_{m2}^2}{g_{m1}}\right). \tag{11.5.7}$$

Thus, to reduce the noise in this inverter circuit, we should increase the transconductance of the input transistor M_1 and reduce the transconductance of the current source M_2. Note, however that, in subthreshold $g_{m2}/g_{m1} = 1$

because the transconductance is invariant with geometry and depends only on bias current, which is equal for both devices. However, the same ideas may be used in an above threshold noise analysis to show that a large W/L for transistor M_1 and a small W/L for transistor M_2 lead to a reduced g_{m2}/g_{m1} ratio that helps attenuate the noise.

Flicker noise is also part of the output noise current. The constant K in Eq. 11.2.7 is different for nFETs and pFETS. Because we operate the circuit in subthreshold with small bias currents thermal noise dominates over flicker noise.

Noise in a Transconductance Amplifier

The circuit for the transconductance amplifier is shown in Fig. 11.13(a). The noise sources associated with each transistor are included with the corresponding transistor in Fig. 11.13(b). We next compute the ac transfer function between each current source and the differential output current. The noise source I_{nb} flows through two symmetrical pathways; one through M_1 and the second through M_2. These currents cancel at the output node. Hence this noise current does not contribute to the output noise. Let us consider next the noise current, I_{n3} as shown in Fig. 11.14(a). The current through M_4 is then $\frac{I_b}{2} - I_{n3}$ so that the total current sums to $\frac{I_b}{2}$. This current is mirrored to the output through M_4. Therefore, $I_{no} = -I_{n3}$. The noise current, I_{n4}, flows directly to the output so $I_{no} = I_{n4}$. A similar analysis can be performed for the noise current I_{n4} in Fig. 11.14(b).

The effect of the noise current from M_1 on the output node is more difficult to analyze. The current I_{n1} flows partially through M_1 and partially through M_2 (Fig. 11.15(a)). To simplify the analysis, we make use of a circuit transformation technique where we split the floating current source I_{n1} into 2 currents (Fig. 11.15(b)). This technique is used when we want to transform some portion of a circuit so that we can apply Thevenin's theorem or Norton's theorem to it (Kelly, 1970; Van Valkenburg and Kinariwala, 1982; Chua et al., 1987). The resulting circuit is shown in Fig. 11.16. The noise at the output is again I_{n1}. A similar analysis can be performed for the noise source I_{n2}. The total output current noise $\overline{I_{n,o}^2}$ is the sum of the squares of the current noise for the remaining four noise sources:

$$\overline{I_{n,o}^2} = \overline{I_{n1}^2} + \overline{I_{n2}^2} + \overline{I_{n3}^2} + \overline{I_{n4}^2} \tag{11.5.8}$$

The input-referred voltage noise per unit bandwidth is given by the total

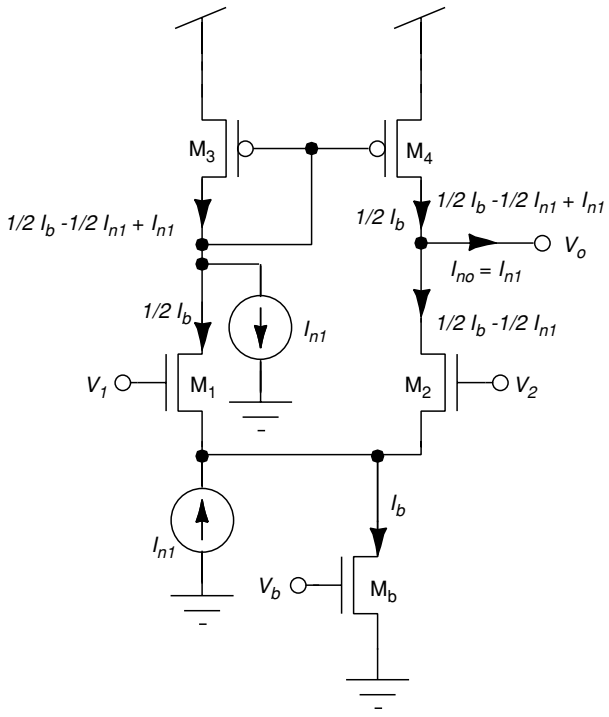


Figure 11.16
Modified circuit for noise current analysis for M_1.

output current noise divided by the transconductance of the amplifier, $g_m = \kappa I_b/2U_T$. Just as in the noise example in the inverter, we have ignored the flicker noise.

12 Layout Masks and Design Techniques

Now that we have seen how to design simple circuits, we describe the different steps that lead to fabrication of the circuit. To fabricate a circuit, we need to generate layout which describes the geometries required for the successive layers of fabrication. The fabrication house uses these layers to generate masks that, for example, determine the different areas on the chip where the polysilicon gates will be deposited; and which areas will be doped p-type or n-type; and where to place the wiring between the nodes on the two-dimensional Si substrate. The fabrication steps are described in Chapter 13. In the first half of the chapter, we describe how to generate the different fabrication layers from a circuit diagram. In the second half of the chapter, we describe some layout techniques that reduce the sources of noise and errors in the fabricated circuit. Successful layout of analog circuits is a difficult task, and a detailed coverage of all the issues involved fills entire books (Hastings, 2001).

12.1 Mask Layout for CMOS Fabrication

Different CMOS processes are similar in that they produce the same types of physical structures. However, each process has its own parameters, such as the thicknesses and different dopings of the different layers. Processes may also differ in some of the processing steps that are used to optimize the performances of the fabricated devices. Furthermore, some processes have more layers and options than others.

For a given fabrication process, a set of physical structures are available. The designer has full control over the placement of some of these structures as projected onto a plane parallel to the semiconductor surface. The position, size and composition of the structures along the dimension perpendicular to the semiconductor surface is specified by the manufacturer in the *process parameters*. These parameters include layer thicknesses and doping profiles.

The designer-specified structures are defined by a set of *binary masks*, where each mask determines the location of a layer in the projection of the circuit onto a plane parallel to the semiconductor surface. These masks correspond to some of the masks that are used for the fabrication of the circuit. Other masks used for processing are dependent on the masks defined by the designer and are generated from them by the manufacturer. The foundry makes available guidelines for the layout of the masks to be generated by the designer.

These guidelines are called *layout rules* or *design rules*. These rules specify the minimum and maximum dimensions, and spacings for the different masks that are recommended to ensure that the circuit conforms to the manufacturer's specifications.

The mask set is used to fabricate electrical circuits near the surface of a uniformly doped *wafer* of semiconductor crystal, which is referred to as the *substrate* or *bulk*. Nowadays, most CMOS processes use a *p*-type substrate. Chapter 13 describes how the mask set is used for the processing of the wafer.

Mask Layers

For a typical process, the designer has to specify a total of about ten masks. The most commonly specified masks are the following:

Well

Wells serve as *local substrates* (or *local bulks*) for MOSFETs. Wells with opposite doping from the substrate may also be used as resistors or to design diodes to the substrate (e.g. photodiodes). Some processes provide a well for only one MOSFET type, while the other MOSFET type is located directly in the substrate (*single-well* or *single-tub* processes). Other processes provide a well for each MOSFET type (*twin-well* or *twin-tub* processes). For example, in a *p*-substrate process, pFETs are fabricated in *n*-wells and nFETs are either fabricated in the substrate or in *p*-wells.

Active

Active regions are all the regions that do not have a thick insulator layer (*field oxide* on the surface of the semiconductor crystal. These are the regions where the structures in the semiconductor are supposed to interact electrically with the ones that are deposited above it. Active regions include the sources, channels, and drains of the MOSFETs, and the contact regions to the wells and substrate. In some processes, active regions can also be used as resistors if they are surrounded by regions of the opposite doping type.

Select

Active regions are implanted by dopants, except where they are blocked by gates. The select mask determines which doping type a given active region receives. The boundaries of the masking regions are usually drawn outside

the areas that are implanted. These areas are defined by the boundaries of the active mask and poly mask (see below). Select typically specifies one doping type, and the active regions that are not covered by select automatically receive doping of the other type, because the manufacturer uses a mask for it that is complementary to the select mask. In some processes, two select masks can be specified; one for each doping type. Active regions without select are then not implanted by either dopant. Some manufacturers do not require the definition of separate active and select masks, but use an n^+ *active* and a p^+ *active* mask instead.

Poly

Poly is a conductor layer that is electrically insulated from the substrate and is used for the gates of the MOSFETs and for interconnects. In some processes, poly may also be used to implement resistors. In modern silicon CMOS processes this layer is made from doped *polycrystalline silicon*, hence the name.

Poly2

Poly 2 is a second conductor layer separated by a thin insulator layer from the poly layer below it and is provided in some processes used for analog circuits. It has the same applications as the poly layer (except that in some processes it may not be used for MOSFET gates). However, poly2 is mainly used to provide the option of designing poly-poly2 capacitors, which have a relatively large capacitance per unit area and have a lower common-mode sensitivity than poly-active capacitors. The poly2 layer is also useful for the fabrication of charge-coupled devices, which typically have overlapping electrodes. Standard CMOS processes used for logic circuits do not usually provide a poly2 layer, because it significantly increases fabrication cost[1].

Metal

The metal layer provides the electrical interconnections between the different circuit structures. It has a higher conductivity than the poly layer and is therefore preferable for longer-range interconnects.

Contact

The contact mask specifies holes in the insulator layer, where the metal layer is to be electrically contacted to the active, poly, or poly2 region beneath it. In

1 Typically by about 5% to 10%.

some cases, different contact masks are specified, depending on which layer is to be contacted.

Metal2

Some processes provide one or more (metal3, metal4, etc.) additional metal layers for electrical interconnections. They are separated by reasonably thick insulator layers from the structures beneath them.

Via

Vias are electrical contacts between metal layers, that is, holes in the separating insulator layers. Processes providing more than two metal layers have more than one via (via2 connects metal2 with metal3, etc.).

Overglass

The entire circuit is covered by a thick insulator layer to prevent the degradation of the fabricated structure by the interaction with its environment. This insulator layer is called a *passivation layer* and in most circuits has holes only at the locations of the electrical contacts of the circuit to the outside world, which are called *bonding pads*. The overglass layer is usually drawn where the holes in the passivation layer are intended. It should then actually be called *overglass cut*. The passivation layer absorbs electromagnetic radiation and additionally has a strongly wavelength-dependent optical transmission due to interference effects. It may therefore be desirable to use overglass cuts at the locations of photodiodes and of areas where exposure to ultraviolet radiation (for example, to facilitate electron tunneling) will occur.

The above masks are sufficient for the designer to implement a circuit in a standard CMOS process[2]. The other masks needed for processing are generated from these masks by the manufacturer according to process-specific rules. Figure 12.1 shows the mask layout of an inverter in a single-well process using an active mask and a select mask that specifies the same doping type that is used for the well. A few additional options may be available. Some processes that do not provide a poly2 layer use a *capacitor implant*[3] into the semiconductor bulk as a capacitor plate to form linear capacitors with poly. Some *silicided* processes (see Chapter 13) provide a *silicide block* mask. This

2 If you are lucky. Some processes require the specification of additional layers.

3 A capacitor implant is an impurity doping implant, but it differs from the source/drain implants by the fact that it happens before the poly deposition and can therefore be located underneath the poly.

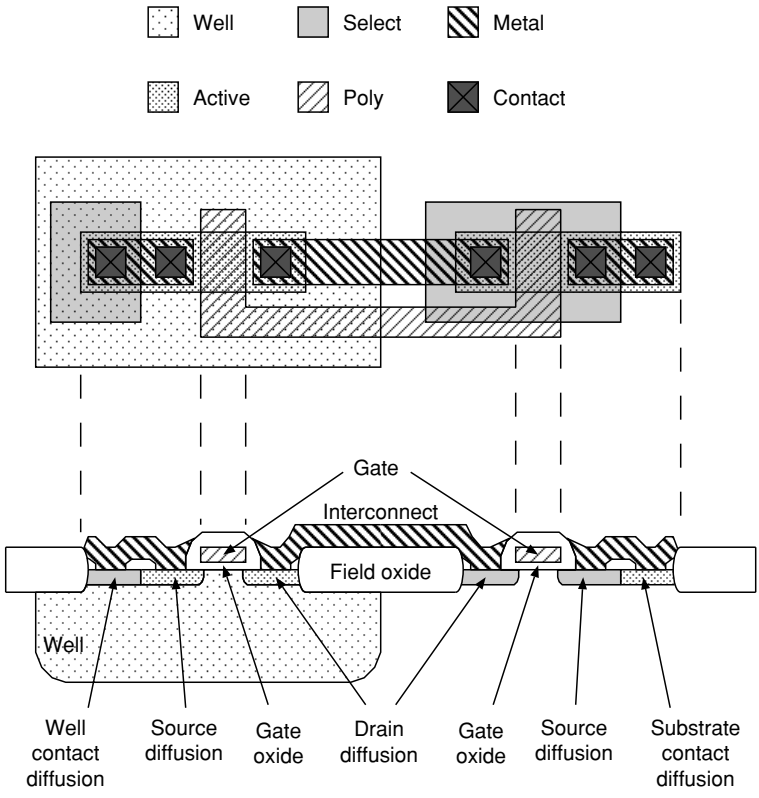


Figure 12.1
Mask layout and cross-section of an inverter in a single-well process. The mask layout uses a single select mask for the active doping of the same type as the well doping. The shown cross-section horizontally cuts through the center of the circuit. The inverter consists of two MOSFETs of opposite types, that have connected gates and drains. The source of the well MOSFET is connected to the well and the source of the MOSFET in the substrate is connected to the substrate.

feature is useful for the design of photosensors, since silicided layers may block most of the incoming light.

There is a class of processes, called *BiCMOS* (**Bi**polar **CMOS**) processes, which offer vertical bipolar junction transistors (BJTs) in addition to MOSFETs, in the same substrate. These processes provide a *base implant* with the same doping type as the substrate. In order to construct a BJT in such a technology, a source/drain implant is used as the *emitter*; the base implant as the *base*; and a well as the *collector*. Only one type of BJT is obtainable by this

simple enhancement and the resulting BJTs do not exhibit a very good performance, given that the emitter and collector layers are not optimized for BJTs.

A few processes also provide the option of implementing *buried-channel charge-coupled devices (BCCDs)*. They offer a lightly doped *BCCD implant* for the buried layer (see Section 10.5). Such CMOS-compatible BCCDs are of a lower quality than commercial ones, which are produced with dedicated CCD processes, and are therefore not suitable for high-resolution imagers. However, CMOS-compatible BCCDs are much better than the surface-channel CCDs that are fabricated using standard CMOS technology, because the density of imperfections at the semiconductor-insulator interface of a typical CMOS process is too large for efficient multi-stage charge transfer along the interface.

The software packages used for circuit layout allow the designer to draw the required masks with a two-dimensional graphical interface that represents the projection of the circuit onto a plane parallel to the semiconductor surface. The different masks are represented by polygons of different colors and/or different stipple patterns. A polygon marks the location where the corresponding binary mask assumes one of its two values (for example, where it is transparent to the processing step), while the absence of such a polygon denotes the other value of the mask. In addition to the mask layers, layout programs use a host of *symbolic layers*, which are ignored by the manufacturer. These layers have different functions. Some of them are logical combinations of the mask layers and are either used for automatic *design rule checks (DRCs)* or for *circuit extraction*, that is, the generation of a symbolic description of the circuit on the device level (for example, in *SPICE* format). Circuit extraction is used mainly to verify if the functionality of a circuit layout matches that of the circuit schematic that it implements. Other symbolic layers are used by the graphical user interface, for example, to mark the locations of design rule violations.

12.2 Layout Techniques for Better Performance

The proper layout techniques should be employed to ensure good device matching and circuit operation on a chip. Fabricated circuits frequently do not work exactly as in simulation. The layout of these circuits is important to minimize parasitic effects and undesirable coupling effects. Every node on a circuit has parasitic capacitances and resistances which should be taken into

account during design. The values of these parasitic elements depend on the layout of the circuit. Additionally, if devices are to be matched in a design, the layout of these devices is important in minimizing the effect of mismatches. Processing variations across a wafer can create mismatches in devices that are to be matched.

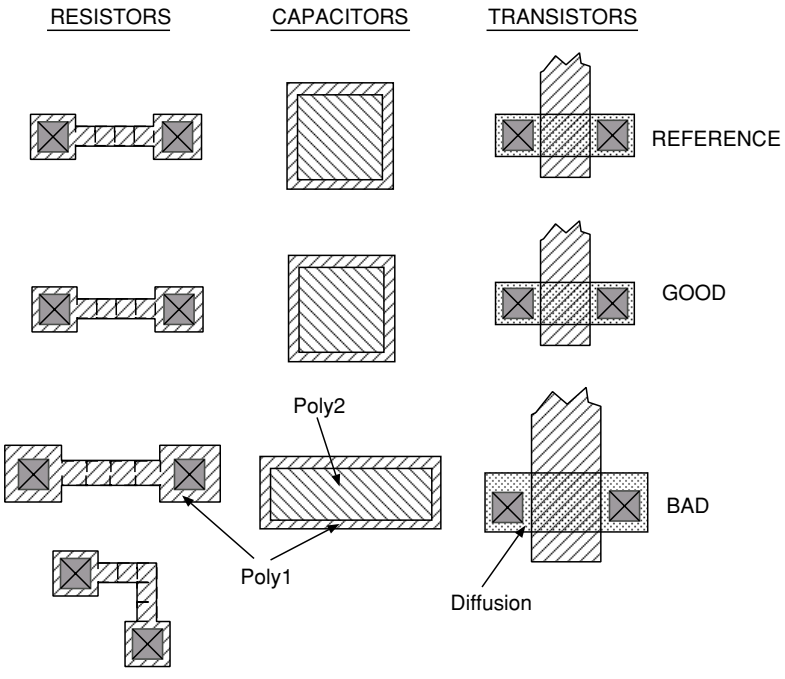


Figure 12.2
Layout elements. The first row shows the reference elements. The second row shows elements that are matched well to the reference set and the third row shows elements that are badly matched to the reference set.

Another major source of problems comes from mixed analog and digital circuits that are placed on the same chip. Analog circuits are sensitive to small changes in voltage; digital circuits typically swing over the entire power supply range when they switch. The different operating ranges of analog and digital circuits should be taken in consideration when planning the layout of the chip. The switching of a digital circuit can couple into an analog circuit and affect its operation. Digital circuits should be have their own power supply lines and

analog circuits should be placed on a different part of the chip, preferably, as far as possible from the digital circuits, to minimize coupling effects.

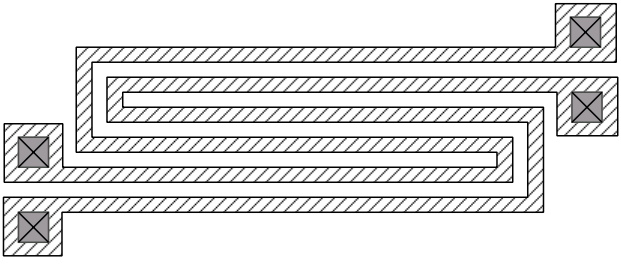


Figure 12.3
Resistors to be matched should be placed closed to one another.

In the following section, we describe some layout techniques that ensure good device matching for better circuit performance. We also describe some suggestions for layout techniques that apply to mixed-mode circuits on the same chip. The techniques described in this chapter have been developed by veteran circuit designers (Vittoz, 1985; Tsividis, 1996).

Device Matching

Device Size Matching Devices to be matched should have the same dimensions. For example, two transistors that should have the same width/length ratio should have the same lengths and same widths (Fig. 12.2). The matching in each dimension is necessary because the dimensions of a fabricated transistor are different from the drawn dimensions. The parasitic capacitances at each node are proportional to the area, so even if two transistors have the same dimension ratio, the parasitic capacitances will be unequal if one transistor has twice the dimensions of the other. The same criterion applies to capacitors and resistors. Capacitors are usually made of a bottom poly1 plate and a top poly2 plate with the interpoly oxide as the insulator. Resistors can be made out of polysilicon or well. Two resistors with unequal dimensions will have different parasitic capacitances. Jogs in resistors are not desirable because etching at sharp corners is not isotropic. Capacitors to be matched should not be drawn with minimum dimensions as the actual dimensions will change after fabrication. To reduce the effects of the variation in the dimensions, two matched capacitors should be of reasonable width and length and have the same di-

mensions. If one capacitor has the same area but a different perimeter, the capacitance values might be unequal after fabrication.

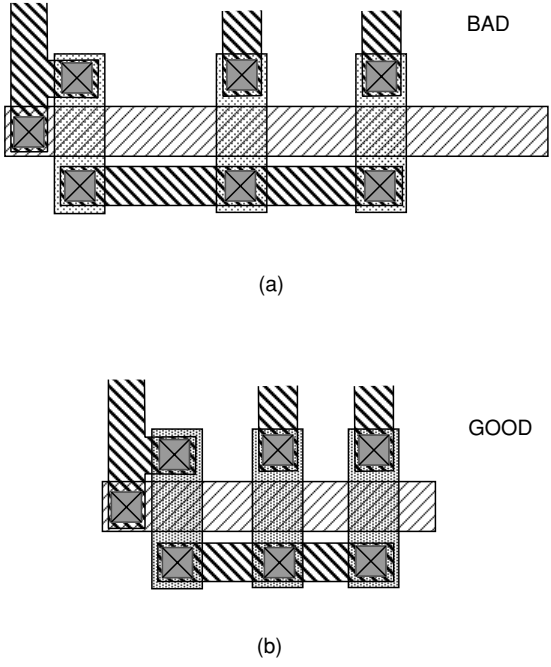


Figure 12.4
Transistors in a current-mirror circuit should be placed closed to one another. (a) Bad layout. (b) Good layout.

Minimum Distance Because of the spatial variation in the process parameters, devices to be matched should be placed next to one another. For example, long resistors should be interleaved together (see Fig. 12.3) and the input transistors in a current mirror circuit should be placed closed together (shown in Fig. 12.4).

Same Orientation Devices should be placed in the same orientation so that they will be subjected to similar effects due to anisotropic process steps, found for example, in mask alignment and anisotropic doping. Other asymmetries come from stress after packaging. For example, transistors in a differential

pair should be matched closely in orientation (Fig. 12.5(c)). Figures 12.5(a) and (b) show some examples of poor layout for a differential pair.

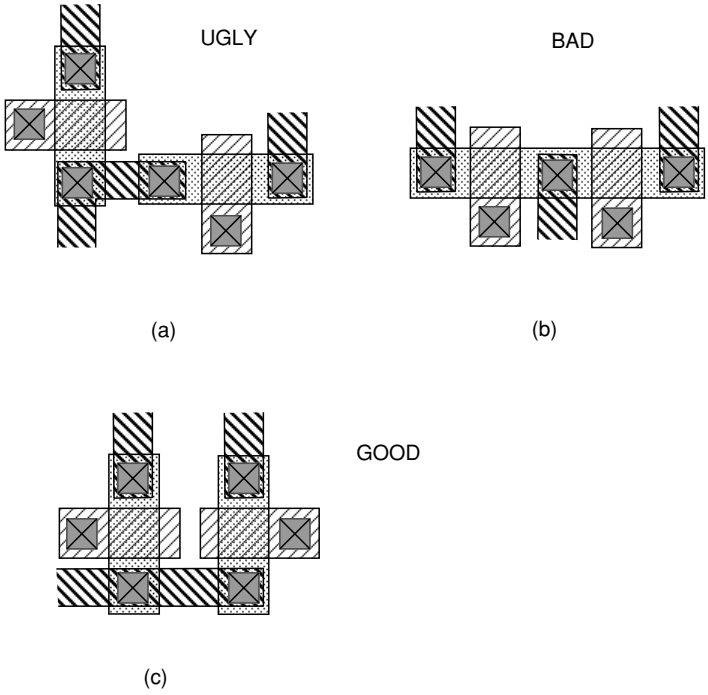


Figure 12.5
To ensure good matching, the input transistors in a differential pair should be placed as close as possible to one another. In addition, the transistors should be in the same orientation so that they are subjected to same alignment mismatch. (a) Poor layout of the transistors. The orientation of these transistors is orthogonal to each other. (b) Better but still bad layout; the current in the two transistors flows in opposite directions. (c) Good layout; the orientation and the current flow are the same for both transistors.

Common-Centroid Geometry Preferably, transistors in a differential pair circuit should be placed in a common centroid arrangement (see Fig. 12.6). However, to satisfy this requirement, the layout cost is large. This type of arrangement is reasonable only for small numbers of transistors.

Common Surroundings Matched devices should have the same surroundings (Fig. 12.7). The edge effects, for example, on capacitors will be different if the capacitors are surrounded by different circuitry. It is best to put dummy capacitors on the edges as shown in the figure.

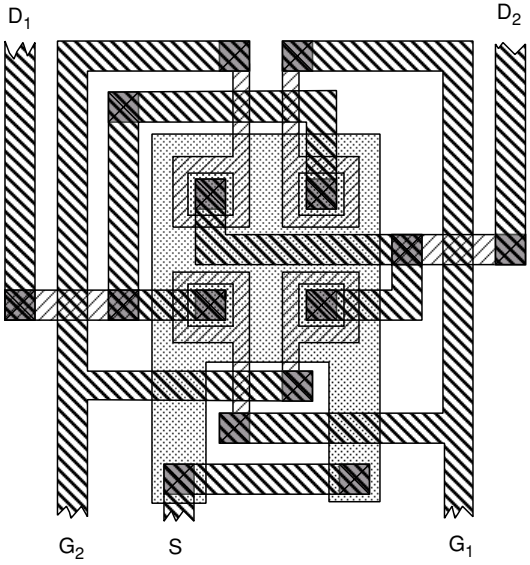


Figure 12.6
Layout example to show the common-centroid arrangement of the input transistors in a differential pair.

Common Temperature If there is a source of heat on the chip, place matched devices on the same isotherm line.

12.3 Short List of Matching Techniques

A summary of these layout techniques for matching devices based on ratios of values is as follows:

1. They should have the same dimensions.
2. They should have the same structure. For example, we should not try to

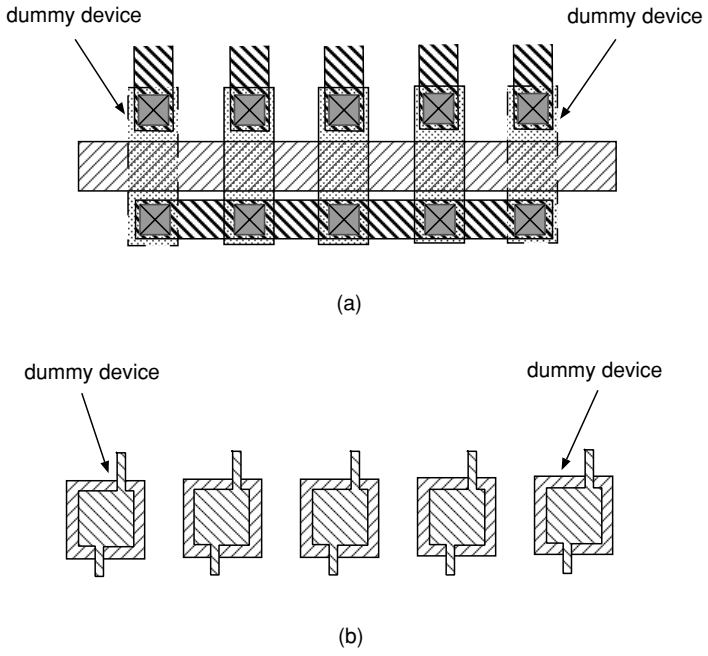


Figure 12.7
To ensure good matching, the matched transistors and capacitors should be surrounded by identical elements.

match nFETs to pFETs.

3. They should not be of minimum size because minimum-sized devices are affected more by random factors in processing.

4. They should be placed close to one another.

5. They should have the same orientation, for example, the current flow should be in the same direction.

6. They should be preferably be laid out in a common-centroid arrangement. This arrangement is especially important for large transistors.

7. The surrounding circuitry should be similar.

8. They should be at the same temperature.

12.4 Parasitic Effects

The capacitances and resistances of the drawn circuits are supplemented by parasitic elements from the circuit layout. The parasitic capacitances come from the inherent capacitance between the different layers of the fabricated circuit. The parasitic resistors exist because of the finite resistivity of the different materials of the layers. These numbers are provided by the fabrication house. Other parasitic effects arise because of the increasingly small geometries in modern processes. For example, the sheet resistances of the drain and source nodes of a transistor cannot be ignored in a 0.25 μm CMOS technology. These parasitic effects can reduce the speed and bandwidth of the circuit, distort the circuit characteristics, reduce the circuit precision, create noise, and add to signal loss. Particular care should be paid to critical nodes to reduce or eliminate the parasitic effects. These parasitic capacitances and resistances should be taken into account when simulating the circuit performance.

One important parasitic capacitance is the capacitance from any node to the substrate because the substrate is common to all circuits on the chip. Any coupling from a node to the substrate can affect many circuits. These coupling effects are especially crucial when considering mixed-mode analog digital circuits. Digital circuits typically switch over the entire power supply range. If care is not taken, these transients can couple into analog circuits which are normally sensitive to small voltage changes. The following are some suggestions to reduce the effects of coupling.

Avoid Common Supply Lines Digital circuits and analog circuits should have separate power supply and ground lines. A bad practice is to run a common supply line to both analog and digital circuits. Large transient spikes can result because digital lines switch quickly over the power supply range. These transient currents create a voltage drop along the power supply line due to the resistance of the line and the inductance of the packaging wires. The power supply to the analog circuits will vary due to the switching in the digital circuits. This variability can be reduced by bringing different lines from the same common pad to the analog and digital circuits. Then, the switching noise from the on-chip digital circuit can be separated from the analog circuit. *An even better solution* is to bring the supply lines to separate pads on the chip, so that the interference due to the inductance of the bonding wire can be reduced. Remember, that the capacitances of the supply lines and the pads; and the inductance of the wires, can form a resonant LC circuit. This impedance will

increase if the frequency of the digital clock matches the resonant frequency of the LC circuit.

The ground line also creates a problem. The switching current of a digital circuit leads to large voltage spikes on the ground line. These voltage spikes affect the operation of circuits with low power-supply rejection ratio, or if the ground line is used to bias an input terminal. It is better to group circuits with common functionality together and have separate ground lines running to each group.

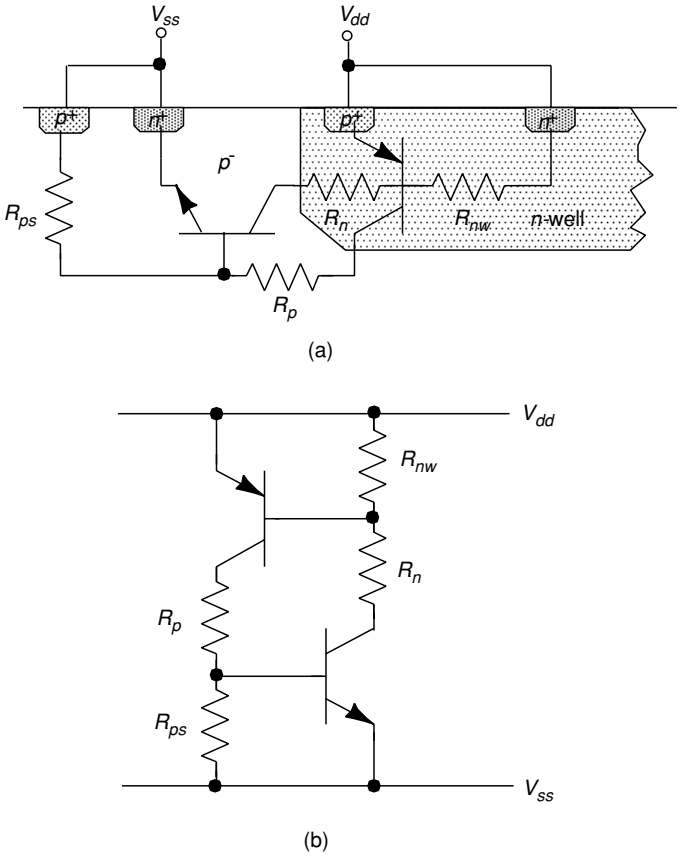


Figure 12.8
An n-well CMOS structure showing the parasitic resistances and bipolar junction transistors that contribute to the latchup mechanism. (a) Cross-section of the CMOS structure. (b) Equivalent circuit that causes latchup.

The power supply and ground lines should be wide so that these lines have low resistance. If there are large current drivers in the circuit, we can estimate the voltage transient by measuring the total transient current and multiplying by the resistance of the supply lines. A side effect of wide lines is that their capacitance increases and can cause undesirable effects like parasitic coupling to the substrate.

12.5 Latchup

CMOS structures are susceptible to a mechanism called *latchup*. This mechanism is aggravated by the parasitic resistances in the different regions of the silicon bulk. There are many structures in the bulk that can trigger *latchup*. When it is triggered, hugh amounts of current will flow through the chip. If the power supply to the chip is not current limited, then the chip will be irreversibly damaged.

Figure 12.8(a) shows the parasitic elements in a cross-section of a CMOS n-well structure that result in latchup. We show only one source/drain region of an nFET and one source/drain region of a pFET. There are two parasitic bipolar junction transistors (BJTs) in the structure. One is a *lateral* npn bipolar junction transistor and is formed by the n^+ source of an nFET, the p^- substrate, and the n-well. The other is a *vertical* pnp bipolar junction transistor and is formed by the p^+ source of a pFET, the n-well, and the p^- substrate. The bulk and well have parasitic resistances, denoted by R_{ps} and R_{nw} respectively. The equivalent circuit consisting of the BJTs and the resistances are shown in Fig. 12.8(b). This $npnp$ structure is also called a *thyristor*. The parasitic resistances are critical for turning on this structure. For example, if there is a big enough transient current on V_{dd}, this current causes a voltage drop across R_{nw} and turns on the pnp BJT. The collector current from this BJT charges up the base of the npn BJT through R_p and R_{ps}, thus also turning on this BJT. The collector current from the npn BJT further lowers the base voltage of the pnp BJT so creating a positive feedback condition and also creating a runaway process: *Latchup*. Huge currents flow in this structure and are limited only by the resistances. If the power supply is not current-limited, the circuits on chip will be destroyed. The same mechanism will occur if there is a transient on V_{ss}. *Remember that there are many structures of this type all over the chip.*

There are several solutions to combat this problem. First, we can make the parasitic resistances small by separating the n^+ and p^+ regions so that the β

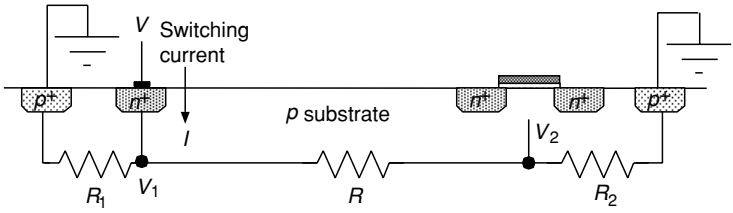


Figure 12.9
Resistive coupling through substrate. Current can be injected into the substrate from different sources. For example, if node V_1 represents the place into which the switching current from a digital transistor flows into the substrate, then node V_2 which represents the substrate of a voltage-sensitive transistor will be affected by this current through resistors, R and R_2. This coupling can be reduced by placing a substrate contact next to the node V_1 so that the substrate resistance, R_1 is small.

of the npn BJT will be small (due to the large base width). Second, substrate contacts should be placed frequently and close to the transistors so that R_{ps} and R_{nw} will be small. Layout rules from the foundry include a minimum spacing between the n^+ well contact and the n^+ region in the substrate to prevent latchup.

Different techniques are used in industry to minimise latchup. One technique is to grow a layer of lightly doped p-type material on top of heavily p-doped starting material. This lightly doped material is called an *epitaxial* (or *epi*) layer . The transistors and the n-well are formed within the *epi* layer. The implanted well is adjacent to the the heavily-doped p^+ layer so that the R_p resistance will be small. Another technique is called the *trench isolation* technique. A vertical trench filled with polysilicon is placed around the well. The trench reaches the heavily-doped p^+ area. Some technologies like the silicon-on-isulator (SOI) technology is completely immune to latchup.

12.6 Substrate Coupling

The substrate is common to all the devices on the chip so any non-desirable coupling into the substrate can create problems. There are different ways that signals can couple into the substrate: Resistive coupling (for example through an n^+–n junction), capacitive coupling, and coupling through a bipolar transistor.

Shielding from Substrate Noise A node of a digital transistor affects the substrate through the junction capacitor between the drain/source of the transistor and the substrate. If the digital transistor is close to an analog transistor, the switching on the node of the digital transistor is coupled into the substrate. Since the substrate is common to both types of transistors, this switching can couple into any sensitive nodes including resistors and capacitors, or into the substrate beneath the analog transistor. The switching affects the transistor current through the body effect. It also couples into the transistor through other parasitic capacitors of the transistor (shown in Figure 12.9). There are several ways of reducing the impact of the coupling from the digital circuit. One method is to separate the analog circuits and the digital circuits so that the resistance R between them is large (see Fig. 12.9). However, this technique is not effective when the process has an epitaxial layer. The heavily doped substrate underneath the epitaxial layer has a low resistance. An option is to use an insulating substrate like SOI.

An effective way of reducing substrate noise from digital ground is to separate the source and substrate wires to the digital transistor so that the high switching currents at the source node do not affect the substrate node.

Reducing Substrate Resistances Another way of reducing the impact of the coupling from the substrate is to reduce R_1 and R_2 in Fig. 12.9. This can be done by placing guard rings around the circuit that is sensitive to interference or that generates interference. The guard rings consist of a ring of diffusion with substrate contacts. Make sure that the ring is tied to a ground that is either brought out separately to the pad or use digital ground or analog ground, whichever one least affects the circuit.

Shielding from Minority Carriers Analog switches can generate minority carriers into the substrate when the switches are turned off. Circuits with phototransistors and photodiodes are also susceptible to minority carriers because the incoming photons generate carriers in the substrate that can diffuse to neighboring circuits if they are not collected in the transducer area. One way of reducing the effects of minority carriers is to build a guard ring around the interfering circuit or sensitive circuit so that minority carriers will be soaked up by the ring. This scenario is shown in Fig. 12.10(a).

A second well ring can also be used to reduce substrate currents, for example in photodiodes. The majority carriers in a p-type substrate are holes: In

an n-well to p-substrate junction, electrons that do not go to the well, generate substrate currents. An n-well guard ring collects these minority carriers in the substrate.

Minority carriers from analog switches can be reduced by putting a guard ring between a sensitive circuit or by placing the switches and sensitive circuits in separate wells (see Fig. 12.10).

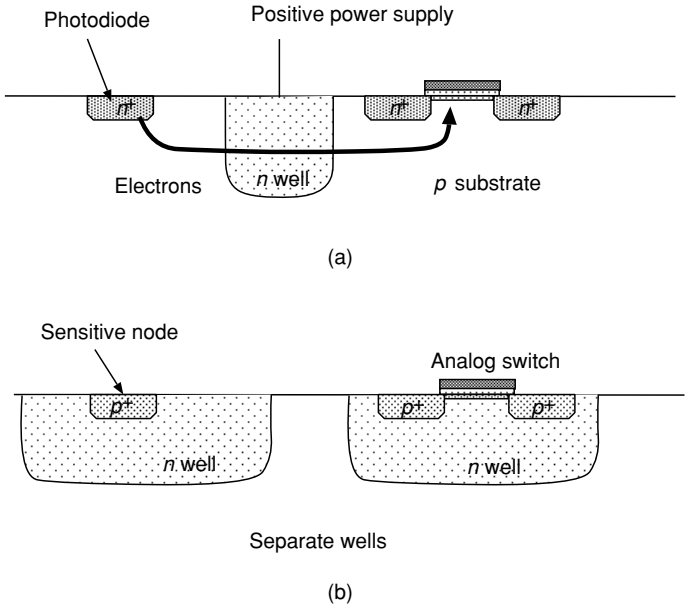


Figure 12.10
Minority carrier shielding. (a) Minority carriers generated at the junction of the photodiode can migrate into the substrate. To reduce this current from affecting other transistors, a well guard ring is placed around the photodiode area. (b) Sensitive nodes can be shielded from analog switches by placing them in separate wells.

Capacitance Shielding Both the top and bottom plates of the capacitor should be shielded from any node with high variance. The shielding of the bottom plate is especially important because the coupling into the substrate from any interfering circuits can cause excursions on the voltage on the bottom plate. One way of doing this is to place a grounded n-well under the capacitor. Shielding of resistors is also important. For example, a poly resistor can be

shielded from the top by the first-layer metal in a two-layer metal process. Transistors that are sensitive to noise should be placed in wells if possible and the wells tied to clean supplies. Remember that there is also capacitive coupling through the air so sensitive nodes should be far away from interfering nodes. Shield analog circuits from digital circuits by placing them in separate wells.

Other Important Tips

Bonding Pads Connect nodes that will brought to the outside to the closest bonding pads. Ensure that the bonding wires from pads to the package pins are short as possible.

Avoid proximity Do not place sensitive lines close to one another or cross them over each other or over any interfering lines. Avoid crossing digital input/output lines over analog lines.

Sensitive Nodes Make lines to sensitive nodes as short as possible to avoid parasitic coupling. Make lines to and from interfering circuits as short as possible: for example, the outputs of digital circuits.

Differential Circuits Make circuits differential where possible. Differential circuits are useful for reducing effects due to parasitic coupling. Make the layout of differential circuits as symmetrical as possible.

Floorplan of Chip Plan the chip before doing the layout. Place analog circuits together and digital circuits together.

12.7 Device Matching Measurements

Models to characterize mismatch in MOS transistors for precision analog design were first developed in the 1980s (Lakshmikumar et al., 1986; Pelgrom et al., 1989). The mismatch between transistors can be characterized by measuring the standard deviation of the current, $\sigma(\frac{\delta I}{I})$. This deviation can be computed from the standard deviation in the β parameter[4], $\sigma(\frac{\delta \beta}{\beta})$, and the standard deviation in the threshold voltage, $\sigma(V_T)$ across transistors[5]. Mismatches

4 $\beta = \mu C_{ox} W/L$.

5 Variations in parameters like κ, C_{ox}, W, L are included in these two parameters.

arise because of fabrication process variations (for example, changes in width and length of transistors, oxide thickness, doping) and spatial "white noise" in the device parameters; for example, local fluctuations in electron mobility (Pavasović et al., 1994). Careless layout of a circuit can create unnecessary mismatches. The standard deviation in V_T varies between 2 and 20 mV and the standard deviation in β varies between 2 and 20% (Enz and Vittoz, 1997). Both analysis and measurements from various sizes of transistors show that both $\frac{\delta\beta}{\beta}$ and $\sigma(V_T)$ vary inversely with $\sqrt{WL}$ where W and L are the width and the length of the transistor respectively (Lakshmikumar et al., 1986; Pelgrom et al., 1989). Hence, one way of reducing mismatch is to make the transistor area large. This also applies to other devices like capacitors. The results in Pelgrom et al. (1989) also show that the current mismatch in a current-mirror transistor pair is largest in subthreshold and decreases when the transistors are operated above threshold. This can be understood by computing how much the variation in the threshold voltage δV_T contributes to $\sigma(\frac{\delta I}{I})$. To do this, we multiply δV_T by $\frac{g_m}{I}$. The multiplicand $\frac{g_m}{I}$ is larger in subthreshold than in above threshold. Hence the mismatch is larger for a current-mirror pair in subthreshold.

The current mismatches measured for arrays of different-sized transistors operating in subthreshold (Pavasović, 1991; Pavasović et al., 1994) show that that mismatches are caused by three main factors: edge effects, spatial striations, and random process effects. Edge effects occur because the transistor characteristics depend on the surrounding devices. The variance in the currents among equal size transistors that are biased with the same gate-to-source voltage vary anywhere from 5% to 50% depending on the size, type, and surroundings of the transistor (Pavasović, 1991; Pavasović et al., 1994; Serrano-Gotarredona and Linares-Barranco, 1999).

The mismatch due to spatial variations in the process also apply for other devices, for example, capacitors. Measurements from capacitors of different sizes (McCreary, 1981; Shyu et al., 1982, 1984; Vittoz, 1985; Gregorian and Temes, 1986; Tuinhout et al., 1996; Minch et al., 1996; Minch, 2000a) show that the relative-capacitance distributions of larger size capacitors across the chip is lower than the smaller size capacitors. Generally, capacitors are very well matched compared to transistors or resistors[6]. Designers who think that all devices work alike are shooting themselves in the foot.

6 That's why switched-capacitor circuits rule!!

13 A Millennium Silicon Process Technology

This chapter is a short description of typical commodity silicon process technology as we enter the new millennium. We describe a state of the art process flow for 0.25 μm CMOS[1] by explaining the different process modules. We conclude with a brief discussion of the scaling of process technology to smaller dimensions. Process technology is extremely complex and very highly developed; we can only touch on some important aspects of it. The interested reader may wish to consult the excellent book (Plummer et al., 2000).

13.1 A typical 0.25 μm CMOS Process Flow

The lithography technology used for critical layers (usually active, poly gate, contact and metal1) is deep UV (DUV[2]) optical lithography. All other layers use I-line (365 nm) lithography.

1. Field Oxide (Isolation) Formation The process begins with prepared wafers that must be patterned with the field oxide transistor isolation structures. The starting material is either $p+/p-$ epi[3] or $p-$ bulk[4]. The process starts with the formation of the isolation oxides. Two different isolation approaches are used in 0.25 μm CMOS processes: The first is called *LOCOS*[5] where a nitride[6] layer locally inhibits the growth of a thick isolation oxide layer (Fig. 13.1) (Kooi et al., 1976; Wolf, 1995).

The major drawback of this traditional approach is the oxidant diffusion under the edges of the nitride layer which results in a smaller than expected active width. The difference between the original active width on the mask and the final active width is named "bird's beak". To accommodate this bird's beak, it is necessary to oversize the active region layer by the amount of two times the bird's beak, leading to a loss in real estate. High-pressure

1 0.25 μm processes entered mass production around 1998.

2 DUV is $<$248 nm, and generally uses 193 nm KrF laser light sources.

3 epi=epitaxially grown. *Epi* means that lightly doped silicon of a few microns is grown on a heavily doped substrate. The lightly doped epi allows formation of n and p wells and the heavily doped substrate reduces latchup problems.

4 Bulk material is sometimes used for commodity products because it is cheaper.

5 LOCOS=**LOC**al **O**xidation of **S**ilicon.

6 Nitride=Silicon Nitride=Si_3N_4. Nitride is an oxidation barrier, among other useful characteristics.

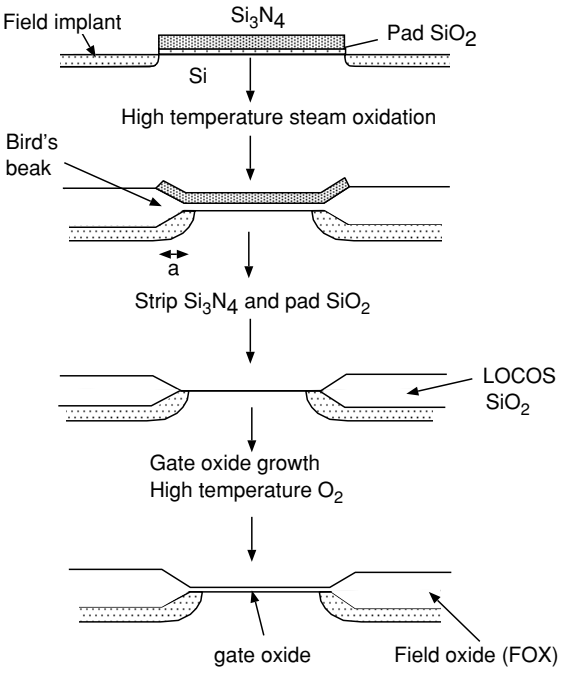


Figure 13.1
Steps in a LOCOS process. A thin layer of sacrificial oxide is grown on top of the Si substrate. Next a layer of silicon nitride is grown on top of the oxide in areas where the field oxide formation is not desired. A shallow field implant is done to inhibit parasitic transistor action between source/drain regions belonging to different transistors which are separated by field oxide. The nFET regions get a p-type field implant, and the pFET regions get a n-type implant. After the oxide has been grown, the silicon nitride and the sacrificial oxide are removed. A layer of clean gate oxide is then grown on top of this area. The field implant diffuses vertically under the field oxide and laterally into the channel. The encroachment of oxide under the edge of the gate is called the bird's beak; together with the lateral diffusion of the field implant, it causes the transistor channel width to effectively be narrower than drawn by a distance a on each side. In most present-day process flows, the field implants are done after field oxide formation, not before, as shown here. Figure adapted from Kooi et al. (1976). Reprinted with permission of The Electrochemical Society, Inc.

oxidation has been shown to reduce the bird's beak, but is not usually used in CMOS processes. At least 100 different complex LOCOS concepts have been developed to minimize the bird's beak: These include semi or fully recessed LOCOS; Poly Buffer LOCOS (PBL) (Han and Ma, 1984); Sealed Interface Local Oxidation (SILO) (Bergemont et al., 1989); and Sidewall Masked Isolation (SWAMI) (Chiu et al., 1982). Details of these different LOCOS concepts can be found in Wolf (1995).

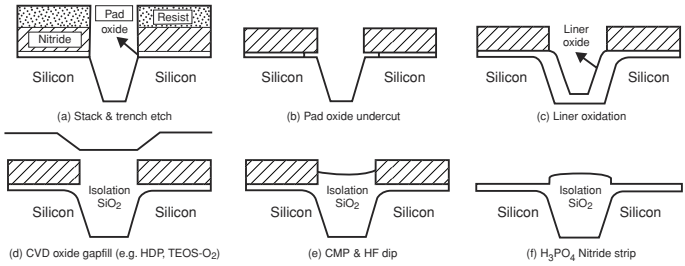


Figure 13.2
Different steps in a Shallow Trench Isolation (STI) process. Figure adapted from Nandakumar et al. (1998), Shallow trench isolation for advanced ULSI CMOS Technologies, International Electron Devices Meeting Technical Digest. © 1998 IEEE.

The second process is called *Shallow Trench Isolation (STI)*, and is the more commonly and recently used approach. In this process, the bird's beak is reduced to near zero and it has the advantage of providing a better planar structure for further processing. The formation of STI (Fig. 13.2) starts with the growth of a thin pedestal oxide, followed by the deposition of a silicon nitride layer (Holloway et al., 1997; Nandakumar et al., 1998; Matsuda et al., 1998; Kuroi et al., 1998). The active mask is used to pattern the STI openings, followed by the etching of nitride/oxide/silicon. The trench etch into silicon is usually around 0.4–0.5 μm. It is important to realize 80 $^\circ$ angle sidewalls

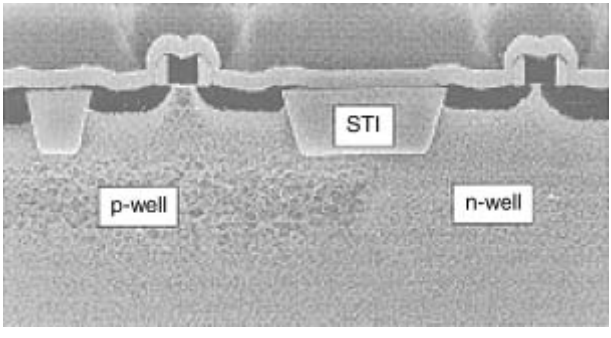


Figure 13.3
Cross section of STI. The transistor and isolation structure are stained to delineate n^+/p^+ junctions and gates. Figure from Holloway et al. (1997), 0.18 μm CMOS Technology for High-Performance, low-power and RF applications, 1997 Symposium on VLSI Technology: Digest of Technical Papers. © 1997 IEEE.

as well as rounded bottom corners to minimize stress induced defects into the silicon. After photoresist strip, a thin thermal oxide is grown on the trench sidewalls to recover any defects and to produce a small round corner at the top of the trench. Then, a dielectric film is deposited to fill the trench. Conformality[7] of deposition is important to avoid the formation of voids inside the trench. High-density plasma (HDP) chemical vapor deposition (CVD) tools are frequently used because they can adequately fill the trench. Then a chemical/mechanical polishing (CMP) step is used to polish back the HDP film until the nitride layer is reached (nitride acts as a CMP stop layer). The dielectric is

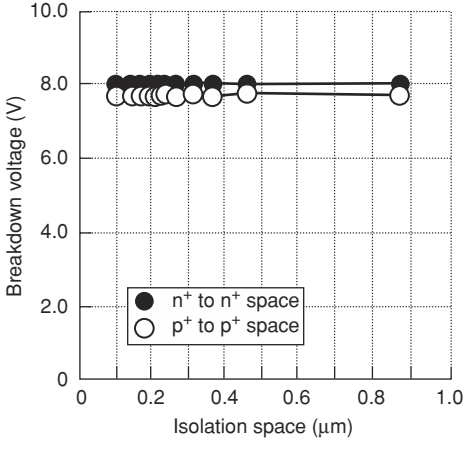


Figure 13.4
Isolation characteristics of STI technology. Figure adapted from Kuroi et al. (1998), Stress analysis of shallow trench isolation for 256M DRAM and beyond, International Electron Devices Meeting Technical Digest. ©1998 IEEE.

then densified at high temperature (1000–1100 °C) to suppress any HDP stress induced defects into the silicon (Kuroi et al., 1998). Then nitride and pedestal oxide are removed, resulting in the final structure shown in Fig. 13.3 (Holloway et al., 1997). Figure 13.4 (Kuroi et al., 1998) shows that STI can maintain good isolation characteristics down to the 0.1 μm range.

7 Conformality means uniformity of thickness of coverage.

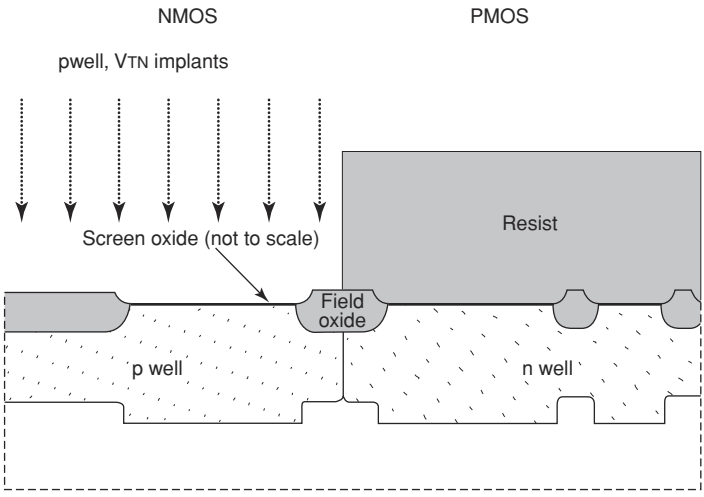


Figure 13.5
Well implantation.

2. Formation of Wells The next step is the implantation of the wells through the field oxide using high energy implants (500 keV for n-Well **Ph**[8] implant and 300 keV for p-Well **B**[9] Implant.). Since the dopants are placed deeply enough by the implant itself, there is no need for high temperature annealing to drive the dopants[10] as in a 0.5 μm CMOS process. Usually the threshold adjust and punchthrough implants[11] are performed at the same time (with the same P- and N- masks). Figure 13.5 shows a process cross section after the wells are formed.

3. Gate oxide formation Figure 13.6 shows the wafer cross section after formation of gate oxide, including the next step of polysilicon deposition and patterning. Gate oxide is continuing to scale down: The typical thickness for 2.5 V, 0.25 μm technology is 45–50 Å. Not only does the oxide need to be reliable from a lifetime point of view, but the oxide also needs to prevent the

8 **Ph** = Phosphorous

9 **B** = Boron

10 Annealing means temperature cycling to heal crystal defects, and driving means heating the wafer to cause rapid diffusion of the dopants.

11 The threshold adjust/punchthrough implants increase both nFET and pFET thresholds to prevent leakage current when the transistor is off.

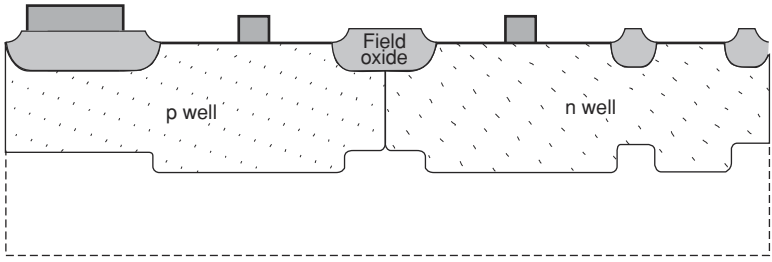


Figure 13.6
Gate oxide and poly formation.

diffusion of dopants (especially **B** from the p+ gate). The addition of nitrogen in the gate oxide has been proven to be a solution for Boron penetration with oxides grown in NO or N_2O ambient.

4. Gate (Polysilicon) Formation Polysilicon deposition follows the gate oxide growth. Figure 13.6 shows a process cross section after gate patterning and plasma etching. The amorphous or polycrystalline polysilicon is undoped, with a typical thickness in the range of 300–400 nm. The doping of the gate is done later, simultaneously with the source/drain implants, allowing surface channel formation for both *n* and *p* channel transistors.

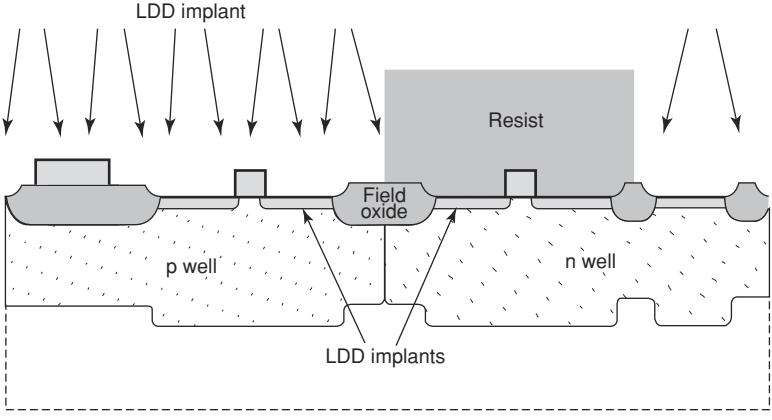
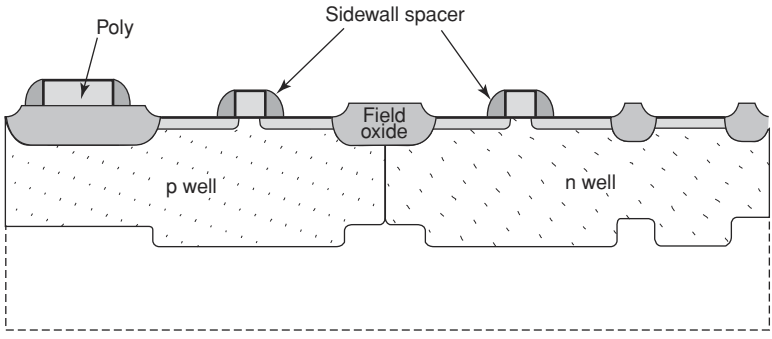


Figure 13.7
Source/drain extension. The first step in full source/drain formation.

5. Source/Drain Extension (SDE) Formation After gate formation, lightly doped source/drain implants (LDD) are realized, self-aligned with the edge



.

Figure 13.8
Lightly-doped drain spacer formation.

of the gate. The primary functions of the SDE are to (1) form a region that reduces the local electric field, to minimize hot carriers effects; (2) realize extremely shallow junctions to minimize short channel effects (Drain-induced

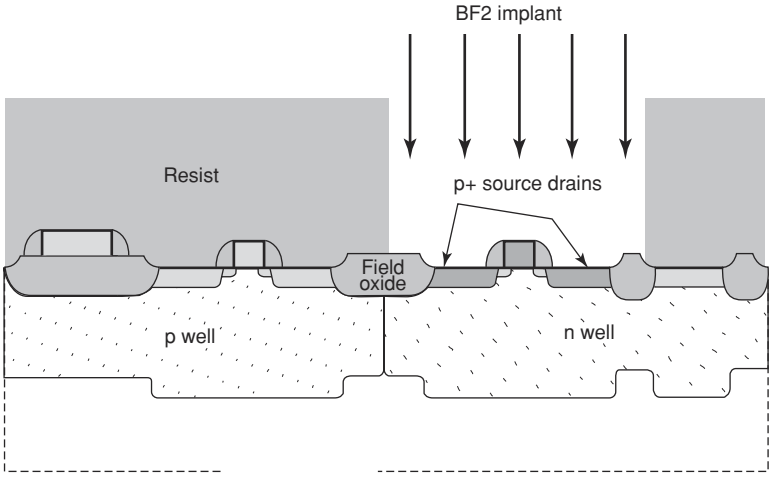


Figure 13.9
Heavily-doped source/drain p+ implant.

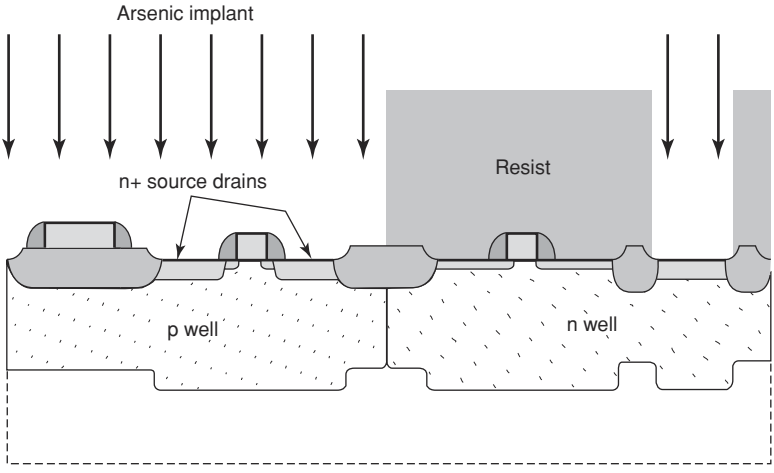


Figure 13.10
Heavily-doped source/drain *n*+ implant.

barrier lowering). Arsenic (**As**) implant is used for NLDD and $\mathbf{BF}_2$ for PLDD regions[12]. Some processes also add halo implants[13] to further reduce the short

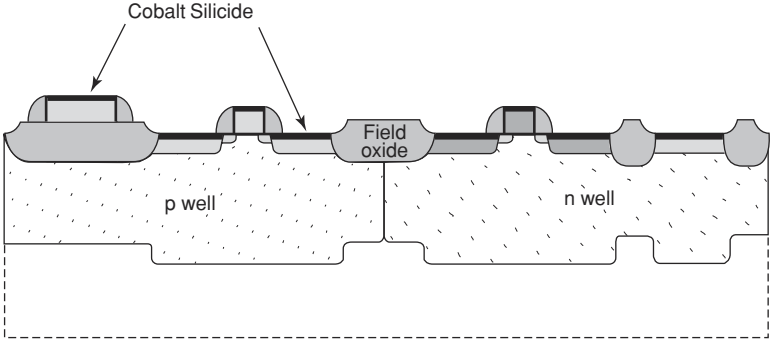


Figure 13.11
Salicide formation.

12 **As** = Arsenic (slow diffusing donor), **B** = Boron (acceptor), **Ph** = Phosphorus (fast diffusing donor).

13 Halo implants are tilted implants to introduce dopants under the gate edges. They help prevent punchthrough for minimum length transistors.

channel effects (**Ph** for pFETs and $\mathbf{BF}_2$ for nFETs). Figure 13.7 shows a cross section after LDD implants.

6. Spacer Formation An oxide or nitride layer is then deposited and anisotropically etched[14] back, to form dielectric spacers (Fig. 13.8). Then the heavily doped source/drain areas are implanted. These regions have deeper junction depths than SDEs; the poly gates are also simultaneously doped with those implants (Figs. 13.9, 13.10). Then a high temperature (1050 °C) rapid thermal anneal is used to activate[15] the dopants.

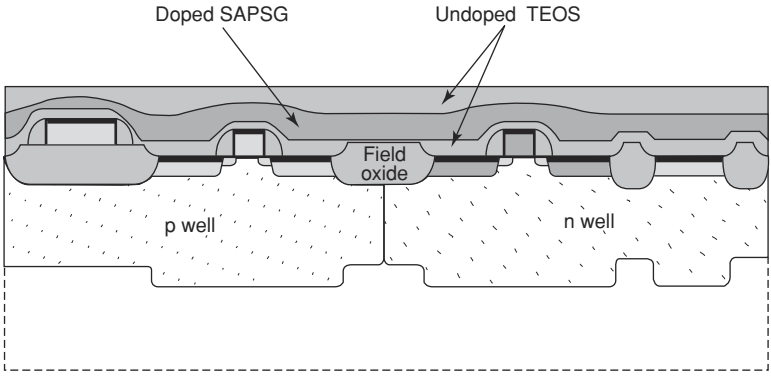


Figure 13.12
Interlevel dielectric deposition and polishing. SAPSG (Sub Atmospheric Phophorous doped Glass) is a silicon phosphate covering material that prevents sodium contamination,TEOS (tetraethooxysilane) is a liquid containing Si and O that is decomposed to form SiO_2, and is undoped to prevent phosphorous contamination.

7. Salicide Formation The next step is the formation of metalized polysilicon called salicide[16], to reduce the sheet resistance of the polysilicon gate and interconnects as well as the source/drain regions[17]. The spacers formed previously on the sides of the poly gates inhibit the formation of the salicide over the spacer region, to avoid source/drain to gate shorts. The sequence to

14 Anisotropic etching etches vertically much more than laterally.

15 Activate means to anneal the silicon structure after implantation to make the dopants part of the single crystal structure.

16 *Sa**licide* means self aligned silicidation; **Si**licide is usually $\mathbf{CoSi_2}$ or $\mathbf{TiSi_2}$

17 Section 12.1 discusses the effect of silicide on photosensitivity.

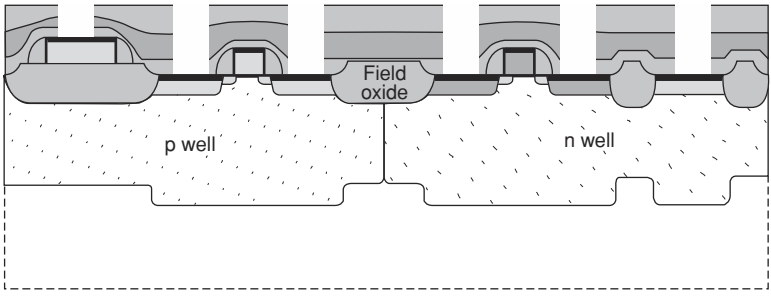


Figure 13.13
Contact openings between active and first metal.

form salicide starts with the removal of any oxide on the source/drain and poly lines. Then the refractory metal[18] (**Ti** or **Co**) is sputtered onto the wafer. A low temperature anneal is next used to react the metal and the silicon exposed on the source/drains and poly lines. The refractory metal that lies on the spacer

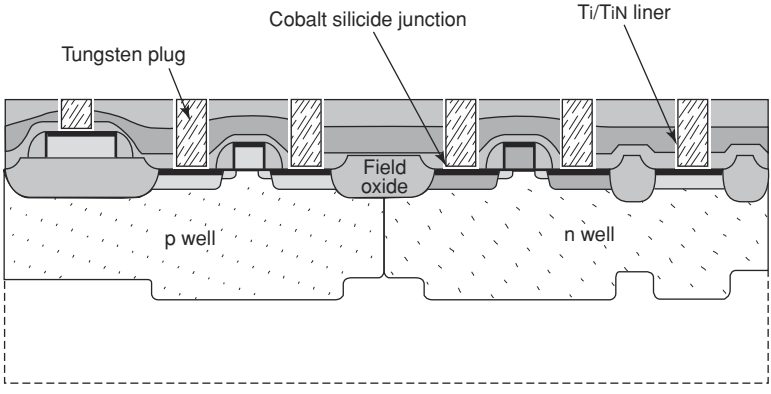


Figure 13.14
Contact filling.

or the field oxide is not converted into salicide during that step, allowing the removal of that unreacted metal in those regions. Then a second anneal is performed to convert the salicide into a more stable and lower resistivity material. Figure 13.11 shows a cross section after salicide formation.

18 Refractory means it has a high melting temperature.

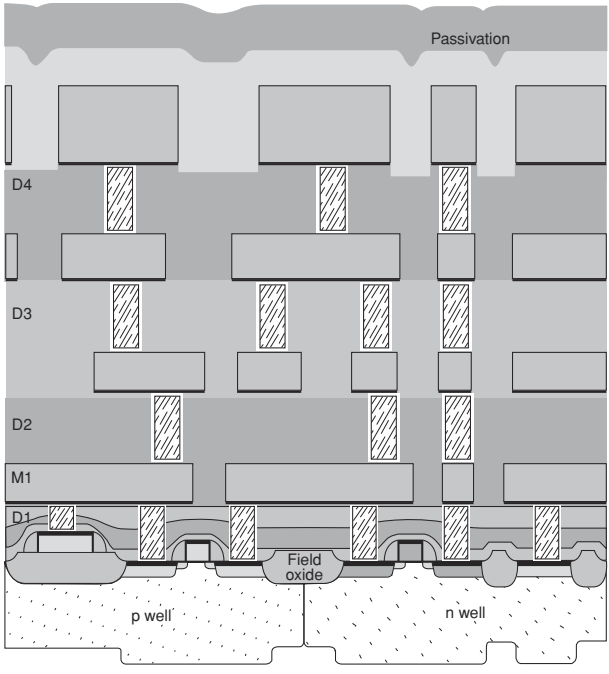


Figure 13.15
Final cross section after passivation.

8. ILD Formation The Interlevel dielectric (ILD) between poly and metal1 is then deposited and polished by CMP[19] to provide a planarized flat surface (Fig. 13.12).

9. Contact Formation A contact mask is used to define openings into the ILD (Fig. 13.13). The contacts connect metal to active or poly areas and also interconnect metal layers. The interconnect itself is generally **Al** (although **Cu** is becoming more popular for high speed logic), but **Al** cannot be directly connected to Si when the junctions are very shallow because spikes of **Al** (wormholes) form, and they short the contacts to the bulk. Hence contact holes are generally filled with another metal, which is then alloyed to the overlying **Al**. Filling the contacts can be tricky. A large aspect ratio[20] is

19 In CMP (chemical mechanical polishing), the wafers are polished by rotating buffing wheels in a chemical slurry.

20 Aspect ratio = interlevel spacing / contact width.

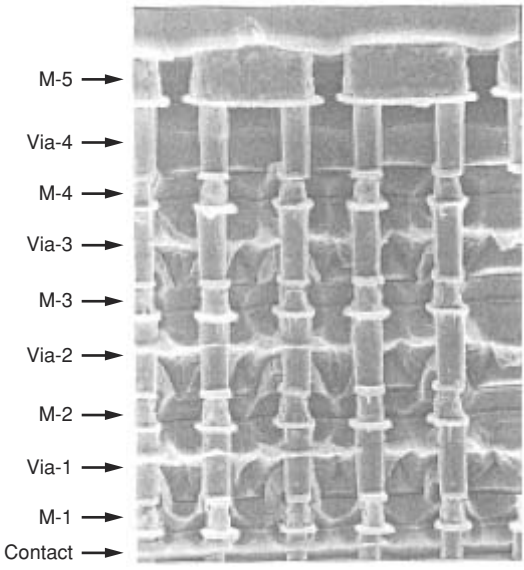


Figure 13.16
Scanning electron microscope image of stacked vias. Figure adapted from Sun et al. (1998), Foundry technology for the next decade, International Electron Devices Meeting Technical Digest. ©1998 IEEE.

desirable because it reduces interlayer capacitance while still maintaining feature density. However, contacts with large aspect ratio are difficult to fill and deep contact cuts are difficult to etch. For these reasons **W**[21] plugs (where **W** is deposited and etched back) or **W** CMP (where **W** is deposited and polished) have been developed (Fig. 13.14). Use of **W** forces the use of barrier metal material[22] (such as **Ti/TiN**) before **W** deposition. The conformality of those barriers inside the contact hole and the method of deposition are other important issues for reliability.

The contact fill scheme is repeated for the filling of vias connecting metal layers. Figure 13.15 shows a final cross section after passivation[23]. Fig-

21 **W**=tungsten.

22 A barrier metal blocks reaction of WF_6 (the gas used to carry **W**) with silicon. This reaction would have bad effect of forming pipes (also named wormholes) into the silicon substrate.

23 Passivation means covering the wafer with a tough, moisture and sodium ion resistant material like phosphorous doped SiO_2. Sodium ions contaminate silicon by making mid-band traps that greatly decrease minority carrier lifetime and make charge stick in the channel. Only the bonding pads are opened up to allow connection.

ure 13.16 shows the use of stacked vias (Sun et al., 1998), which are now commonly available and which greatly simplify logic layout.

13.2 Scaling Limits for Conventional Planar CMOS Architectures

Process scaling is largely driven by the demands for faster and denser logic, not by demands for more compact, lower power, and more precise analog. Analog designers are usually compelled to use the logic scaling results whether they want to or not. It will be clear from the following discussion of scaling how logic scaling is focused on three objectives: Density, speed, and lower power. A more general discussion of process scaling that focuses on long-term trends is given in Chapter 14.

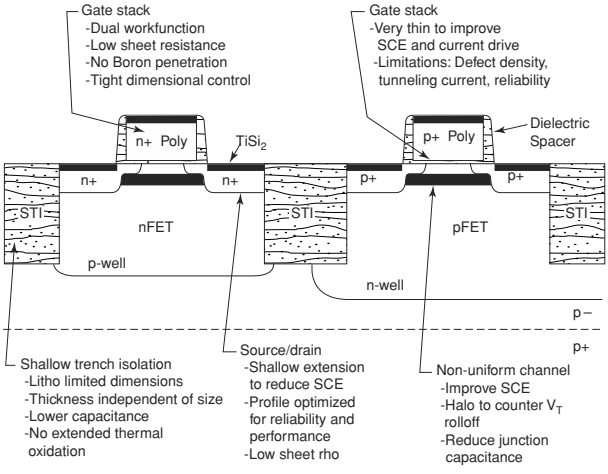
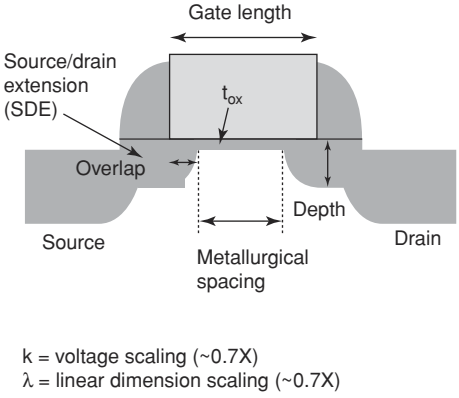


Figure 13.17
State of the art CMOS and some of the features to consider for scaling. Figure adapted from Davari et al. (1995), CMOS scaling for high performance and low power – the next ten years, Proceedings of the IEEE. ©1995 IEEE.

As gate length scales down and below 0.25 μm, short channel effects determine the scaling limits. These limits including finite gate leakage currents in the transistor off state and the increased resistance of sources and drains. All of these limits lead to degraded device performance. Figure 13.17 (Davari et al., 1995) shows a cross-section of a state of the art CMOS process and some of the important features to consider for continual scaling down of

feature sizes. The details of how scaling is limited by the different technology parameters are discussed in Chapter 14. In Fig. 13.18 (Thompson, 1999), we



		λ = k
Dimension(linear)	λ	k
Potential	k	k
Impurity Con.	k/λ^2	1/k
Electric Field	k/λ	1
C_{ox}	λ	k
Current	k^2/λ	k
Power	k^3/λ	k^2
Power * delay	λ^3/k	k

Figure 13.18
The historical scaling scenario. The table on the right shows how different circuit parameters are affected when the voltage supply and all dimensions are scaled down by k and by λ respectively. The right column of the table shows the scenario for constant field scaling. Figure adapted from Thompson (1999), Sub 100 nm CMOS: Technology performances, trends and challenges, International Electron Devices Meeting (IEDM) short course, Washington D.C.©1999 IEEE.

see the historical scaling scenario of transistors. These scenarios have been used over the years as guidelines for each new process generation. The left column of the table shows the changes in circuit parameters as the dimensions are scaled down by λ and the voltage supply by k. For example, if the voltage supply is not scaled down, (k=1), the electric field increases as $1/\lambda$. The right column shows the constant-field scaling scenario where the voltage supply is scaled down by the same factor ($\lambda = k$) as the dimensions.

Threshold Drift

For analog and mixed-signal applications in advanced CMOS technologies, the main limitation for MOSFET scaling is the threshold shift of the pFET due to negative bias temperature instability (NBTI) and hot-carrier injection (HCI). The nitridation of gate oxide which is done to prevent *B* penetration from the pFET gate enhances NBTI. The threshold drift due to NBTI and HCI is shown in Fig. 13.19 in a 0.18 μm process (Chaparala et al., 2000). A surface channel pFET was stressed under NBTI conditions ($V_g = V_{\text{stress}}$ and V_d=V_s=V_{sub}=0)

and HCI conditions ($V_g = V_d = V_{\mathrm{stress}}$ and $V_s = V_{\mathrm{sub}} = 0$). The threshold drift is as much as 80 mV after 500 hours of stress @ 150 °C (Chaparala et al., 2000; Kimizuka et al., 2000). This drift is due to the generation of interface traps and the formation of fixed positive charges in the oxide under stress conditions.

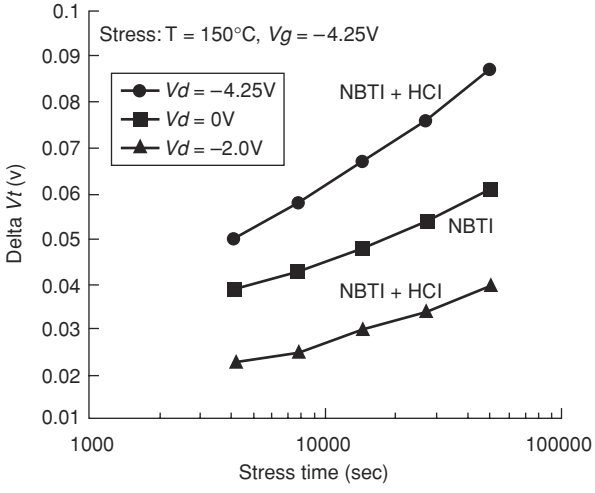


Figure 13.19
Threshold voltage drift over stress time during NBTI and HCI stress conditions. The devices were stressed at 150 °C. Figure adapted from Chaparala et al. (2000), Threshold voltage drift in pFETs due to NBTI and HCI, IEEE International Integrated Reliability Workshop, ©2000 IEEE.

Gate Current Limitations due to Gate Oxide Scaling

As the MOS feature size is scaled down, the gate oxide thickness scales down along with the feature size. However, the oxide thickness can only be scaled down to about 2 nm before the gate leakage current becomes unacceptably large (around 1 A/cm^2 with V_{dd}=1 V) for device performance. This limit comes about not because of the feasibility of manufacturing oxide thickness that is smaller; but rather, it is set by the gate-to-channel tunneling leakage current. The gate leakage current is due to direct tunneling through the oxide. This tunneling occurs from the inversion layers and the accumulation layers (SDE to gate overlap). The increase in gate current density as a function of the gate voltage for different oxide thickness can be seen in Fig. 13.20 (Lo et al., 1997; Thompson, 1999). The oxide thickness limit is considered to be reached when

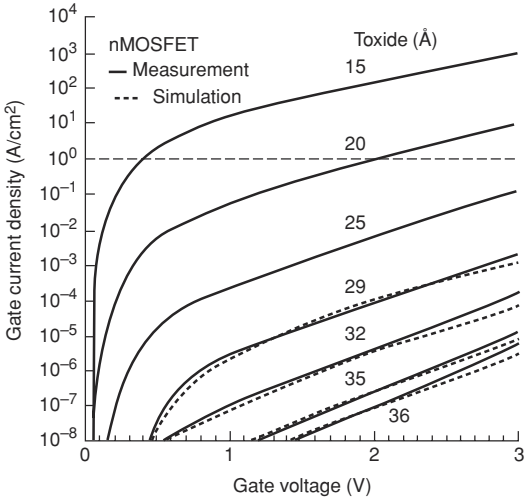


Figure 13.20
Simulated gate current density versus gate voltage from a nFET in inversion. Figure adapted from Lo et al. (1997), Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's, IEEE Electron Device Letters, ©1997 IEEE.

the tunneling current is equal to the off-state drain-to-source subthreshold leakage (around 1 nA/μm). This limit on the gate oxide thickness sets the limit

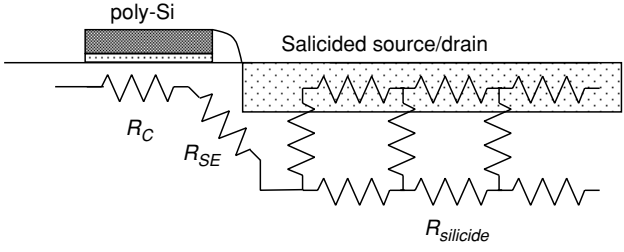


Figure 13.21
Components of series resistance at the source/drain of a MOS transistor.

on the channel length of the transistor, because making the channel shorter without decreasing the gate oxide thickness eventually results in a transistor whose current is controlled more by the source and drain nodes than by the gate.

Alternatives High dielectric constant materials like Ta_2O_5 are currently being considered as a method of reducing the tunneling current through the insulator. Such materials permit thicker dielectrics to be used for the same inversion charge. These materials unfortunately have their own problems. They need an SiO_2 buffer between the dielectric and the substrate. They also need a metal gate (**TiN/W** or **TiN/A**l) to prevent a reaction between the poly Si gate and the dielectric.

Source/Drain Junction Scaling

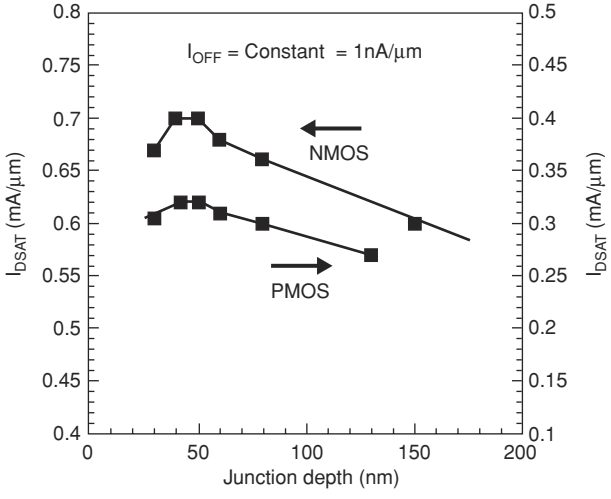


Figure 13.22
I_{DSAT} versus SDE depth for an nFET and pFET. The offset spacer is at 0 nm. Figure adapted from Thompson et al. (1998), Source/drain extension scaling for 0.1 μm and below channel length MOSFETs, Symposium on VLSI Technology: digest of technical papers, ©1998 IEEE.

As gate lengths scale down, the SDE depth and gate overlap are also scaled down. The SDE depth is scaled down so that the drain charge does not significantly control the amount of channel charge. For 0.25 μm process technologies, junction depths are around 50–100 nm. However, these shallow SDEs gave rise to smaller I_{DSAT}, as was predicted. The shallow SDE has a side effect in that it leads to a larger series resistance. For a salicided LDD MOSFET, the series resistance of the source and drain (R_s and R_d respectively) is the sum of various components as shown in Fig. 13.21 (Asai and Wada, 1997; Thompson,

1999); R_s, $R_d = R_{SE} + R_{silicide}$, where R_{SE} is the sheet resistance due the shallow SDE. The R_{SE} can be of the same order of magnitude as the channel resistance R_c especially in a pFET. Figure 13.22 (Taur et al., 1997; Thompson

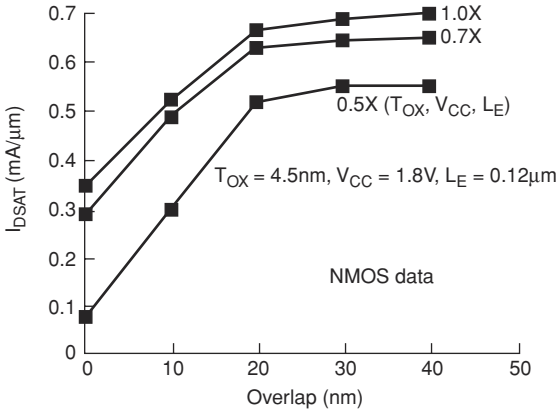


Figure 13.23
I_{DSAT} versus SDE to gate overlap for an NMOS transistor. Figure adapted from Thompson et al. (1998), Source/drain extension scaling for 0.1 μm and below channel length MOSFETs, Symposium on VLSI Technology: digest of technical papers, ©1998 IEEE.

et al., 1998) shows that the maximum I_{DSAT} for both nFETs and pFETs at a constant off-state leakage (1 nA/μm) is degraded when the junction depth is below 35–40 nm because of the increased SDE resistance. On the other hand, I_{DSAT} degrades when the junction depth increases above 35–40 nm because of charge sharing.

The SDE to gate overlap also affects the saturation current through the transistor. If the overlap is reduced, the current spreads out into a lower doped region of the SDE. This increases the series resistance and also the accumulation. Figure 13.23 (Taur et al., 1997; Thompson et al., 1998) shows how I_{DSAT} degrades when the SDE overlap is less than 20 nm for a transistor in a 0.25 μm technology and for transistors whose gate oxide, power supply, and gate length are scaled down by both 0.7 and $(0.7)^2$.

Transistor Off-State Leakage Currents and V_{dd}/V_T Scaling.

The dependence of V_T on the gate length is stronger compared to other factors that also cause V_T fluctuation like random dopant distribution. As gate lengths

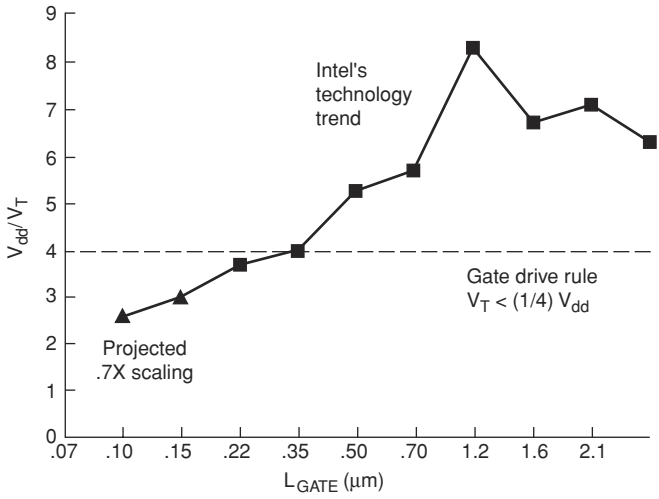


Figure 13.24
V_{dd}/V_T trend as devices scaled down in Intel's process technology. Figure adapted from Thompson (1999), Sub 100 nm CMOS: Technology performances, trends and challenges, International Electron Devices Meeting (IEDM) short course, Washington D.C. ©1999 IEEE.

scale below 0.1 μm, both the power supply and the threshold voltage are also scaled down. Reducing the power supply is necessary to decrease the power consumption. However, this leads to reduced gate overdrive for the same V_T. Hence, threshold voltages also have to scale down. A side effect of this is that the off-state leakage increases due to the increased subthreshold leakage currents. A general rule of thumb is to have V_{dd} about $4V_T$, to maintain gate overdrive at the same off-state leakage. Figure 13.24 (Thompson, 1999) shows the gate overdrive problems that will be encountered at 0.1 μm by using the historical transistor scaling scenario. At V_{dd}=1 V, a threshold voltage of 0.25 V will be required to maintain good performance! Another important factor is the limitation due to larger V_T variations especially at low V_T. This limitation is illustrated in Fig. 13.25 (Sun and Tsui, 1994), which shows a ring oscillator propagation delay varying wildly when V_{dd} is scaled down towards V_T.

Poly-Gate Depletion

The polysilicon gate can be depleted by the voltage applied to it. The effect of this depletion is to effectively decrease C_{ox}, lowering the maximum current. This effect was not a concern in the past because the drain and source junctions

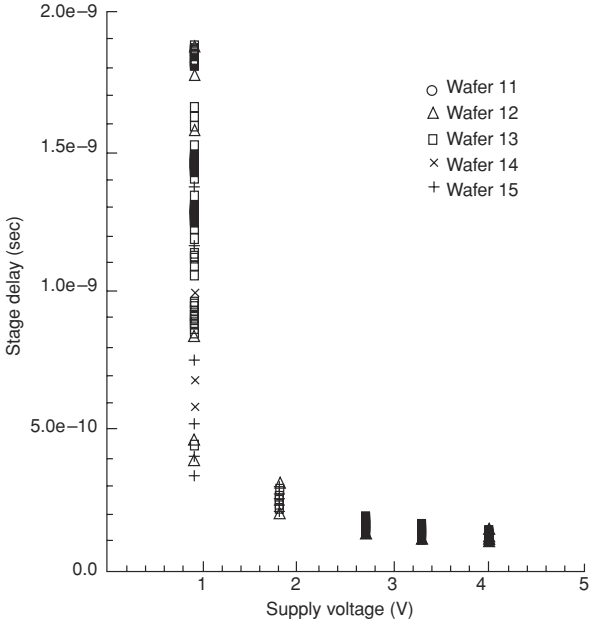


Figure 13.25
Ring oscillator frequency variation with supply voltage scaling. Figure adapted from Sun and Tsui (1994), Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation, IEEE Custom Integrated Circuits Conference, ©1994 IEEE.

and the polysilicon gate were doped separately, so the gate could be doped heavily enough to prevent its own depletion. Because the source/drain and the gate poly are now implanted at the same time, there is a compromise between poly depletion (tied to poly thickness), depth of the junctions (tied to the final thermal processing used to activate the dopants) and poly thickness (to avoid B penetration through the gate oxide). A heavy gate implant and high temperature is needed to reduce poly-gate depletion, but the B penetration through the gate oxide is enhanced. With the extremely thin active junctions now common and the requirement for very small thermal cycling budgets, the gate can be insufficiently doped to prevent its own depletion. Figure 13.26 shows how gate doping concentration affects transistor operation (Arora et al., 1995; Ricco et al., 1996).

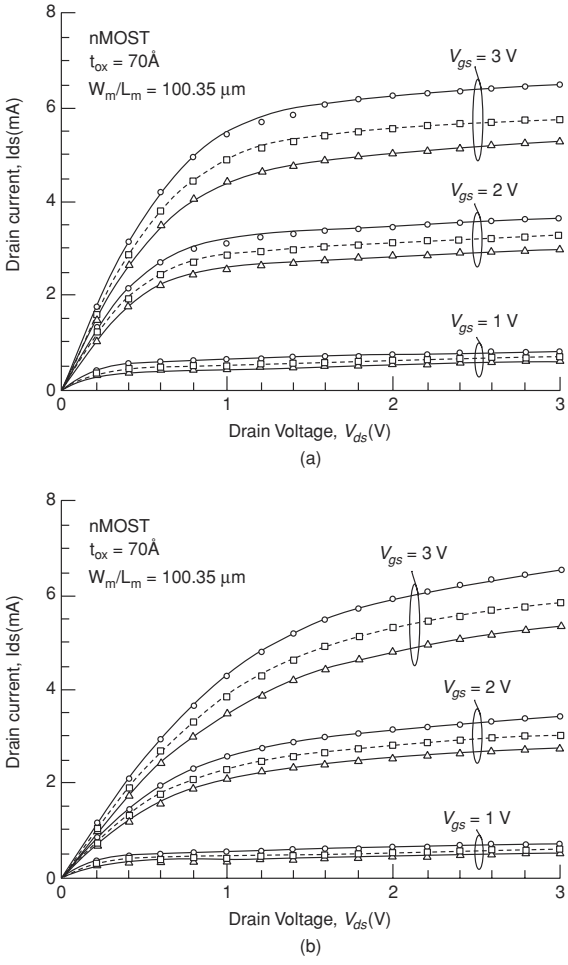


Figure 13.26
Poly-gate depletion. The drain current versus drain voltage for 3 different polysilicon gate doping concentrations (N_α=5 $\times$ 10^{19}, 1 $\times 10^{19}$ and 0.5 $\times 10^{19}$ cm^{-3}) and three gate to source voltages. (a) for nMOST (nFET). (b) for pMOST (pFET). The lines are fits from a model and the symbols are 2-D simulated data. Figure adapted from Arora et al. (1995), Modeling the polysilicon depletion effect and its impact on submicrometer CMOS circuit performance, IEEE Transactions on Electron Devices, ©1995 IEEE.

A Steep Subthreshold Slope

A steep subthreshold slope (which is good for analog circuits because it makes for large g_m in subthreshold) will be difficult to maintain with further process

scaling, unless the starting material is switched from bulk Si to silicon on insulator (SOI). The ever-increasing bulk doping means that κ will tend to decrease.

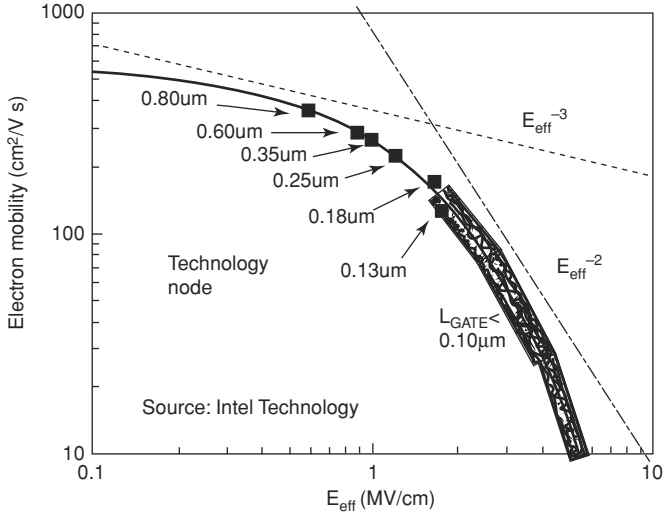


Figure 13.27
Mobility degradation with channel length scaling. Figure adapted from Thompson (1999), Sub 100 nm CMOS: Technology performances, trends and challenges, International Electron Devices Meeting (IEDM) short course, Washington D.C. © 1999 IEEE.

Mobility Degradation

The electron mobility degrades with channel length scaling as shown in Fig. 13.27 (Thompson, 1999). This degradation is the result of impurity scattering due to the extremely highly doped channel region. The degradation can be reduced by growing undoped thin epitaxial layers on top of the highly doped substrate.

13.3 Conclusions and Guidelines for New Generations

New devices and process technologies are constantly being developed, some radically different than the conventional (year 2000) flow described here. When the time is right for new concepts and architectures, everything is

possible. A good cook is innovative by definition! For a new concept to be successful, it has to be

1. Done yesterday or face the ”I need a job!” dilemma (Fig. 13.28).
2. Cheap.
3. Done right and easy to manufacture: Tastes good first time, every time!
4. Done by a good cook (kitchen engineer): Learn how to be one of those, start with the taste of freedom!

You’re gonna ask: what else?! This is the art of cooking. Bon appetit!

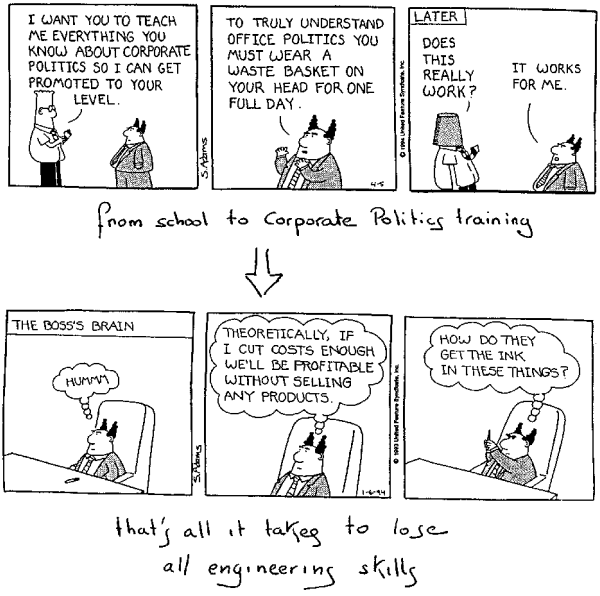


Figure 13.28
Staying sane Adams (1996b,a, 1998b,a). Specified text excerpts from pages 12 and 64 from THE DILBERT PRINCIPLE by SCOTT ADAMS. Copyright (c) 1996 by United Features Syndicate, Inc. Reprinted by permission of HarperCollins Publishers, Inc.

This page intentionally left blank

14 Scaling of MOS Technology to Submicrometer Feature Sizes

It is always difficult to predict the future; few attempts to do so have met with resounding success. One remarkable example of successful prediction is the exponential increase in complexity of integrated circuits, first noted by Gordon E. Moore[1]. As we contemplate the ongoing evolution of this great technology, many questions arise: Can the trend continue? Will single-chip systems attain levels of complexity that render present system architectures unworkable (Mead and Conway, 1980)? Will digital techniques completely replace analog methods (Mead, 1989)? The answers to these questions depend critically on the properties of the individual transistors that provide the essential active functions, without which no interesting system behavior is possible. Integrated-circuit density is increased by a reduction in the size of elementary features of the underlying structures. Therefore, any discussion of the capabilities of future technologies must rely on an understanding of how the properties of transistors evolve as the transistors' dimensions are made smaller[2]

Elsewhere (Hoeneisen and Mead, 1972b), we described the factors that limit how small an MOS transistor can be and still operate properly. That discussion will not be repeated here, but we will outline the major issues:

1. For the device current to be primarily controlled by the gate, the device should not be punched through; that is, the sum of the source and drain depletion layers should be less than the geometric channel length. As a direct consequence of this requirement, the bulk doping must increase as dimensions are decreased.

2. Increasing the bulk doping has two important consequences: (a) Junction breakdown voltage is lowered; and (b) a larger electric field is required in the gate oxide to obtain a given change in surface potential.

Because of 2(a), the operating voltage must be reduced. So that sufficient electric field can be obtained with a lower operating voltage, the gate oxide must be made thinner. Thus, it is inevitable that, as the minification process is continued, both drain depletion layer and gate oxide will become thin enough that electron tunneling through them will become comparable with

1 According to Moore's Law, the number of transistors on a chip doubles every 18 months.

2 This chapter is slightly modified from a paper by Mead (1994), Scaling of MOS technology to submicrometer feature sizes, Journal of VLSI Signal Processing. Reprinted with permission from Kluwer Academic/Plenum Publishers.

other device currents. In 1971, when our original study (Hoeneisen and Mead, 1972b) was written, we described a device of 0.15 micrometers (μm) channel length, having a 50 Angstrom (Å) gate oxide. Although we were confident that a device of this size could be made to work, we were not at all sure that smaller devices could be made viable.

Over the ensuing 22 years, feature sizes have evolved from 6 μm to 0.6 μm, and the trend shows no sign of abating (Nagata, 1992; Davari et al., 1992; Chang et al., 1992; Bryant et al., 1992; Yan et al., 1992; Yamaguchi et al., 1993; Iwase et al., 1993). In this chapter, we shall examine what we have learned from the past 22 years of technology evolution, and shall discuss to what extent these same trends may continue into the future. We conclude that at least one more order of magnitude of scaling can be obtained with a concomitant increase in both density and performance. Several of the conclusions of this study were reached independently by Hu (1993).

14.1 Scaling Approach

The historic trend of gate-oxide thickness t_{ox} as a function of l, the minimum feature size of the process, is plotted in Fig. 14.1. The trend can be expressed accurately as

$$t_{ox} = 210\, l^{0.77},$$

where the feature size is in μm, and the gate-oxide thickness is in Å. This observation suggests that it may be fruitful to express all important process parameters as powers of the feature size, and to determine whether there is a scaling of this form that allows sensible process evolution to dimensions well below 0.1 μm. To prevent the gate-oxide thickness from becoming thinner than a single atomic layer, we have chosen a scaling of the form

$$t_{ox} = max\left(210\, l^{0.77}, 140\, l^{0.55}\right). \tag{14.1.1}$$

This expression is plotted as the solid line in Fig. 14.1. In reviewing the historic trend, it is clear that we previously expressed (Hoeneisen and Mead, 1972b) more concern with gate-oxide tunneling than has been justified by the experience accumulated through the intervening years. It is conceivable that the same bit of paranoia occurs here. In any case, if oxide thickness continues to decrease at the present rate, the resulting devices will be somewhat more capable than those we present.

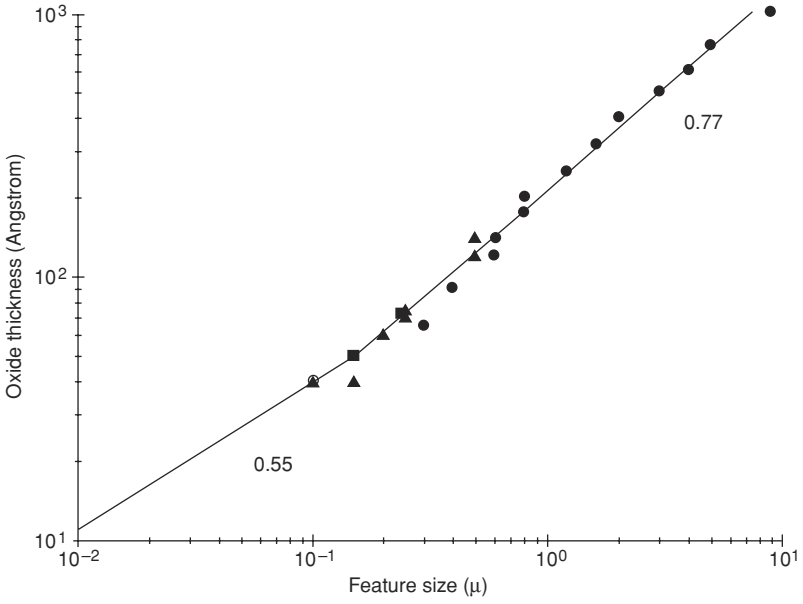


Figure 14.1
Gate-oxide thickness as a function of feature size. The solid circles are production processes in silicon-gate technology, starting in 1970. Triangles are processes reported in the literature. Solid squares are the two most advanced devices described in our previous study (Hoeneisen and Mead, 1972b). The solid line is the analytic expression used in this chapter (Eq. 14.1.1).

The oxide thickness and feature size together determine the gate-oxide capacitance C_g of a minimum-sized device:

$$C_g = \epsilon_{ox} \frac{l^2}{t_{ox}}.$$

The historic trend in supply voltage V is shown in Fig. 14.2. This trend is not as smooth as the trend in oxide thickness, due to the long period of standardization at 5 volts (V). It is clear, however, that modern submicrometer devices operate better on lower voltages (Bryant et al., 1992; Lyon, 1993), and that this trend to lower voltages must continue. The scaling used here is

$$V = 5\, l^{0.75}. \tag{14.1.2}$$

This expression is plotted as the solid line in Fig. 14.2.

Once the gate-oxide capacitance and supply voltage are known, the energy W_g stored on the gate of a minimum-sized transistor at any given feature size

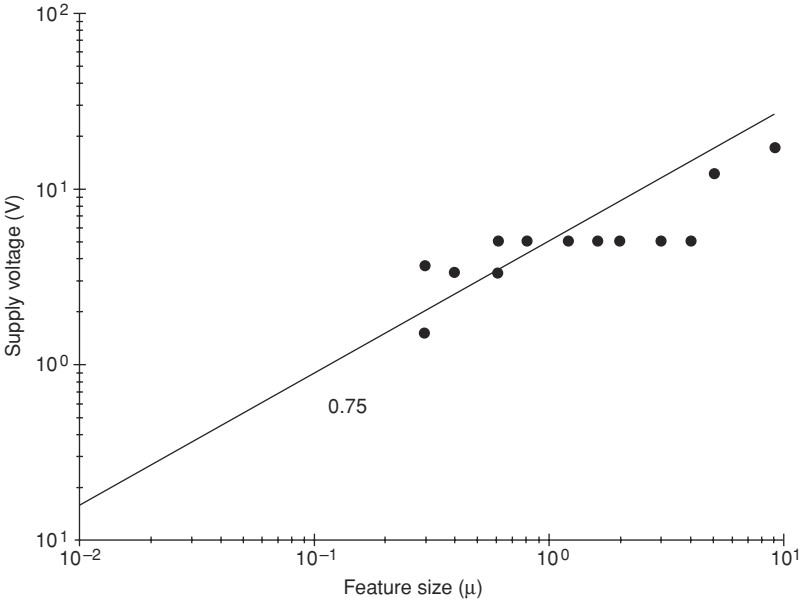


Figure 14.2
Power-supply voltage as a function of feature size. The solid line is the analytic expression used in this chapter (Eq. 14.1.2).

can be estimated. The stored energy is slightly overestimated as

$$W_g = \frac{1}{2} C_g V^2. \tag{14.1.3}$$

For the scaling laws given here, the stored energy (in Joules) is

$$W_g = 2.2 \times 10^{-14}\, l^{2.75}. \tag{14.1.4}$$

This expression is plotted as the long solid line in Fig. 14.3. Even with the slight "kink" introduced by Eq. 14.1.1, this expression is a good abstraction of the actual energy over the entire range of the plot. In the central section of historic data, however, the constant 5 V power-supply voltage has established a trend with much less dependence on feature size.

This shorter trend is well-represented by the expression

$$W_5 = 2 \times 10^{-14}\, l^{1.22}.$$

Also shown for reference on Fig. 14.3 is the thermal energy kT, and the

spacing of levels in the channel with momenta in the direction of current flow. It is clear that the stored energy is more than 10 kT even at feature sizes of 0.01 μm.

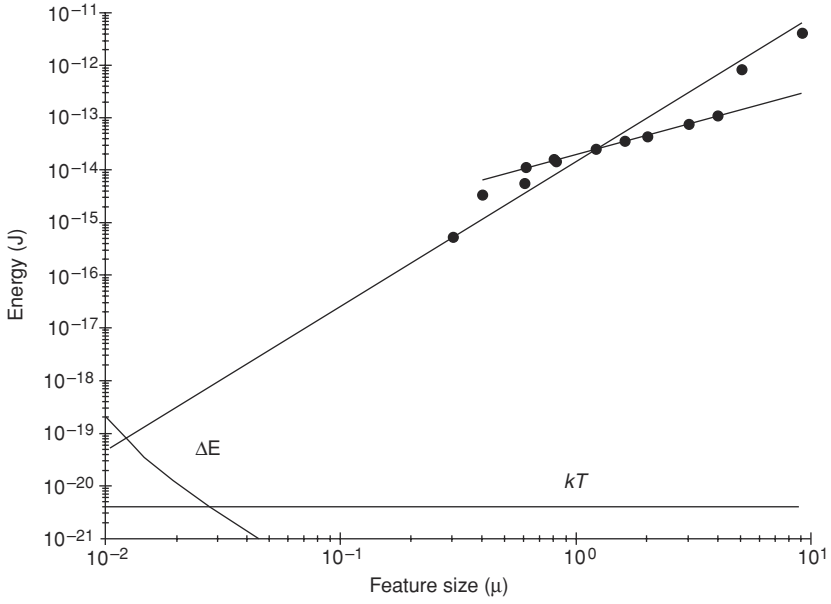


Figure 14.3
Energy stored on the gate of a minimum-sized transistor as a function of feature size. We compute the points from Eq. 14.1.3 using oxide thickness values from Fig. 14.1 and the supply-voltage values from Fig. 14.2. The solid line is the analytic expression used in this chapter (Eq. 14.1.4). Also shown for reference are the thermal energy kT at room temperature, and the quantum-level spacing for electrons in the channel with momenta in the direction of current flow.

The minimum stored energy is an interesting quantity because it sets the scale for the switching energy dissipated in a digital system. The energy per operation of compute-intensive digital chips is compared with the minimum stored energy in Fig. 14.4. The system energy per operation is 4 to 6 orders of magnitude higher than the minimum stored energy, and can be bounded by the two solid trend lines

$$W_{max} = 1.15 \times 10^{-8}\, l^{3.4} \tag{14.1.5}$$

$$W_{min} = 2.5 \times 10^{-10}\, l^{3.25}. \tag{14.1.6}$$

The overall system trend is steeper than that for minimum stored energy, presumably because designers have become more skilled over the years, and processes have an ever increasing set of features on which designers can draw (multiple levels of metal, for example). A 5 V sub-trend is clearly discernible in the system data as well.

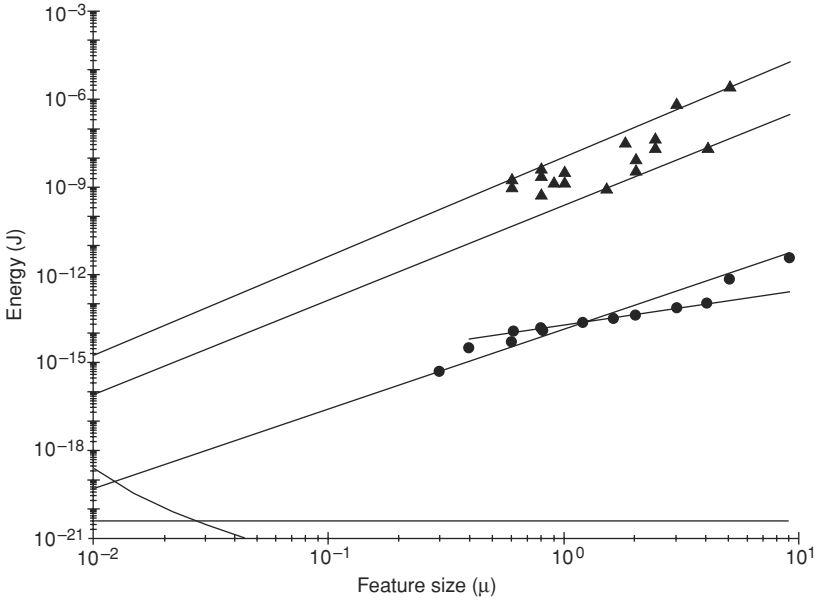


Figure 14.4
Energy dissipated per operation at the chip level. Filled triangles are data taken from the literature and from manufacturers' data sheets. Examples are all compute-intensive single chips, such as multipliers, digital signal processors, and similar devices. So that the data could be plotted on a single scale, all values were normalized to 8×8 multiply-add operations, assuming that the energy is proportional to the product of the word lengths of the multiplicand and multiplier. Minimum and maximum trend lines shown are Eqs. 14.1.5 and 14.1.6. Also shown for reference are the data of Fig. 14.3.

With the information on hand, we can determine the tunneling current density J_{ox} through the gate oxide (Lenzlinger and Snow, 1969; Hori et al., 1992; Suné et al., 1992), making the worst-case assumption that the entire supply voltage appears across the entire gate area:

$$J_{ox} = J_0 E_{ox}^2 e^{-kt_{ox}} \tag{14.1.7}$$

where $J_0 = 6.5 \times 10^{10}\mathrm{A}/(\mathrm{V}/\mathrm{cm})^2$ was adjusted to match experimental data, as shown in Fig. 14.5. The imaginary part of the wave vector k is given by

$$k = \frac{2k_0}{3}\frac{\phi}{V}\left[1 - \left(1 - min\left(1, \frac{V}{\phi}\right)\right)^{3/2}\right].$$

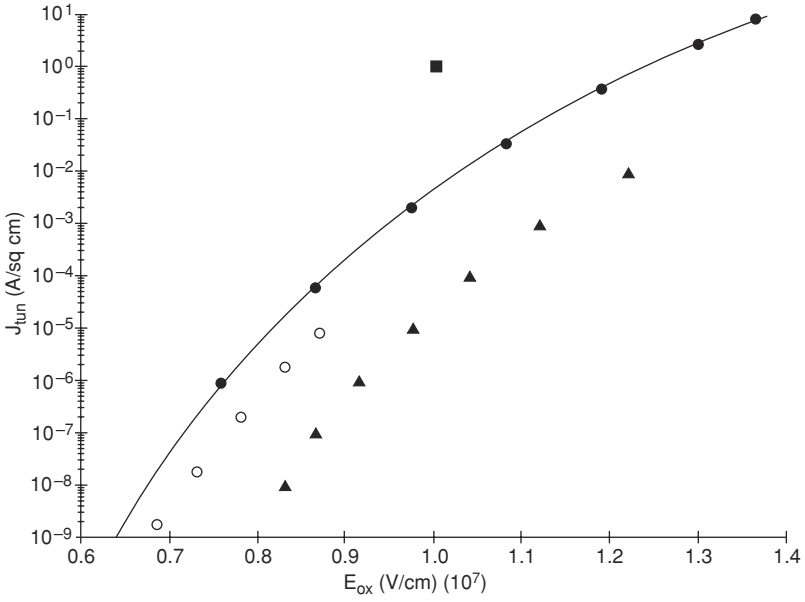


Figure 14.5
Oxide tunneling current as a function of electric field. The open circles are from the original work of Lenzlinger and Snow (1969). Filled circles are from the recent work of Suné et al. (1992). Filled triangles are from Hori et al. (1992). The solid line is the analytical expression used in this chapter (Eq. 14.1.7). The filled square is inferred from Iwase et al. (1993), but is not directly comparable with the other data because it was taken from a transistor drain characteristic, and may be corrupted with other effects such as gate-enhanced drain tunneling. The gate current was not reported separately, so this value shown represents a worst-case estimate.

These expressions are valid for voltages both above and below the barrier potential ϕ, which was taken to be 3.2 V. The pre-exponential constant $k_0 = 1.2\ \text{Å}^{-1}$ was used. It is comforting to note that oxide tunneling data are available over the entire range of electric fields that will be encountered down to the smallest dimensions studied here. It will be helpful, however, to have

actual experimental data in the 10 Å range. For these extremely thin oxides, it will be essential to take into account the quantum corrections discussed in Suné et al. (1992).

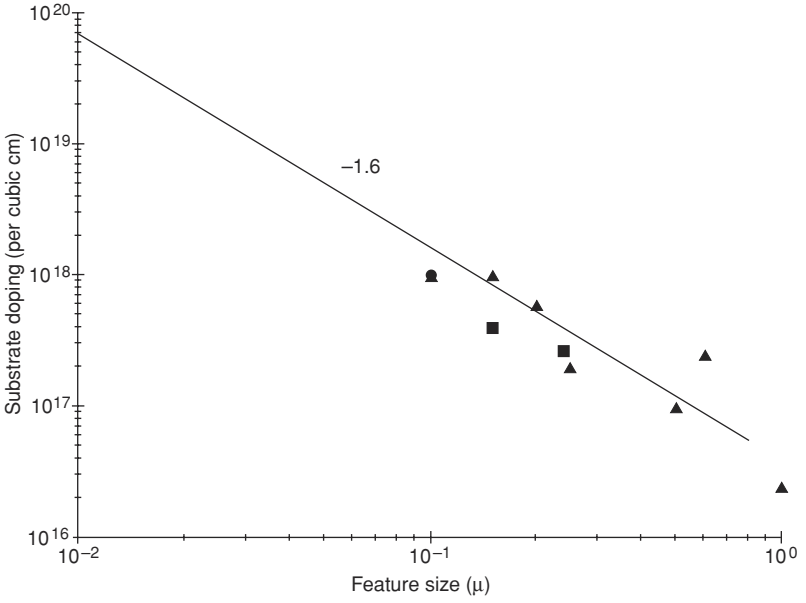


Figure 14.6
Substrate doping as a function of feature size. The solid line is the analytical expression used in this chapter (Eq. 14.1.8). Filled triangles represent processes reported in the literature. The two solid squares are the two smallest transistor designs shown our earlier work (Hoeneisen and Mead, 1972b).

The other major source of parasitic current is tunneling through the drain junction. The junction-tunneling current density J_j is critically dependent on the substrate acceptor concentration n, which must be increased to avoid punch-through as device dimensions are decreased (Chynoweth et al., 1960; Logan and Chynoweth, 1963; Krieger, 1966; Fair and Wivell, 1976; Stork and Isaac, 1983; Hackbarth and Tang, 1988; Reisch, 1990). The scaling law used here is plotted in Fig. 14.6:

$$n = 4 \times 10^{16}\, l^{-1.6}. \tag{14.1.8}$$

Given the doping density n, we can compute the depletion-layer thickness x

for any potential ψ relative to substrate using the usual step-junction approximation

$$x = \sqrt{\frac{2\epsilon_{si}\psi}{qn}}. \tag{14.1.9}$$

The corresponding depletion-layer capacitance C is given by

$$C = \frac{\epsilon_{si}}{x}.$$

We can determine the maximum electric field in the drain junction, from the junction voltage, which in the worst case will be the supply voltage plus the built-in voltage:

$$E_j = \sqrt{\frac{2qn(V + V_b)}{\epsilon_{si}}}.$$

We could alternatively use a graded-junction approximation, such as that used by Fair and Wivell (1976). For our purposes, the two approaches are nearly equivalent, so the simpler step-junction expression with the junction built-in voltage V_b =1.1 V is used. In either case, the tunneling current density is a function of the maximum electric field:

$$J_j = G_0 V \frac{E_j}{E_0} e^{-E_0/E_j}. \tag{14.1.10}$$

The constant $E_0 = 2.9 \times 10^7\ \mathrm{V/cm}$ was taken from Fair and Wivell (1976), and the pre-exponential factor $G_0 = 3 \times 10^9\ \mathrm{A/Vcm^2}$ was chosen to fit the experimental data plotted in Fig. 14.7. It is significant that experimental data exist that allow us to predict the tunneling currents in junctions of devices down to 0.03 μm feature sizes. Previously (Hoeneisen and Mead, 1972b), we pointed out that the "drain corner" tunneling occurs at lower voltage than that across the junction area, a fact that has received considerable attention (Li et al., 1988). For the present study, we use Eq. 14.1.10 for area tunneling, both for simplicity and because considerable cleverness on the part of process designers as this phenomenon becomes limiting can be expected. Caution, however, that corner effects may significantly increase the drain tunneling over the values shown in the following figures.

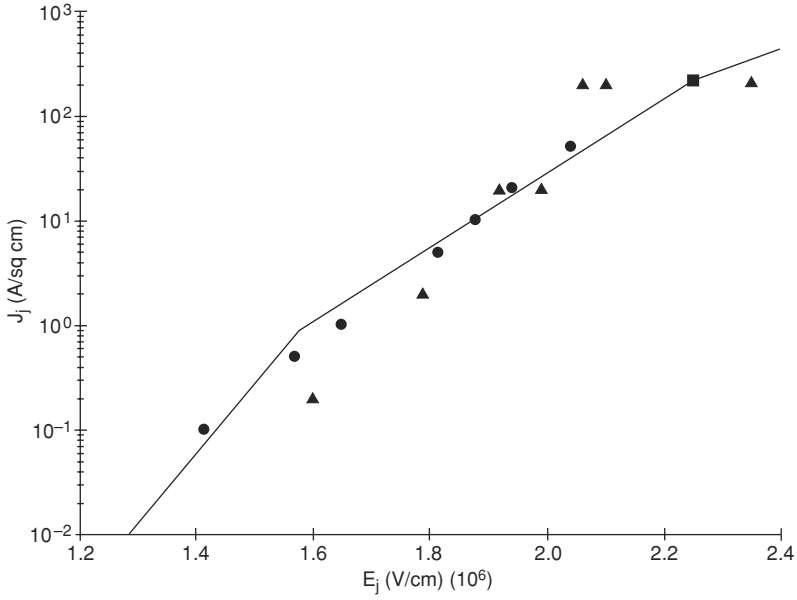


Figure 14.7
Junction-tunneling current density as a function of peak electric field in the junction. The filled triangles are from alloy tunnel diodes, which were reported as step junctions by Chynoweth et al. (1960). The filled circles are from diffused emitter-base junctions reported as graded junctions by Fair and Wivell (1976). These were the only references that we were able to locate for electric fields in the range encountered in the finest feature sizes considered in this chapter. Some data are shown by Reisch (1990), but not enough information is given to allow direct comparison with the other data. For reference, the solid square represents the parameters encountered in the 0.03 μm device described in this chapter. The solid line is the analytical expression used here (Eq. 14.1.10).

14.2 Threshold Scaling

To determine the detailed properties of small devices, we must take into account the short-channel properties, most notable of which are carrier-velocity saturation and drain-induced barrier lowering (the precursor to punch-through). Previously (Mead, 1989), we developed a model that gives closed-form expressions for the current in short-channel devices, including the effects of velocity saturation . To apply the model, we need some abstraction of the vertical doping profile under the gate. The most widely used such abstraction is the threshold voltage V_T. We therefore proceed by choosing a nominal threshold voltage of the form

$$V_T = 0.55\, l^{0.23}. \tag{14.2.1}$$

The actual threshold voltage will be lower than the nominal one by the amount of drain-induced barrier lowering (DIBL) (Troutman, 1979; Bakker, 1991; Deen and Yan, 1992; Van der Tol and Chamberlain, 1993). The expression given by Fjeldly and Shur (1993) is

$$\text{DIBL} = V\,\frac{x_c}{\lambda}\,\frac{\sinh\left(\dfrac{x_s}{\lambda}\right)}{\cosh\left(\dfrac{l-x_d}{\lambda}\right)-\cosh\left(\dfrac{x_s}{\lambda}\right)} \tag{14.2.2}$$

where x_s and x_d are the classical depletion-layer thicknesses of the source and drain junctions. We have used a surface potential of 0.5 V in Eq. 14.1.9 to compute x_c, the thickness of the depletion layer under the channel. The distance scale λ is given by

$$\lambda = x_c\left(1+\frac{C_{ox}}{C_c}\right)^{-\frac{1}{2}}$$

where the depletion-layer capacitance per unit area C_c from channel to substrate is

$$C_c = \frac{\epsilon_{si}}{x_c}$$

and the oxide capacitance per unit area C_{ox} from gate to channel is

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}.$$

The nominal threshold voltage; the actual threshold voltage, including DIBL; and the supply voltage are plotted as a function of feature size in Fig. 14.8. For the scaling parameters used here, DIBL does not become a serious problem until feature sizes are less than 0.03 μm.

14.3 Device Characteristics

Threshold is defined as the gate voltage at which mobile charge Q_s at the source end of the channel changes the surface potential by kT/q (Mead, 1989). The channel charge at threshold is

$$Q_t = \frac{kT}{q}\left(C_{ox}+C_c\right). \tag{14.3.1}$$

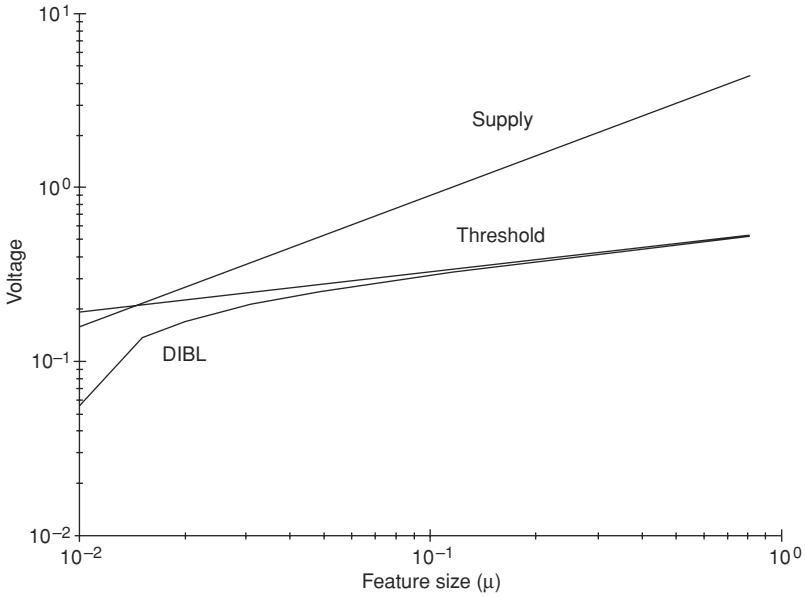


Figure 14.8
Threshold voltage. The middle curve is the nominal threshold voltage, given by Eq. 14.2.1. The bottom curve is the actual threshold voltage, which is lowered from the nominal value by drain-induced barrier lowering (DIBL), given by Eq. 14.2.2. The top curve is the nominal supply voltage from Eq. 14.1.2.

For higher gate voltages, essentially all charge on the gate attracts equal and opposite counter-charge of mobile carriers in the channel. Thus, we can form an excellent estimate of the channel charge Q_s at the source end of the channel:

$$Q_s = C_{ox}\,(V - V_T)\,. \tag{14.3.2}$$

For gate voltages below V_T, channel current decreases exponentially with decreasing gate voltage. At zero gate voltage, the channel charge is

$$Q_s = Q_t\, e^{-q\kappa V_T/kT} \tag{14.3.3}$$

where

$$\kappa = \frac{C_{ox}}{C_c + C_{ox}}.$$

Given Q_t and Q_s, we can compute the saturated channel current for a minimum-sized transistor of any given channel length using Eq. B.28 from Mead

(1989):

$$I_{sat} = Q_s\, v_0 + Q_t\, v_0 \left(\frac{l}{l_0} + 1\right) \left(1 - \sqrt{1 + \frac{2Q_s l}{Q_t l_0} \left(\frac{l}{l_0} + 1\right)^{-2}}\right) \quad (14.3.4)$$

where v_0, the saturated velocity of electrons in silicon, is taken to be 10^7 cm/s (Noor Mohammad, 1992), and $l_0 = D/v_0$ can be thought of as the mean free path of the carrier, which is taken to be 0.007 μm (Mead, 1989).

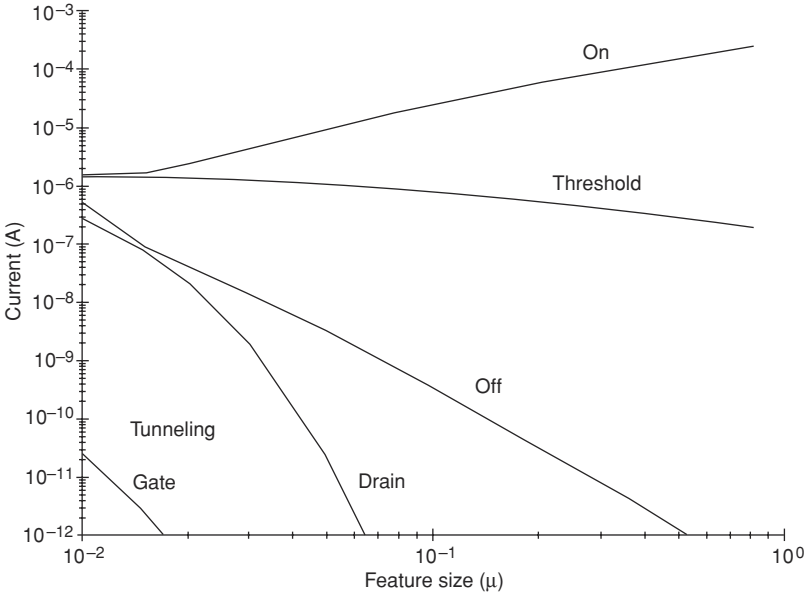


Figure 14.9
Currents characteristic of minimum-sized devices as a function of feature size. We obtain the threshold current I_t by substituting $Q_s = Q_t$ from Eq. 14.3.1 into Eq. 14.3.4. We obtain the on current I_{on} by substituting Q_s from Eq. 14.3.2 into Eq. 14.3.4, using the threshold voltage lowered only by the built-in junction voltage, rather than by the total junction voltage. We obtain the off current I_{off} by substituting Q_s from Eq. 14.3.3 into Eq. 14.3.4, using the threshold voltage as lowered by the full supply voltage. The junction tunneling current was computed from Eq. 14.1.10, assuming the drain area is the square of the feature size. The gate-oxide tunneling current was computed from Eq. 14.1.7, assuming that the full supply voltage is present across the full gate area (the square of the feature size).

We obtain the threshold current I_t by substituting $Q_s = Q_t$ from Eq. 14.3.1 into Eq. 14.3.4. We obtain the on current I_{on} by substituting Q_s

from Eq. 14.3.2 into Eq. 14.3.4, using the threshold voltage lowered only by the built-in junction voltage, rather than by the total junction voltage. We obtain the off current I_{off} by substituting Q_s from Eq. 14.3.3 into Eq. 14.3.4, using the threshold voltage as lowered by DIBL. These expressions thus represent a conservative characterization of the transistor performance, because the on current will be somewhat underestimated.

The several currents associated with a minimum-sized transistor are shown as a function of feature size in Fig. 14.9. The tradeoffs mentioned in the introduction are immediately apparent in this plot. As features become smaller, substrate doping must increase to prevent punch-through. The increase in substrate doping increases the junction electric field, thereby increasing drain-junction tunneling current into the substrate. To limit the tunneling current to a reasonable value, we reduce the supply voltage, thereby reducing the ratio of channel on current to channel off current. The most remarkable conclusion from Fig. 14.9 is that transistors of 0.03 μm channel length still function essentially as do present-day devices. With proper scaling of all parameters of the process, device miniaturization is alive and well. Many issues will arise in the development of ever finer-scale fabrication, but, in the end, the endeavor will prevail.

Given that devices at least 1 order of magnitude smaller than today's are feasible, we may enquire what their characteristics may be. Figure 14.10 shows several quantities of interest. It is clear that discreteness of all quantities will become increasingly important at smaller feature sizes—particularly that of doping ions in the substrate. We have given elsewhere a simple discussion of the effects of discrete substrate charge (Hoeneisen and Mead, 1972b); a recent analysis is presented by Nishinohara et al. (1992).

Perhaps the single most important aspect of device performance is the speed of logic fabricated from any particular technology. We can estimate the time τ required for an elementary logic element to drive another like it:

$$\tau = \frac{V\, C_{tot}}{I_{on}} \tag{14.3.5}$$

where the total capacitance C_{tot} is taken to be three times the sum of the oxide and drain junction capacitances. This delay should correspond rather directly to the delay per stage measured for ring oscillators in any given process, and is plotted along with several experimental points in Fig. 14.11. It is remarkable that, despite the reduction in supply voltage at small feature sizes, logic performance continues to improve. Several authors have emphasized the

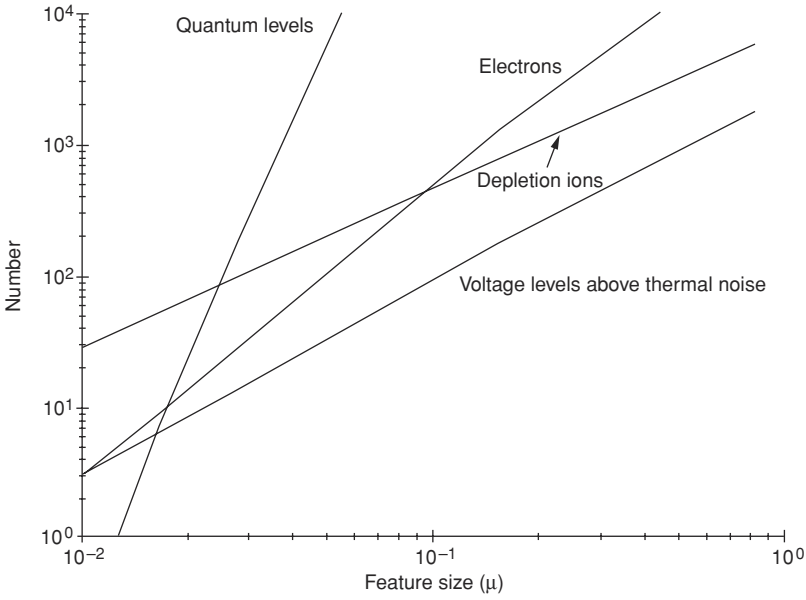


Figure 14.10
Number of signal levels resolvable by a minimum-sized device according to the scaling laws used in this chapter. Thermal noise limits the analog depth representable by a single voltage. The number of voltage levels above thermal noise was taken to be the square root of the minimum stored energy shown in Fig. 14.3, expressed in units of kT. The quantum-level separation was taken to be the energy spacing of states in a one-dimensional box of length $l - x_s - x_d$. The number of electrons under the gate was taken to be the on-value of Q_s multiplied by the gate area (a slight overestimate). The number of depletion ions was taken to be the doping density n given by Eq. 14.1.8, multiplied by the gate area and the depletion depth x from Eq. 14.1.9, using 1 V for ψ. As the number of depletion ions becomes smaller, the range of threshold voltages encountered across a single chip increases. In analog systems, adaptation techniques can mitigate or eliminate the variation among transistors.

improvement in speed that we can make available by reducing threshold and power-supply voltages (Burr and Peterson, 1991; Liu and Svensson, 1993; Murphy, 1993; Lyon, 1993).

The primary effect behind this somewhat counter-intuitive trend is velocity saturation, an excellent recent account of which can be found in Noor Mohammad (1992). We gave an early treatment of the effect of velocity saturation on device characteristics (Hoeneisen and Mead, 1972a); an extended analysis appears in Appendix B of a previous work (Mead, 1989).

The supply voltage V affects the performance of standard CMOS digital logic in three ways: The channel charge is proportional to $V - V_T$; the electric

field in the channel is proportional to V; and the logic swing is proportional to V. For long-channel devices, the carrier velocity is proportional to the electric field in the channel. The channel current is the product of the channel charge and the carrier velocity. Therefore, the device current has a quadratic dependence on the supply voltage. This current must charge the load capacitance to approximately one-half of the supply voltage to achieve a logic transition. This factor cancels one of the V terms in the current, leaving the circuit speed linear in the supply voltage.

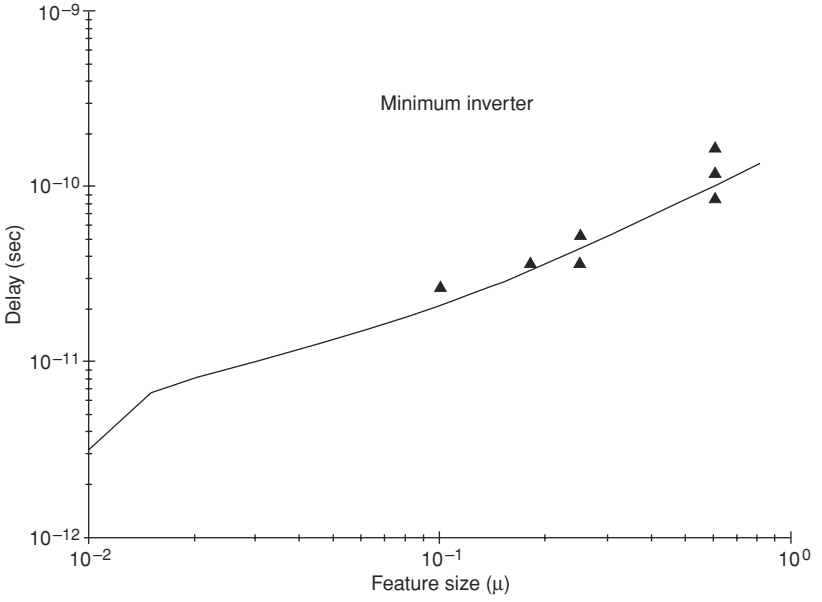


Figure 14.11
Delay of minimally loaded inverter as a function of feature size. Filled triangles are experimental results from ring oscillators reported in the literature. Solid line is the expression given in Eq. 14.3.5.

Once the carrier velocity is saturated, however, increasing the electric field in the channel no longer increases the channel current. Both the charge in transit and the voltage to be traversed by the output are increased by the same factor. In this regime, the only effect of increased supply voltage is an increase in the switching energy, with virtually no increase in performance. Just how close devices of the present day come to this limit can be seen in the delay-

versus-voltage plots in the recent literature; see, for example, (Hori et al., 1992; Chang et al., 1992; Iwase et al., 1993).

Because we have at our disposal the currents associated with all terminals of the transistor, we can evaluate the conductances associated with these currents. For logic devices to function properly, it is necessary that an elementary logic circuit have a gain greater than unity, which in turn requires that the transconductance g_m of the transistor be larger than the sum of all contributions to the drain conductance. As feature size decreases below 0.1 μm, both DIBL and drain-junction tunneling make rapidly increasing contributions to the drain conductance, as can be seen in Fig. 14.12. Despite these parasitic effects, the device is still capable of providing greater than unity gain down to the smallest feature sizes investigated.

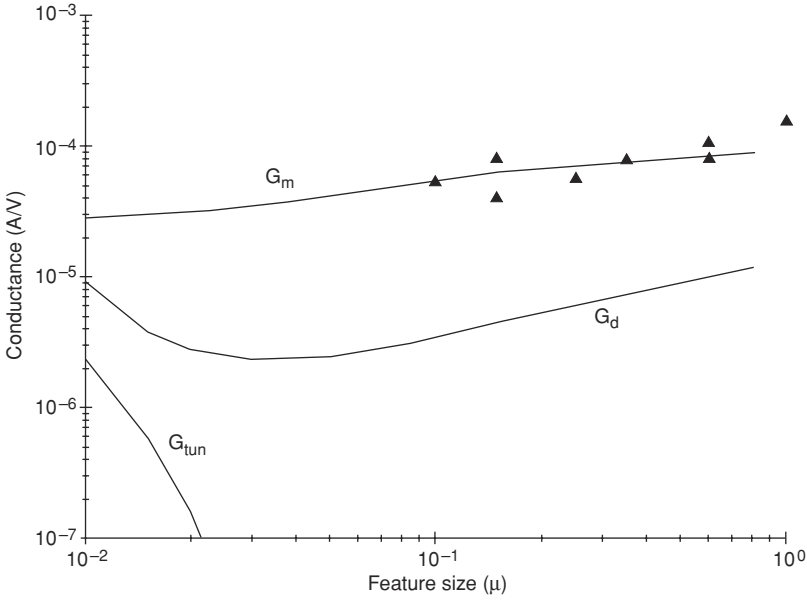


Figure 14.12
Several conductances associated with minimum-sized transistors, as a function of feature size. The top curve is the transconductance. The filled triangles are experimental values given in the literature, normalized to a minimum-sized device at the reported dimension. The second curve is the drain conductance due to DIBL, computed by evaluating Eq. 14.3.4 at a drain voltage equal to V and at 0.9 V, and dividing the difference by 0.1 V. The current through this conductance flows from drain to source. The bottom curve is the drain conductance due to drain-junction tunneling. Current through this conductance flows from drain to substrate.

14.4 System Properties

The enormous effect of device scaling on computational capability becomes apparent only when viewed from the system level. We can estimate the system-level capabilities of digital chips fabricated with advanced processes by extrapolation from present-day systems. The first such extrapolation is the number of devices per unit area. If every transistor in a modern digital chip were to be shrunk to minimum size, the entire active area would cover approximately 2 % of the chip area. If we assume that this coverage factor can be maintained in future designs, the density of active elements scales with feature size, as shown in Fig. 14.13. The system-clock period in today's processors is approximately 100 τ. Even today, it is becoming more economical to break each chip into several processors that can operate in parallel, than it is to merely build larger "dinosaur" processors. For purposes of extrapolation, we can assume that each processor contains 10^6 transistors. The computation available under these clearly oversimplified assumptions is plotted versus feature size in Fig. 14.14. If we further assume that all devices are in fact of minimum size, and that they are clocked at the system-clock frequency, we can estimate the power that will be dissipated by chips built in these advanced technologies. The power attributable to useful switching, and the dissipations of various parasitic currents that do not depend on clock speed, are shown in Fig. 14.15. Down to about 0.03 μm feature size, most of the energy supplied to the chip is dissipated in real, useful computation. Only below this scale do the parasitic currents overwhelm the energy consumed in performing real computation.

14.5 Conclusions

The MOS transistor has become the workhorse of modern microelectronics; it has survived many generations of process scaling to finer feature sizes. In this chapter, we have explored the extent to which the MOS device, as we know it today, can be scaled to still smaller dimensions. There is no data available that provides experimental support for the tunneling currents that will be encountered in the heavily doped source and drain junctions of devices down to 0.03 μm. There is no comparable data to support the theory for oxides in the 10 Å range, nor is there direct experimental verification of the effect of statistical fluctuations on very small structures built in heavily doped material. As such data become available, we will be better able to chart the

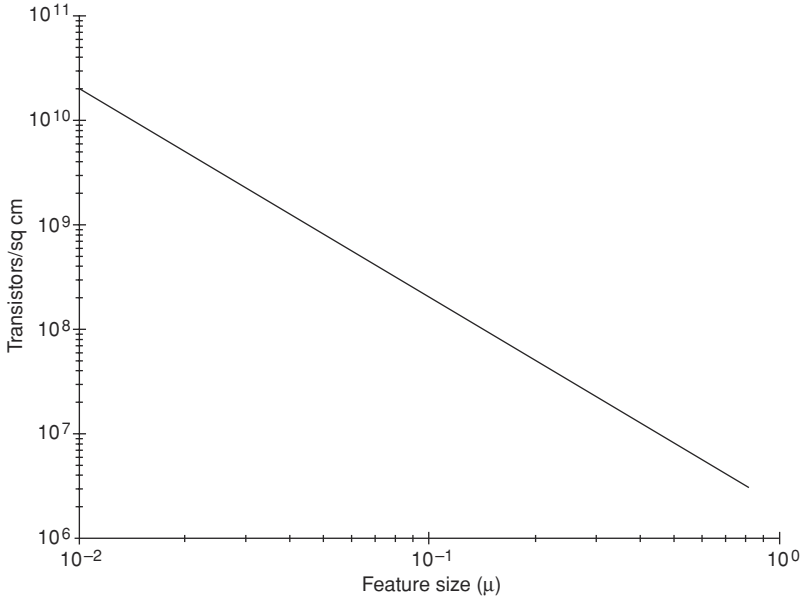


Figure 14.13
Assumed number of active devices per square centimeter of chip area. If all devices are of minimum size, active (transistor channel) area is 2 % of total area.

course of future minification, of which the present study is only an outline. It is already clear that MOS circuits will be integrated to upward of 10^9 devices per square centimeter merely by scaling, without any major change in the conceptual framework that we use today. There are many challenges involved in this technology evolution (Nagata, 1992), but show-stoppers are not expected. The prospect of very high levels of integration was daunting in 1971 when our earlier study was written (Hoeneisen and Mead, 1972b), and is far more daunting today. Whereas massive parallelism is possible in present-day technology, it will clearly become mandatory if we are to realize even a fraction of the potential of more highly evolved technology. Even as this chapter was written, there was far more potential in a square centimeter of silicon than we have developed the paradigms to use, as has often been the case in periods of rapid technological evolution.

We should clarify the "limits" considered here. It is clear that devices much smaller than those treated here can be made to show useful characteristics. Conventional MOS devices can be fabricated on insulating substrates (SOI–

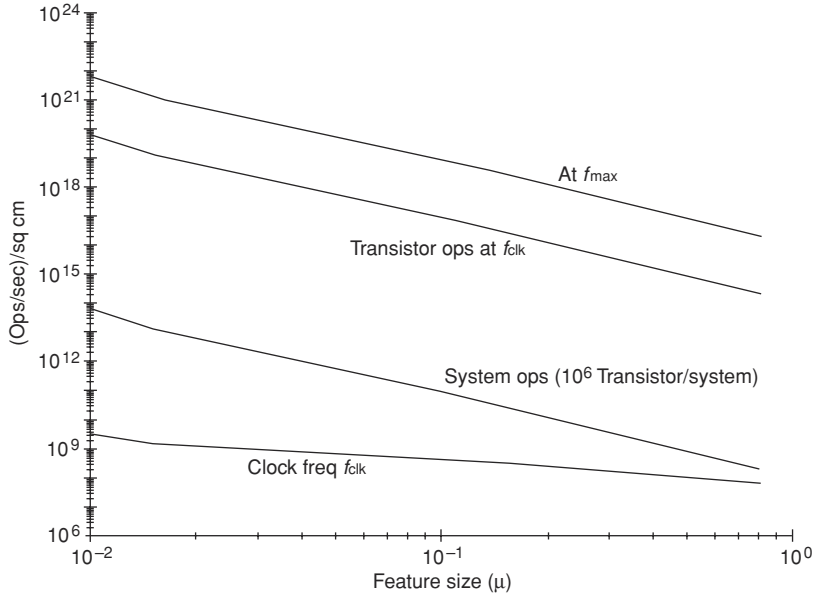


Figure 14.14
Several measures of computation capability per unit area as a function of feature size. The bottom curve is a typical processor clock frequency, the clock period assumed to be 100 times the inverter delay shown in Fig. 14.11. The second curve is the number of systems (of 10^6 transistors each) per square centimeter multiplied by the clock frequency. The third curve is the number of transistors per square centimeter shown in Fig. 14.13 multiplied by the clock frequency. The top curve is the number of transistors per square centimeter multiplied by the reciprocal of the inverter delay shown in Fig. 14.11.

SOS), thereby removing the constraint imposed by substrate tunneling. Much smaller devices are possible at molecular scale. The most obvious example of an extremely small device is an electron-transfer reaction occurring along an amino-acid path, the potential of which is determined by the charge on a nearby atomic site. Such arrangements are thought to occur in many biological systems. The physics of such a transfer corresponds directly to that of an MOS transistor operating in weak inversion (below threshold). Imagining a device that functions is easy; building a device that works is much harder; and having a process by which billions of devices can be constructed in a single physical structure is many orders of magnitude harder still. This study is limited to the consideration of direct extensions to existing technology.

Finally, we emphasize that only the properties of transistors themselves have been considered, and we have not even touched many other important

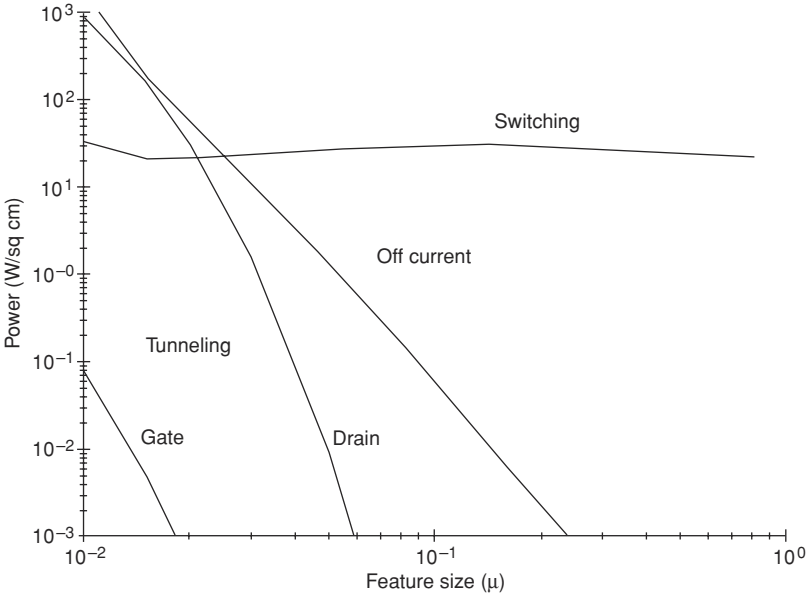


Figure 14.15
Several contributions to the power dissipated by typical digital systems as a function of feature size. The curve labeled **Switching** was obtained by multiplying the number of transistors per unit area shown in Fig. 14.13 by the switching energy shown in Fig. 14.3 and by the clock frequency shown in Fig. 14.14. This power contributes to the performance of computation: It scales directly with clock frequency. In addition to the switching power, there are several parasitic mechanisms by which power is wasted, each being the result of one of the parasitic currents shown in Fig. 14.9. These parasitic mechanisms are present even at zero clock frequency, and perform no useful work. The values shown assume that all devices are of minimum size, and have the full voltage V on their drains. All values depend critically on the assumptions embodied in the scaling laws of Eqs. 14.1.1, 14.1.2, 14.1.8, and 14.2.1. Even slightly different scaling can lead to substantially different results for the smallest feature sizes. The particular laws discussed in this chapter were fine tuned to produce reasonable results down to 0.02 μm. For example, a slight increase in doping density markedly decreases the off current by reducing DIBL, while dramatically increasing the drain-junction tunneling current. Similar tradeoffs can be made with other parameters.

aspects of the technology. Of the latter, interconnect (both within a single chip and across chip boundaries) is obviously a key concern. We have given elsewhere a preliminary discussion of the global scaling properties of a single-chip interconnect network for ultra-dense technology (Mead and Conway, 1980). The topic of interconnect, along with many other issues, such as the fabrication technology itself, deserve a great deal of consideration as the technology evolves. Whatever complications arise, however, it is clear that *the technology will evolve*. It will evolve because that evolution is possible,

because there is so much to be gained at the system level by that evolution, and because the same energy and will on the part of bright, energetic, devoted people that has overcome enormous obstacles in the past will overcome those that lie ahead.

A Appendix: Units and symbols

The SI units

Basic units

Quantity	Unit	Sym.
Length	Meter	m
Mass	Kilogram	kg
Time	Second	s
Therm. temp.	Kelvin	K
Electrical current	Ampere	A
Luminous intensity	Candela	cd
Amount of subst.	Mol	mol

Derived units with special names

Quantity	Unit	Sym.	Derivation
Frequency	Hertz	Hz	s^{-1}
Force	Newton	N	$\mathrm{kg} \cdot \mathrm{m} \cdot \mathrm{s}^{-2}$
Pressure	Pascal	Pa	$\mathrm{N} \cdot \mathrm{m}^{-2}$
Energy	Joule	J	$\mathrm{N} \cdot \mathrm{m}$
Power	Watt	W	$\mathrm{J} \cdot \mathrm{s}^{-1}$
Charge	Coulomb	C	$\mathrm{A} \cdot \mathrm{s}$
Electrical Potential	Volt	V	$\mathrm{W} \cdot \mathrm{A}^{-1}$
Electrical Capacitance	Farad	F	$\mathrm{C} \cdot \mathrm{V}^{-1}$
Electrical Resistance	Ohm	Ω	$\mathrm{V} \cdot \mathrm{A}^{-1}$
Electrical Conductance	Siemens	S	$\mathrm{A} \cdot \mathrm{V}^{-1}$
Magnetic flux	Weber	Wb	$\mathrm{V} \cdot \mathrm{s}$
Magnetic flux density	Tesla	T	$\mathrm{Wb} \cdot \mathrm{m}^{-2}$
Inductance	Henry	H	$\mathrm{Wb} \cdot \mathrm{A}^{-1}$
Luminous flux	Lumen	lm	$\mathrm{cd} \cdot \mathrm{sr}$
Illuminance	Lux	lx	$\mathrm{lm} \cdot \mathrm{m}^{-2}$

Physical Constants

Name	Symbol	Value	Unit
Angstrom Unit	Å	10^{-10}	m
Number π	π	3.141592653589793238	
Number e	e	2.718281828459	
Speed of light in vacuum	c	$2.99792458 \cdot 10^{8}$	m/s (def)
Permittivity of the vacuum	ε_0	$8.854187 \cdot 10^{-12}$	F/m
Permeability of the vacuum	μ_0	$4\pi \cdot 10^{-7}$	H/m
Planck's constant	h	$6.6260755 \cdot 10^{-34}$	Js
Dirac's constant	$\hbar = h/2\pi$	$1.0545727 \cdot 10^{-34}$	Js
Molar gasconstant	R	8.31441	J/mol
Avogadro's constant	N_{A}	$6.0221367 \cdot 10^{23}$	mol^{-1}
Boltzmann's constant	k	$1.380658 \cdot 10^{-23}$	J/K
Elementary charge	q	$1.60217733 \cdot 10^{-19}$	C
Electron mass	m_{e}	$9.1093897 \cdot 10^{-31}$	kg
Electron volt	eV	$1.60218 \cdot 10^{-19} J$	eV

Properties of Si

Name	Symbol	Value	Unit
Breakdown field	E_B	$3 \cdot 10^5$	V/cm
Diffusion coefficient:			
Electron	D_n	34.91	cm^2/s
Hole	D_p	12.41	cm^2/s
Effective mass:			
Electron	m_n^*/m_0	1.18	
Hole	m_n^*/m_0	0.81	
Electron affinity	$q\chi$	4.05	eV
Energy gap	E_g	1.124	eV
Intrinsic carrier concentration	n_i	$1.45 \cdot 10^{10}$	cm^{-3}
Intrinsic resistivity	ρ	$3.16 \cdot 10^5$	$\Omega - cm$
Minority carrier lifetime	τ	$2.5 \cdot 10^{-3}$	s
Mobility:			
Electron	μ_n	1350	$cm^2/V\text{-}s$
Hole	μ_p	480	$cm^2/V\text{-}s$
Relative permittivity	ϵ_r	11.9	
Effective density of states:			
Conduction band	N_c	$2.8 \cdot 10^{19}$	cm^{-3}
Valence band	N_v	$1.04 \cdot 10^{19}$	cm^{-3}

List of Symbols

Symbol	Description	Unit
a	Lattice Constant	Å
c	Speed of light in vacuum	m/s
C	Capacitance	F
D	Diffusion coefficient	m^2/s
E	Energy	eV
E_C	Bottom of conduction band	eV
E_F	Fermi level energy	eV
E_V	Top of valence band	eV
$\mathcal{E}$	Electric Field	V/m
$\mathcal{E}_c$	Critical Field	V/m
$\mathcal{E}_m$	Maximum Field	V/m
f	Frequency	Hz
$F(E)$	Fermi-Dirac distribution function	
I	Current	A
J	Current density	A/m^2
L	Transistor's length	m
m_0	Electron rest mass	kg
m^*	Effective mass	kg
n	Density of free electrons	m^{-3}
n_i	Intrinsic density	m^{-3}
N	Doping concentration	m^{-3}
N_A	Acceptor impurity density	m^{-3}
N_C	Effective density of states in conduction band	m^{-3}
N_D	Donor impurity density	m^{-3}
N_V	Effective density of states in valence band	m^{-3}
p	Density of free holes	m^{-3}
q	Magnitude of electronic charge	C
R	Resistance	Ω
T	Absolute temperature	K
v	Carrier velocity	m/s
v_s	Saturation velocity	m/s
v_{th}	Thermal velocity	m/s

continued on next page

Symbol	Description	Unit
V_{bi}	Built-in potential	V
V_{EB}	Emitter-base voltage	V
V_B	Breakdown voltage	V
W	Transistor width	m
W_b	Base thickness	m
ϵ_0	Permettivity in vacuum	F/m
ϵ_s	Semiconductor permettivity	F/m
ϵ_i	Insulator permettivity	F/m
ϵ_s/ϵ_0	Dielectric constant	
ϵ_i/ϵ_0	Dielectric constant	
μ_0	Permeability in vacuum	H/m
μ_n	Electron mobility	m^2/V-s
μ_p	Hole mobility	m^2/V-s
ρ	Resistivity	Ω-m
U_T	Thermal voltage	V
ω	Angular frequency ($2\pi f$)	Hz

Prefixes

Name	Symbol	Value	Name	Symbol	Value
exa	E	10^{18}	deci	d	10^{-1}
peta	P	10^{15}	centi	c	10^{-2}
tera	T	10^{12}	milli	m	10^{-3}
giga	G	10^{9}	micro	μ	10^{-6}
mega	M	10^{6}	nano	n	10^{-9}
kilo	k	10^{3}	pico	p	10^{-12}
hecto	h	10^{2}	femto	f	10^{-15}
deca	da	10	atto	a	10^{-18}

The ∇ (*nabla*) symbol

The symbol ∇ represents a *vector differential operator*:

$$\vec{\nabla} = \frac{\partial}{\partial x}\vec{e}_x + \frac{\partial}{\partial y}\vec{e}_y + \frac{\partial}{\partial z}\vec{e}_z$$

where $\vec{e}_x, \vec{e}_y$, and $\vec{e}_z$ are the base vectors of unit length directed along the positive directions of the x, y, and z axes respectively.

The symbol ∇^2 is called the *Laplacian operator*:

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

If $f(\cdot) = f(x, y, z)$ represents a *scalar field* with continuous first partial derivatives, then

$$\vec{\nabla} f = \text{grad} f = \frac{\partial f}{\partial x}\vec{e}_x + \frac{\partial f}{\partial y}\vec{e}_y + \frac{\partial f}{\partial z}\vec{e}_z$$

is called the *gradient* of the scalar function f. The gradient operator defines a vector.

If $\vec{v}(x, y, z)$ is a differentiable vector function, with components v_1, v_2, and v_3, then

$$\vec{\nabla} \cdot \vec{v} = \text{div}\ \vec{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$

is called the *divergence* of the vector field $\vec{v}$. The divergence operator defines a scalar.

If $\vec{v}(x, y, z)$ is a differentiable vector function defined as

$$\vec{v}(x, y, z) = v_1\vec{e}_x + v_2\vec{e}_y + v_3\vec{e}_z$$

then

$$\vec{\nabla} \times \vec{v} = \text{curl}\ \vec{v} = \left(\frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z}\right)\vec{e}_x + \left(\frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x}\right)\vec{e}_y + \left(\frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y}\right)\vec{e}_z$$

is called the *curl* of the vector function $\vec{v}$. The curl operator $\text{curl}\ \vec{v}$, also denoted by $\text{rot}\ \vec{v}$, defines a vector.

Basic relations for the mathematical operators

If f and g are two scalar functions, and $\vec{v}$ and $\vec{u}$ are two vector functions, the following relations hold:

$$\vec{\nabla}(fg) = f\vec{\nabla}g + g\vec{\nabla}f, \qquad \vec{\nabla}(f/g) = (1/g^2)(g\vec{\nabla}f - f\vec{\nabla}g)$$

$$\nabla^2 f = \mathrm{div}(\vec{\nabla}f), \qquad \nabla^2(fg) = g\nabla^2 f + 2\vec{\nabla}f \cdot \vec{\nabla}g + f\nabla^2 g$$

$$\mathrm{div}(f\vec{v}) = f\mathrm{div}\ \vec{v} + \vec{v} \cdot \vec{\nabla}, \qquad \mathrm{div}(f\vec{\nabla}g) = f\nabla^2 g + \vec{\nabla}f \cdot \vec{\nabla}g$$

$$\mathrm{curl}(f\vec{v}) = \vec{\nabla}f \times \vec{v} + f\mathrm{curl}\ (\vec{v}), \qquad \mathrm{div}(\vec{u} \times \vec{v}) = \vec{v} \cdot \mathrm{curl} - \vec{u} \cdot \mathrm{curl}\ \vec{v}$$

$$\mathrm{curl}(\vec{\nabla}f) = 0, \qquad \mathrm{div}(\mathrm{curl}\ \vec{v}) = 0$$

This page intentionally left blank

References

Adams, R. W. (1979). Filtering in the log domain, Preprint 1470. In *Audio Engineering Society Convention 63*.

Adams, S. (1996a). *The Dilbert Principle*. Harper Business, New York.

Adams, S. (1996b). *Fugitive from the cubicle police*. Andrews and McMeel, Kansas City, MO.

Adams, S. (1998a). *The Dilbert Future, Thriving on stupidity in the 21st century*. Harper Business, New York.

Adams, S. (1998b). *The Joy of Work*. Harper Business, New York.

Allen, P. E. and Holberg, D. R. (2002). *CMOS Analog Circuit Design*. Oxford University Press, 2nd edition.

Alspector, J. and Allen, R. B. (1987). A neuromorphic VLSI learning system. In Losleben, P., editor, *Proceedings of the 1987 Stanford Conference on Advanced Research in VLSI*, pages 313–349, Cambridge, MA. MIT Press.

Amelio, G. F., Bertram, Jr., W. J., and Tompsett, M. F. (1971). Charge-coupled imaging devices. *IEEE Transactions on Electron Devices*, ED-18:986–992.

Andreou, A. G. and Boahen, K. A. (1994). Neuromorphic information processing II. In Ismail, M. and Fiez, T., editors, *Analog VLSI : Signal and Information Processing*, chapter 8, pages 358–413. McGraw-Hill, New York.

Andreou, A. G. and Boahen, K. A. (1996). Translinear circuits in subthreshold MOS. *Analog Integrated Circuits and Signal Processing*, 9(2):141–166.

Andreou, A. G., Boahen, K. A., Pouliquen, P. O., Pavasovic, A., Jenkins, R. E., and Strohbehn, K. (1991). Current-mode subthreshold MOS circuits for analog VLSI neural systems. *IEEE Transactions on Neural Networks*, 2(2):205–213.

Aritome, S., Shirota, R., Hemink, G., Endoh, T., and Masuoka, F. (1993). Reliability issues of flash memory cells. *Proceedings of the IEEE*, 81(5):776–788.

Arnold, M. G., Bailey, T. A., Cowles, J. R., and Cupal, J. J. (1990). Redundant logarithmic arithmetic. *IEEE Transactions on Computers*, 39(8):1077–1086.

Arora, N. D., Rios, R., and Huang, C.-L. (1995). Modeling the polysilicon depletion effect and its impact on submicrometer CMOS circuit performance. *IEEE Transactions on Electron Devices*, 42(5):935–943.

Asai, S. and Wada, Y. (1997). Technology challenges for integration near and below 0.1 μm. *Proceedings of the IEEE*, 85(4):505–520.

Ashok, S. (1976a). Integrable sinusoidal frequency doubler. *IEEE Journal of Solid-State Circuits*, SC-11(2):341–343.

Ashok, S. (1976b). Translinear root-difference-of-squares circuit. *Electronics Letters*, 12(8):194–195.

Bakker, J. G. C. (1991). Simple analytical expressions for the fringing field and fringing-field-induced transfer time in charge-coupled devices. *IEEE Transactions on Electron Devices*, 38(5):1152–1161.

Barker, R. W. J. and Hart, B. L. (1974). Root-law circuit using monolithic bipolar-transistor arrays. *Electronics Letters*, 10(21):439–440.

Ben-Yishai, R., Lev Bar-Or, R., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the USA*, 92(9):3844–3848.

Bergemont, A., Deleonibus, S., Guegan, G., Guillaumot, B., Laurens, M., and Martin, F. (1989). A high performance CMOS process for submicron 16Mb EPROM. In *International Electron Devices Meeting*, pages 591–594.

Bertsekas, D. P. (1982). *Constrained optimization and Lagrange multiplier methods*. Academic Press, New York.

Boahen, K. A. (1997). *Retinomorphic Vision Systems: Reverse Engineering the Vertebrate Retina*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Boahen, K. A. (1998). Communicating neuronal ensembles between neuromorphic chips. In Lande, T. S., editor, *Neuromorphic Systems Engineering*, pages 229–259. Kluwer, Boston, MA.

Boahen, K. A. and Andreou, A. G. (1992). A contrast sensitive silicon retina with reciprocal synapses. In Moody, J. E., Hanson, S. J., and Lippman, R. P., editors, *Advances in neural information processing systems*, volume 4, pages 764–772, San Mateo, CA. Morgan Kaufmann.

Boyle, W. S. and Smith, G. E. (1970). Charge coupled semiconductor devices. *The Bell System Technical Journal*, 49(4):587–593.

Brajovic, V. and Kanade, T. (1998). Computational sensor for visual tracking with attention. *IEEE Journal of Solid-State Circuits*, 33(8):1199–1207.

Brüggemann, H. (1970). Feedback stabilized four-quadrant analog multiplier. *IEEE Journal of Solid-State Circuits*, SC-5(4):150–159.

Bryant, A., El-Kareh, B., Furukawa, T., Noble, W. P., Nowak, E. J., Schwittek, W., and Tonti, W. (1992). A fundamental performance limit of optimized 3.3-V sub-quarter-micrometer fully overlapped LDD MOSFET's. *IEEE Transactions on Electron Devices*, 39(5):1208–1215.

Burr, J. B. and Peterson, A. M. (1991). Energy considerations in multichip-module based multiprocessors. In *IEEE International Conference on Computer Design*, pages 593–600.

Carlson, A. B. (1986). *Communication Systems*. McGraw-Hill, New York, 3rd edition.

Cauwenberghs, G. and Bayoumi, M. A., editors (1999). *Learning on Silicon: Adaptive VLSI Neural Systems*. Kluwer, Boston, MA.

Chamberlain, S. G. and Lee, J. P. Y. (1983). A novel wide dynamic range silicon photodetector and linear imaging array. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 441–445.

Chang, W.-H., Davari, B., Wordeman, M. R., Taur, Y., Hsu, C. C.-H., and Rodriguez, M. D. (1992). A high-performance 0.25-μm CMOS technology. I. Design and characterization. *IEEE Transactions on Electron Devices*, 39(4):959–966.

Chaparala, P., Shibley, J., and Lim, P. (2000). Threshold voltage drift in PMOSFETs due to NBTI and HCI. In *IEEE International Integrated Reliability Workshop*.

Chiu, K. Y., Moll, J. L., and Manoliu, J. (1982). A bird's beak free local oxidation technology feasible for VLSI circuits fabrication. *IEEE Transactions on Electron Devices*, ED-29(4):536–540.

Choi, J. and Sheu, B. J. (1993). A high-precision VLSI winner-take-all circuit for self-organizing neural networks. *IEEE Journal of Solid-State Circuits*, 28(5):576–584.

Chua, L. O., Desoer, C. A., and Kuh, E. S. (1987). *Linear and nonlinear circuits*, pages 696–700. McGraw-Hill, New York.

Chung, S. S., Kuo, S. N., Yih, C. M., and Chao, T. S. (1997). Performance and reliability evaluations of *p*-channel flash memories with different programming schemes. In *International Electron Devices Meeting technical digest*, pages 295–298.

Chute, C. (2000). *CMOS Versus CCD: Battle for Market Share*. IDC, Framingham, MA.

Chynoweth, A. G., Feldmann, W. L., Lee, C. A., Logan, R. A., and Pearson, G. L. (1960). Internal field emission at narrow silicon and germanium *p-n* junctions. *Physical Review*, 118(2):425–434.

Cohen, M. H. and Andreou, A. G. (1992). Current-mode subthreshold MOS implementation of the Herault-Jutten autoadaptive network. *IEEE Journal of Solid-State Circuits*, 27(5):714–727.

Davari, B., Chang, W. H., Petrillo, K. E., Wong, C. Y., Moy, D., Taur, Y., Wordeman, M. R., Sun, J. Y.-C., Hsu, C. C.-H., and Polcari, M. R. (1992). A high-performance 0.25-μm CMOS technology: II. Technology. *IEEE Transactions on Electron Devices*, 39(4):967–975.

Davari, B., Dennard, R. H., and Shahidi, G. G. (1995). CMOS scaling for high performance and low power – the next ten years. *Proceedings of the IEEE*, 83(4):595–606.

Deen, M. J. and Yan, Z. X. (1992). DIBL in short-channel NMOS devices at 77K. *IEEE*

Transactions on Electron Devices, 39(4):908–915.

Delbrück, T. (1989). A chip that focuses an image on itself. In Mead, C. and Ismail, M., editors, *Analog VLSI Implementation of Neural Systems*, pages 171–188. Kluwer, Boston, MA.

Delbrück, T. (1993). *Investigations of analog VLSI visual transduction and motion processing*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Delbrück, T. and Mead, C. A. (1994). Adaptive photoreceptor with wide dynamic range. In *1994 IEEE International Symposium On Circuits and Systems*, volume 4, pages 339–342. ISCAS '94: London, England, 30 May–2 June.

Delbrück, T. and Mead, C. A. (1995). Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits. In Koch, C. and Li, H., editors, *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, pages 139–161. IEEE Computer Society Press, Los Alamitos, CA.

Demosthenous, A., Smedley, S., and Taylor, J. (1998). A CMOS analog winner-take-all network for large-scale applications. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 45(3):300–304.

Diorio, C. (1997). *Neurally Inspired Silicon Learning: From Synapse Transistors to Learning Arrays*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Diorio, C. (2000). A p-channel MOS synapse transistor with self-convergent memory writes. *IEEE Transactions on Electron Devices*, 47(2):464–472.

Diorio, C., Hasler, P., Minch, B. A., and Mead, C. (1998a). Floating-gate MOS synapse transistors. In Lande, T. S., editor, *Neuromorphic Systems Engineering: Neural Networks in Silicon*, pages 315–338. Kluwer, Boston, MA.

Diorio, C., Hasler, P., Minch, B. A., and Mead, C. A. (1996). A single-transistor silicon synapse. *IEEE Transactions on Electron Devices*, 43(11):1972–1980.

Diorio, C., Hasler, P., Minch, B. A., and Mead, C. A. (1997a). A complementary pair of four-terminal silicon synapses. *Analog Integrated Circuits and Signal Processing*, 13(1–2):153–166.

Diorio, C., Hasler, P., Minch, B. A., and Mead, C. A. (1997b). A floating-gate MOS learning array with locally computed weight updates. *IEEE Transactions on Electron Devices*, 44(12):2281–2289.

Diorio, C. J., Hasler, P. E., Minch, B. A., and Mead, C. A. (1997c). Semiconductor structure for long term learning. U.S. Patent No. 5,627,392, Issued May 6.

Diorio, C. J., Hasler, P. E., Minch, B. A., and Mead, C. A. (1998b). Three-terminal silicon synaptic device. U.S. Patent No. 5,825,063, Issued October 20.

Doorenbosch, F. and Goinga, Y. (1976). Integrable, wideband, automatic volume control (A.V.C.) using Pythagoras's Law for amplitude detection. *Electronics Letters*, 12(16):418–420.

Drakakis, E. M., Payne, A. J., and Toumazou, C. (1997). Bernoulli operator: A low-level approach to log-domain processing. *Electronics Letters*, 33(12):1008–1009.

Drakakis, E. M., Payne, A. J., and Toumazou, C. (1998). Multiple feedback log-domain filters. In *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems*, volume 1, pages 317–320. ISCAS '98: Monterey, CA, 31 May–3 June.

Drakakis, E. M., Payne, A. J., and Toumazou, C. (1999). "Log-domain state-space": A systematic transistor-level approach for log-domain filtering. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46(3):290–305.

Dunlap, Jr., W. C. (1957). *An Introduction to Semiconductors*. Wiley, New York.

Edgar, A. D. and Lee, S. C. (1979). FOCUS microcomputer number system. *Communications of the ACM*, 22(3):166–177.

Enz, C. and Punzenberger, M. (1999). 1-V Log-domain filters. In Huijsing, J., van de Plassche, R., and Sansen, W., editors, *Analog Circuit Design: Volt Electronics; Mixed-Mode Systems; Low-Noise and RF Power Amplifiers for Telecommunication*, pages 33–67. Kluwer, Boston, MA.

Enz, C. C. (1989). *High Precision CMOS Micropower Amplifiers*. Ph.D. thesis, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland. No. 802.

Enz, C. C., Krummenacher, F., and Vittoz, E. A. (1995). An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications. *Analog Integrated Circuits and Signal Processing*, 8(1):83–114.

Enz, C. C. and Vittoz, E. A. (1997). MOS transistor modeling for low-voltage and low-power analog IC design. *Microelectronic Engineering*, 39:59–76.

Etienne-Cummings, R., Van der Spiegel, J., and Mueller, P. (1996). VLSI model of Primate visual smooth pursuit. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 706–712, Cambridge, MA. MIT Press.

Fabre, A. (1984). An integrable multiple output translinear current converter. *International Journal of Electronics*, 57(5):713–717.

Fabre, A. (1985). Translinear current conveyors implementation. *International Journal of Electronics*, 59(5):619–623.

Fabre, A. (1986). The translinear operational current amplifier: A new building block. *International Journal of Electronics*, 60(2):275–279.

Fabre, A. (1988). Translinear Current-Controlled Current Amplifier. *Electronics Letters*, 24(9):548–549.

Fabre, A. and Rochegude, P. (1987). Current processing circuits with translinear operational current amplifiers. *International Journal of Electronics*, 63(1):9–28.

Fair, R. B. and Wivell, H. W. (1976). Zener and avalanche breakdown in As-implanted low-voltage Si n-p junctions. *IEEE Transactions on Electron Devices*, ED-23(5):512–518.

Fjeldly, T. A. and Shur, M. (1993). Threshold voltage modeling and the subthreshold regime of operation of short-channel MOSFET's. *IEEE Transactions on Electron Devices*, 40(1):137–145.

Fossum, E. R. (1989). Architectures for focal plane image processing. *Optical Engineering*, 28(8):865–871.

Fossum, E. R. (1993). Active pixel sensors: Are CCDs dinosaurs? In Blouke, M. M., editor, *Charge-Coupled Devices and Solid State Optical Sensors III, Proceedings of the SPIE*, volume 1900, pages 2–14.

Fossum, E. R. (1997). CMOS image sensors: Electronic camera-on-a-chip. *IEEE Transactions on Electron Devices*, 44(10):1689–1698.

Frey, D. R. (1993). Log-domain filtering: An approach to current-mode filtering. *IEE Proceedings G: Circuits, Devices and Systems*, 140(6):406–416.

Frey, D. R. (1996a). Explicit Log Domain Root-Mean-Square Detector. U.S. Patent No. 5,585,757, Issued December 17.

Frey, D. R. (1996b). Exponential state space filters: A generic current-mode design strategy. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 43(1):34–42.

Fried, R. and Enz, C. C. (1996). CMOS parametric current amplifier. *Electronics Letters*, 32(14):1249–1250.

Frohman-Bentchkowsky, D. (1971). Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure. *Applied Physics Letters*, 18(8):332–334.

Fujita, O. and Amemiya, Y. (1993). A floating-gate analog memory device for neural networks. *IEEE Transactions on Electron Devices*, 40(11):2029–2035.

Genin, R. and Konn, R. (1979). Sinusoidal frequency doubler. *Electronics Letters*, 15(2):47–48.

Gilbert, B. (1968a). A DC-500 MHz Amplifier/Multiplier Principle. In Raper, J. A. A., editor, *1968 International Solid-State Circuits Conference Digest of Technical Papers*, volume XI, pages 114–115, New York. L. Winner. Philadelphia, PA, 14–16 February.

Gilbert, B. (1968b). A new wide-band amplifier technique. *IEEE Journal of Solid-State Circuits*, SC-3(4):353–365.

Gilbert, B. (1968c). A precise four-quadrant multiplier with subnanosecond response. *IEEE Journal of Solid-State Circuits*, SC-3(4):365–373.

Gilbert, B. (1974). A high-performance monolithic multiplier using active feedback. *IEEE Journal of Solid-State Circuits*, SC-9(6):364–373.

Gilbert, B. (1975). Translinear circuits: A proposed classification. *Electronics Letters*, 11(1):14–16. See also Errata, 11(6):136, March 1975.

Gilbert, B. (1976). High-accuracy vector-difference and vector-sum circuits. *Electronics Letters*, 12(11):293–294.

Gilbert, B. (1983). A four-quadrant analog divider/multiplier with 0.01 % distortion. In *Digest of Technical Papers of the 1983 IEEE International Solid-State Circuits Conference*, pages 248–249. Philadelphia, PA.

Gilbert, B. (1990). Current-mode circuits from a translinear viewpoint: A tutorial. In Tomazou, C., Lidgey, F. J., and Haigh, D. G., editors, *Analogue IC design: the current-mode approach*, chapter 2, pages 11–91. Peregrinus, Stevenage, Herts., UK.

Gilbert, B. (1993). Translinear circuits – 25 years on. Part I: The foundations. *Electronic Engineering*, 65(800):21–24.

Gilbert, B. (1996). Translinear circuits: An historical review. *Analog Integrated Circuits and Signal Processing*, 9(2):95–118.

Gilbert, B. and Counts, L. W. (1976). A monolithic RMS-DC converter with Crest-Factor compensation. In *Digest of Technical Papers of the 1976 IEEE International Solid-State Circuits Conference*, pages 110–111. Philadelphia, PA.

Gilbert, B. and Holloway, P. (1980). A wideband two-quadrant analog multiplier. In *Digest of Technical Papers of the IEEE International Solid-State Circuits Conference*, pages 200–201. San Francisco, CA.

Godfrey, M. D. (1992). CMOS device modeling for subthreshold circuits. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 39(8):532–539.

Gray, P. R., Hurst, P. J., Lewis, S. H., and Meyer, R. G. (2001). *Analysis and Design of Analog Integrated Circuits*. Wiley, New York, 4th edition.

Gray, P. R. and Meyer, R. G. (1993). *Analysis and design of analog integrated circuits*, pages 716–717. Wiley, New York, 3rd edition.

Gregorian, R. (1999). *Introduction to CMOS OP-AMPs and Comparators*. Wiley, New York.

Gregorian, R. and Temes, G. C. (1986). *Analog MOS Integrated Circuits for Signal Processing*. Wiley, New York.

Grossberg, S. (1978). Competition, decision, and consensus. *Journal of Mathematical Analysis and Applications*, 66(2):470–493.

Grove, A. S. (1967). *Physics and Technology of Semiconductor Devices*. Wiley, New York.

Hackbarth, E. and Tang, D. D.-L. (1988). Inherent and stress-induced leakage in heavily doped silicon junctions. *IEEE Transactions on Electron Devices*, 35(12):2108–2118.

Häfliger, P. and Mahowald, M. (1998). Weight vector normalization in an analog VLSI artificial neuron using a backpropagating action potential. In *Neuromorphic Systems: Engineering Silicon from Neurobiology*, chapter 16, pages 191–196. World Scientific.

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.

Hall, E. L., Lynch, D. D., and Dwyer, III, S. J. (1970). Generation of products and quotients using approximate binary logarithms for digital filtering applications. *IEEE Transactions on Computers*, C-19(2):97–105.

Han, Y. P. and Ma, B. (1984). Isolation process using polysilicon buffer layer for scaled MOS/VLSI. *Journal of the Electrochemical Society*, 131(3):85C. Abstract 67, Cincinnati, Ohio meeting, May 6-11.

Hasler, P., Andreou, A. G., Diorio, C., Minch, B. A., and Mead, C. A. (1998). Impact ionization and hot-electron injection derived consistently from Boltzmann transport. *VLSI Design*, 8(1–4):455–461.

Hasler, P. E. (1997). *Foundations of Learning in Analog VLSI*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Hastings, A. (2001). *The Art of Analog Layout*. Prentice Hall, Upper Saddle River, NJ.

Hawkins, G. A. (1985). Lateral profiling of interface states along the sidewalls of channel-stop isolation. *Solid State Electronics*, 28(9):945–956.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Reading, MA.

Hoeneisen, B. and Mead, C. A. (1972a). Current-voltage characteristics of small size MOS transistors. *IEEE Transactions on Electron Devices*, ED-19(3):382–383.

Hoeneisen, B. and Mead, C. A. (1972b). Fundamental limitations in microelectronics — I. MOS technology. *Solid-state Electronics*, 15:819–829.

Holler, M., Tam, S., Castro, H., and Benson, R. (1989). An electrically trainable artificial neural network (ETANN) with 10240 floating gate synapses. In *Proceedings of the 1989 IEEE INNS International Joint Conference on Neural Networks*, volume 2, pages 191–196. Washington, D.C.

Holloway, T. C., Dixit, G. A., Grider, D. T., Ashburn, S. P., Aggarwal, R., Shih, A., Zhang, X., Misium, G., Esquivel, A. L., Jain, M., Madan, S., Breedijk, T., Singh, A., Thakar, G., Shinn, G., Riemenschneider, B., O'Brien, S., Frystak, D., Kittl, J., Amerasekera, A., Aur, S., Nicollian, P., Aldrich, D., Eklund, B., Appel, A., Bowles, C., and Parrill, T. (1997). 0.18 μm CMOS technology for high-performance, low-power and RF applications. In *1997 Symposium on VLSI Technology: digest of technical papers*, pages 13–14.

Hori, T., Akamatsu, S., and Odake, Y. (1992). Deep-submicrometer CMOS technology with reoxidized or annealed nitrided-oxide gate dielectrics prepared by rapid thermal processing. *IEEE Transactions on Electron Devices*, 39(1):118–126.

Horiuchi, T. K. and Koch, C. (1999). Analog VLSI-based modeling of the primate oculomotor system. *Neural Computation*, 11(1):243–265.

Horowitz, P. and Hill, W. (1989). *The Art of Electronics*. Cambridge University Press, 2nd edition.

Hu, C. (1993). Future CMOS scaling and reliability. *Proceedings of the IEEE*, 81(5):682–689.

Huijsing, J. H. (2001). *Operational Amplifiers: Theory and Design*. Kluwer, Dordrecht, The Netherlands.

Huijsing, J. H., Lucas, P., and de Bruin, B. (1982). Monolithic analog multiplier-divider. *IEEE Journal of Solid-State Circuits*, SC-17(1):9–15.

Indiveri, G. (2000). Modeling selective attention using a neuromorphic analog VLSI device. *Neural Computation*, 12(12):2857–2880.

Indiveri, G. (2001a). A current-mode hysteretic winner-take-all network, with excitatory and inhibitory coupling. *Analog Integrated Circuits and Signal Processing*, 28(3):279–291.

Indiveri, G. (2001b). A neuromorphic VLSI device for implementing 2-D selective attention systems. *IEEE Transactions on Neural Networks*, 12(6):1455–1463.

Indiveri, G., Kramer, J., and Koch, C. (1996). System implementations of analog VLSI velocity sensors. *IEEE Micro*, 16(5):40–49.

Ismail, M. and Fiez, T. (1994). *Analog VLSI : Signal and Information Processing*. McGraw-Hill, New York.

Iwase, M., Mizuno, T., Takahashi, M., Niiyama, H., Fukumoto, M., Ishida, K., Inaba, S., Takigami, Y., Sanda, A., Toriumi, A., and Yoshimi, M. (1993). High-performance 0.10-μm CMOS devices operating at room temperature. *IEEE Electron Device Letters*, 14(2):51–53.

Johns, D. A. and Martin, K. (1997). *Analog integrated circuit design*. Wiley, New York.

Kahng, D. (1967). Semipermanent memory using capacitor charge storage and IGFET read-out. *The Bell System Technical Journal*, XLVI(6):1296–1300.

Kahng, D. and Sze, S. M. (1967). A floating-gate and its applications to memory devices. *The Bell System Technical Journal*, XLVI(6):1288–1295.

Kaski, S. and Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks*, 7(6/7):973–984.

Kelly, R. D. (1970). Electronic circuit analysis and design by driving-point impedance techniques. *IEEE Transactions on Education*, E-13(3):154–167.

Kerns, D. A., Tanner, J. E., Sivilotti, M. A., and Luo, J. (1991). CMOS UV-writeable non-volatile analog storage. In Séquin, C. H., editor, *Advanced Research in VLSI*, pages 245–261. MIT Press, Cambridge, MA.

Kimizuka, N., Yamaguchi, K., Imai, K., Iizuka, T., Liu, C. T., Keller, R. C., and Horiuchi, T. (2000). NBTI enhancement by nitrogen incorporation into ultrathin gate oxide for 0.10μm gate CMOS generation. In *Digest of Technical Papers / 2000 Symposium on VLSI Technology*, pages 92–93, Piscataway, NJ. IEEE Electron Devices Society.

Kingsbury, N. G. and Rayner, P. J. W. (1971). Digital filtering using logarithmic arithmetic. *Electronics Letters*, 7(2):56–58.

Kittel, C. (1996). *Introduction to Solid State Physics*. Wiley, New York, 7th edition.

Konn, R. and Genin, R. (1979). High-performance aperiodic frequency multiplying. *Electronics Letters*, 15(6):187–189.

Kooi, E., van Lierop, J. G., and Appels, J. A. (1976). Formation of Silicon Nitride at a Si-SiO_2 interface during local oxidation of silicon and during heat-treatment of oxidized silicon in NH_3 gas. *Solid-State Science and Technology, Journal of the Electrochemical Society*, 123(7):1117–1120.

Kramer, J., Sarpeshkar, R., and Koch, C. (1997). Pulse-based analog VLSI velocity sensors. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 44(2):86–101.

Krieger, J. B. (1966). Theory of electron tunneling in semiconductors with degenerate band structure. *Annals of Physics*, 36:1–60.

Kuroi, T., Uchida, T., Horita, K., Sakai, M., Inoue, Y., and Nishimura, T. (1998). Stress analysis of shallow trench isolation for 256M DRAM and beyond. In *International Electron Devices Meeting technical digest*, pages 141–144.

Kurokawa, T. and Mizukoshi, T. (1991). Computer Graphics Using Logarithmic Number Systems. *The Transactions of the Institute of Electronics, Information and Communication Engineers E*, 74(2):447–451.

Lai, F. (1991). A 10ns hybrid number system data execution unit for digital signal processing systems. *IEEE Journal of Solid-State Circuits*, 26(4):590–599.

Lai, F.-S. and Wu, C.-F. E. (1991). A hybrid number system processor with geometric and complex arithmetic capabilities. *IEEE Transactions on Computers*, 40(8):952–962.

Lakshmikumar, K. R., Hadaway, R. A., and Copeland, M. A. (1986). Characterization and modeling of mismatch in MOS transistors for precision analog design. *IEEE Journal of Solid-State Circuits*, SC-21(6):1057–1066.

LaMaire, R. O. and Lang, J. H. (1986). Performance of digital linear regulators which use logarithmic arithmetic. *IEEE Transactions on Automatic Control*, AC-31(5):394–400.

Lang, J. H., Zukowski, C. A., LaMaire, R. O., and An, C. H. (1985). Integrated-circuit logarithmic arithmetic units. *IEEE Transactions on Computers*, C-34(5):475–483.

Lau, K. T. and Lee, S. T. (1998). A CMOS winner-takes-all circuit for self-organizing neural networks. *International Journal of Electronics*, 84(2):131–136.

Lazzaro, J. (1990). *Silicon Models of Early Audition*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Lazzaro, J. and Mead, C. A. (1989). A silicon model of auditory localization. *Neural Computation*, 1(1):47–57.

Lazzaro, J., Ryckebusch, S., Mahowald, M. A., and Mead, C. A. (1989). Winner-take-all networks of $O(n)$ complexity. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems*, volume 1, pages 703–711, San Mateo, CA. Morgan Kaufmann.

Lenzlinger, M. and Snow, E. H. (1969). Fowler-Nordheim tunneling into thermally grown SiO_2. *Journal of Applied Physics*, 40(1):278–283.

Lewis, D. M. (1995). 114 MFLOPS logarithmic number system arithmetic unit for DSP applications. *IEEE Journal of Solid-State Circuits*, 30(12):1547–1553.

Li, G. P., Hackbarth, E., and Chen, T.-C. (1988). Identification and implication of a perimeter tunneling current component in advanced self-aligned bipolar transistors. *IEEE Transactions on Electron Devices*, 35(1):89–95.

Liu, D. and Svensson, C. (1993). Trading speed for low power by choice of supply and threshold voltages. *IEEE Journal of Solid-State Circuits*, 28(1):10–17.

Liu, S.-C. (1996). Silicon model of motion adaptation in the fly visual system. In *Proceedings of the Third Joint UCSD/Caltech Symposium*, pages 1–10.

Liu, S.-C. (1999). Silicon retina with adaptive filtering properties. *Analog Integrated Circuits and Signal Processing*, 18(2/3):243–254.

Liu, S.-C. (2000). A winner-take-all circuit with controllable soft max property. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 717–723. MIT Press, Cambridge, MA.

Lo, S.-H., Buchanan, D. A., Taur, Y., and Wang, W. (1997). Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's. *IEEE Electron Device Letters*, 18(5):209–211.

Logan, R. A. and Chynoweth, A. G. (1963). Effect of degenerate semiconductor band structure on current-voltage characteristics of silicon tunnel diodes. *Physical Review*, 131(1):89–95.

Lyon, R. F. (1993). Cost, power, and parallelism in speech signal processing. In *Proceedings of the IEEE 1993 Custom Integrated Circuits Conference*, pages 15.1.1–15.1.9. San Diego, CA.

Maher, M. A. C. (1989). *A charge-controlled model for MOS transistors*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Mahowald, M. (1994). *An Analog VLSI System for Stereoscopic Vision*. Kluwer, Boston, MA.

Maloberti, F. (2001). *Analog Design for CMOS VLSI Systems*. Kluwer, Dordrecht, The Netherlands.

Matsuda, S., Sato, T., Yoshimura, H., Takegawa, Y., Sudo, A., Mizushima, I., Tsunashima, Y., and Toyoshima, Y. (1998). Novel corner rounding process for shallow trench isolation utilizing MSTS (Micro-Structure Transformation of Silicon). In *International Electron Devices Meeting technical digest*, pages 137–140.

McCreary, J. L. (1981). Matching properties, and voltage and temperature dependence of MOS capacitors. *IEEE Journal of Solid-State Circuits*, SC-16(6):608–616.

Mead, C. A. (1989). *Analog VLSI and Neural Systems*. Addison-Wesley, Reading, MA.

Mead, C. A. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE*, 78(10):1629–1636.

Mead, C. A. (1994). Scaling of MOS technology to submicrometer feature sizes. *Journal of VLSI Signal Processing*, 8(1):9–25.

Mead, C. A. and Conway, L. A. (1980). *Introduction to VLSI Systems*. Addison-Wesley,

Reading, MA.

Mendis, S. K., Kemeny, S. E., Gee, R. C., Pain, B., Staller, C. O., Kim, Q., and Fossum, E. R. (1997). CMOS Active Pixel Image Sensors for highly integrated imaging systems. *IEEE Journal of Solid-State Circuits*, 32(2):187–197.

Minch, B. A. (1997). *Analysis, Synthesis, and Implementation of Networks of Multiple-Input Translinear Elements*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Minch, B. A. (2000a). Floating-gate techniques for assessing mismatch. In *Emerging technologies for the 21st century: Proceedings of the 2000 IEEE International Symposium on Circuits and Systems*, volume 4, pages 385–388. ISCAS 2000 Geneva, Switzerland, 28–31 May.

Minch, B. A. (2000b). Synthesis of dynamic multiple-input translinear element networks. In *Emerging technologies for the 21st century: Proceedings of the 2000 IEEE International Symposium on Circuits and Systems*, volume 1, pages 483–486. ISCAS 2000 Geneva, Switzerland, 28–31 May.

Minch, B. A., Diorio, C., Hasler, P., and Mead, C. A. (1996). Translinear circuits using subthreshold floating-gate MOS transistors. *Analog Integrated Circuits and Signal Processing*, 9(2):167–180.

Minch, B. A., Hasler, P., and Diorio, C. (1998). The multiple-input translinear element: A versatile circuit element. In *Proceedings of the 1998 IEEE International Symposium on Circuits and Systems*, volume 1, pages 527–530. ISCAS '98: Monterey, CA, 31 May–3 June.

Minch, B. A., Hasler, P., and Diorio, C. (1999). Synthesis of multiple-input translinear element networks. In *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems*, volume 2, pages 236–239. ISCAS '99: Orlando, FL, 30 May–2 June.

Mitchell, Jr., J. N. (1962). Computer multiplication and division using binary logarithms. *IRE Transactions on Electronic Computers*, EC-11(4):512–517.

Montoro, C. G., Schneider, M. C., and Cunha, A. I. A. (1999). A current-based MOSFET model for integrated circuit design. In Sánchez-Sinencio, E. and Andreou, A. G., editors, *Low-Voltage/Low-Power Integrated Circuits and Systems*, pages 7–55. IEEE Press, Piscataway, NJ.

Moss, T. S., editor (1980). *Handbook on Semiconductors*, volume 1–4. North-Holland, Amsterdam.

Mudra, R., Hahnloser, R., and Douglas, R. J. (1999). Integrating neuromorphic action-oriented perceptual inputs to generate a navigation behaviour for a robot. *International Journal of Neural Systems*, 9(5):411–416.

Mulder, J., Serdijn, W. A., van der Woerd, A. C., and van Roermund, A. H. M. (1996). Dynamic translinear RMS-DC converter. *Electronics Letters*, 32(22):2067–2068.

Mulder, J., Serdijn, W. A., van der Woerd, A. C., and van Roermund, A. H. M. (1997a). A current-mode synthesis method for translinear companding filters. In *Proceedings of the Fourth IEEE International Conference on Electronics, Circuits, and Systems*, volume 3, pages 1419–1422. Cairo.

Mulder, J., Serdijn, W. A., van der Woerd, A. C., and van Roermund, A. H. M. (1997b). Dynamic translinear circuits—An overview. In *Proceedings of the 2nd IEEE-CAS Region 8 Workshop on Analog and Mixed IC Design*, pages 65–72. Baveno, Italy.

Mulder, J., Serdijn, W. A., van der Woerd, A. C., and van Roermund, A. H. M. (1999). *Dynamic Translinear and Log-Domain Circuits: Analysis and Synthesis*. Kluwer, Boston, MA.

Mulder, J., van der Woerd, A. C., Serdijn, W. A., and van Roermund, A. H. M. (1995). Application of the back gate in MOS weak inversion translinear circuits. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 42(11):958–962.

Mulder, J., van der Woerd, A. C., Serdijn, W. A., and van Roermund, A. H. M. (1997c). An RMS-DC converter based on the dynamic translinear principle. *IEEE Journal of Solid-State Circuits*, 32(7):1146–1150.

Murphy, B. T. (1993). Minimization of transistor delay at a given power density. *IEEE Transactions on Electron Devices*, 40(2):414–420.

Nagata, M. (1992). Limitations, innovations, and challenges of circuits and devices into a half micrometer and beyond. *IEEE Journal of Solid-State Circuits*, 27(4):465–472.

Nandakumar, M., Chatterjee, A., Sridhar, S., Joyner, K., Rodder, M., and Chen, I.-C. (1998). Shallow trench isolation for advanced ULSI CMOS technologies. In *International Electron Devices Meeting technical digest*, pages 133–136.

Nass, M. M. and Cooper, L. N. (1975). A theory for the development of feature detecting cells in visual cortex. *Biological Cybernetics*, 19:1–18.

Nishinohara, K., Shigyo, N., and Wada, T. (1992). Effects of microscopic fluctuations in dopant distributions on MOSFET threshold voltage. *IEEE Transactions on Electron Devices*, 39(3):634–639.

Nixon, R. H., Kemeny, S. E., Staller, C. O., and Fossum, E. R. (1995). 128x128 CMOS photodiode-type active pixel sensor with on-chip timing, control and signal chain electronics. In Blouke, M. M., editor, *Charge-Coupled Devices and Solid State Optical Sensors V. Proceedings of the SPIE*, volume 2415, pages 117–123.

Noble, P. J. W. (1968). Self-scanned silicon image detector arrays. *IEEE Transactions on Electron Devices*, ED-15(4):202–209.

Noor Mohammad, S. (1992). Unified model for drift velocities of electrons and holes in semiconductors as a function of temperature and electric field. *Solid-State Electronics*, 35(10):1391–1396.

Normand, G. (1985). Translinear current conveyors. *International Journal of Electronics*, 59(6):771–777.

Nyquist, H. (1928). Thermal agitation of electric charge in conductors. *Physical Review*, 32(1):110–113.

Oppenheim, A. V., Willsky, A. S., and Nawab, S. H. (1996). *Signals & Systems*. Prentice Hall, Upper Saddle River, NJ, 2nd edition.

Paterson, W. L. (1963). Multiplication and logarithmic conversion by operational amplifier–transistor circuits. *The Review of Scientific Instruments*, 34(12):1311–1316.

Pavasović, A. (1991). *Subthreshold region MOSFET mismatch analysis and modeling for analog VLSI systems*. Ph.D. thesis, The Johns Hopkins University, Baltimore, MD.

Pavasović, A., Andreou, A. G., and Westgate, C. R. (1994). Characterization of subthreshold MOS mismatch in transistors for VLSI systems. *Journal of VLSI Signal Processing*, 8(1):75–85.

Payne, A. and Thanachayanont, A. (1997). Translinear circuit for phase detection. *Electronics Letters*, 33(18):1507–1509.

Payne, A., Thanachayanont, A., and Papavassilliou, C. (1998). A 150-MHz translinear phase-locked loop. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 45(9):1220–1231.

Pelgrom, M. J. M., Duinmaijer, A. C. J., and Welbers, A. P. G. (1989). Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1440.

Perry, D. and Roberts, G. W. (1996). The design of log-domain filters based on the operational simulation of *LC* ladders. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 43(11):763–774.

Plummer, J. D., Deal, M., and Griffen, P. B. (2000). *Silicon VLSI Technology: Fundamentals, Practice and Modeling*. Prentice Hall, Upper Saddle River, NJ.

Pookaiyaudom, S. and Mahattanakul, J. (1995). A 3.3 Volt high-frequency capacitorless electronically-tunable log-domain oscillator. In *1995 IEEE International Symposium on Circuits and Systems*, volume 2, pages 829–832. ISCAS '95: Seattle, WA, 30 April–3 May.

Poularikas, A. D. and Seely, S. (1994). *Signals and Systems*. PWS-KENT, Boston, MA, 2nd edition.

Punzenberger, M. and Enz, C. (1996). A new 1.2 V BiCMOS log-domain integrator for companding current-mode filters. In *1996 IEEE International Symposium on Circuits and Systems*, volume 1, pages 125–128. ISCAS '96: Atlanta, GA, 12–15 May.

Razavi, B. (2001). *Design of analog CMOS integrated circuits*. McGraw-Hill, Boston, MA.

Reimbold, G. (1984). Modified 1/f trapping noise theory and experiments in MOS transistors biased from weak to strong inversion—influence of interface states. *IEEE Transactions on Electron Devices*, 31(9):1190–1197.

Reisch, M. (1990). Tunneling-induced leakage currents in *pn* junctions. *AEÜ: Archiv für Elektronik und Übertragungstechnik*, 44(5):368–376.

Ricco, B., Versari, R., and Esseni, D. (1996). Characterization of polysilicon-gate depletion in MOS structures. *IEEE Electron Device Letters*, 17(3):103–105.

Robinson, F. N. H. (1974). *Noise and Fluctuations in Electronic Devices and Circuits*. Oxford University Press.

Rose, A. (1973). *Vision: Human and Electronic*. Plenum Press, New York.

Sanchez, J. J. and DeMassa, T. A. (1991). Review of carrier injection in the silicon/silicon-dioxide system. *IEE Proceedings G: Circuits, Devices and Systems*, 138(3):377–389.

Sarpeshkar, R. (1997). *Efficient Precise Computation with Noisy Components: Extrapolating from an Electronic Cochlea to the Brain*. Ph.D. thesis, California Institute of Technology, Pasadena, CA.

Sarpeshkar, R. (1998). Analog versus digital: Extrapolating from electronics to neurobiology. *Neural Computation*, 10(7):1601–1638.

Sarpeshkar, R., Delbrück, T., and Mead, C. A. (1993). White noise in MOS transistors and resistors. *IEEE Circuits and Devices Magazine*, 9(6):23–29.

Schlotzhauer, K. G. and Viswanathan, T. R. (1972). New bipolar analogue multiplier. *Electronics Letters*, 8(16):425–427.

Schutte, C. and Rademeyer, P. (1992). Subthreshold 1/f noise measurements in MOS transistors aimed at optimizing focal plane array signal processing. *Analog Integrated Circuits and Signal Processing*, 2(3):171–177.

Sedra, A. and Smith, K. C. (1970). A second generation current conveyor and its applications. *IEEE Transactions on Circuit Theory*, CT-17(1):132–134.

Seevinck, E. (1981). *Analysis and Synthesis of Translinear Integrated Circuits*. D.Sc. thesis, University of Pretoria, Pretoria, South Africa.

Seevinck, E. (1988). *Analysis and Synthesis of Translinear Integrated Circuits*. Elsevier, Amsterdam.

Seevinck, E. (1990). Companding current-mode integrator: A new circuit principle for continuous-time monlithic filters. *Electronics Letters*, 26(24):2046–2047.

Seevinck, E., Wassenaar, R. F., and Wong, H. C. K. (1984). A wide-band technique for vector summation and RMS–DC conversion. *IEEE Journal of Solid-State Circuits*, SC-19(3):311–318.

Seitz, P., Leipold, D., Kramer, J., and Raynor, J. M. (1993). Smart optical and image sensors fabricated with industrial CMOS/CCD semiconductor processes. In Blouke, M. M., editor, *Charge-Coupled Devices and Solid State Optical Sensors III. Proceedings of the SPIE*, volume 1900, pages 21–30.

Serdijn, W. A., Mulder, J., Poort, P., Kouwenhoven, M., van Staveren, A., and van Roermund, A. H. M. (1999). Dynamic Translinear Circuits. In Huijsing, J., van de Plassche, R., and Sansen, W., editors, *Analog Circuit Design: Volt Electronics; Mixed-Mode Systems; Low-Noise and RF Power Amplifiers for Telecommunication*, pages 3–32. Kluwer, Boston, MA.

Serdijn, W. A., Mulder, J., van der Woerd, A. C., and van Roermund, A. H. M. (1997a). Design of wide-tunable translinear second-order oscillators. In *Proceedings of the 1997 IEEE International Symposium on Circuits and Systems*, volume 2, pages 829–832. ISCAS '97: Hong

Kong, 9–12 June.

Serdijn, W. A., Mulder, J., van der Woerd, A. C., and van Roermund, A. H. M. (1998). A wide-tunable translinear second-order oscillator. *IEEE Journal of Solid-State Circuits*, 33(2):195–201.

Serdijn, W. A., Mulder, J., and van Roermund, A. H. M. (1997b). Shortening the Analog Design Trajectory by Means of the Dynamic Translinear Principle. In *Proceedings of the ProRISC Workshop on Circuits, Systems and Signal Processing*, pages 483–489.

Serrano, T. and Linares-Barranco, B. (1995). A modular current-mode high-precision winner-take-all circuit. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 42(2):132–134.

Serrano-Gotarredona, T. and Linares-Barranco, B. (1999). Systematic width-and-length dependent CMOS transistor mismatch characterization and simulation. *Analog Integrated Circuits and Signal Processing*, 21(3):271–296.

Serrano-Gotarredona, T., Linares-Barranco, B., and Andreou, A. G. (1999). A general translinear principle for subthreshold MOS transistors. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 46(5):607–616.

Shibata, T. and Ohmi, T. (1992). A functional MOS transistor featuring gate-level weighted sum and threshold operations. *IEEE Transactions on Electron Devices*, 39(6):1444–1455.

Shockley, W. (1961). Problems related to *p-n* junctions in silicon. *Solid-State Electronics*, 2(1):35–67.

Shyu, J.-B., Temes, G. C., and Krummenacher, F. (1984). Random error effects in matched MOS capacitors and current sources. *IEEE Journal of Solid-State Circuits*, SC-19(6):948–955.

Shyu, J.-B., Temes, G. C., and Yao, K. (1982). Random errors in MOS capacitors. *IEEE Journal of Solid-State Circuits*, SC-17(6):1070–1076.

Sin, C.-K., Kramer, A., Hu, V., Chu, R. R., and Ko, P. K. (1992). EEPROM as an analog storage device, with particular applications in neural networks. *IEEE Transactions on Electron Devices*, 39(6):1410–1419.

Singh, J. (2001). *Semiconductor Devices: Basic Principles*. Wiley, New York.

Smith, K. C. and Sedra, A. (1968). The current conveyor — a new circuit bulding block. *Proceedings of the IEEE*, 56(8):1368–1369.

Smith, R. A. (1979). *Semiconductors*. Cambridge University Press, 2nd edition.

Starzyk, J. A. and Fang, X. (1993). CMOS current mode winner-take-all circuit with both excitatory and inhibitory feedback. *Electronics Letters*, 29(10):908–910.

Stork, J. M. C. and Isaac, R. D. (1983). Tunneling in base-emitter junctions. *IEEE Transactions on Electron Devices*, ED-30(11):1527–1534.

Sun, J. Y. C., Chiang, S.-Y., and Liu, M. (1998). Foundry technology for the next decade. In *International Electron Devices Meeting technical digest*, pages 321–324.

Sun, S. W. and Tsui, P. G. Y. (1994). Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation. In *1994 IEEE Custom Integrated Circuits Conference*, pages 267–270.

Suné, J., Olivo, P., and Riccò, B. (1992). Quantum-mechanical modeling of accumulation layers in MOS structure. *IEEE Transactions on Electron Devices*, 39(7):1732–1739.

Surakampontorn, W., Jutaviriya, S., and Apajinda, T. (1988). Dual translinear sinusoidal frequency doubler and full-wave rectifier. *International Journal of Electronics*, 65(6):1203–1208.

Swartzlander, Jr., E. E. and Alexopoulos, A. G. (1975). The sign/logarithm number system. *IEEE Transactions on Computers*, C-24(12):1238–1242.

Swartzlander, Jr., E. E., Chandra, D. V. S., Nagle, Jr., H. T., and Starks, S. A. (1983). Sign/logarithm arithmetic for FFT implementation. *IEEE Transactions on Computers*, C-32(6):526–534.

Sze, S. M. (1981). *Physics of Semiconductor Devices*. Wiley, New York, 2nd edition.

Takeda, E., Yang, C. Y., and Miura-Hamada, A. (1995). *Hot-Carrier Effects in MOS Devices*. Academic Press, San Diego, CA.

Tam, S., Ko, P. K., and Hu, C. (1984). Lucky-electron model of channel hot-electron injection in MOSFET's. *IEEE Transactions on Electron Devices*, 31(9):1116–1125.

Taur, Y., Buchanan, D. A., Chen, W., Frank, D. J., Ismail, K. E., Lo, S.-H., Sai-Halasz, G. A., Viswanathan, R. G., Wann, H.-J. C., Wind, S. J., and Wong, H. S. (1997). CMOS scaling into the nanometer regime. *Proceedings of the IEEE*, 85(4):486–504.

Taylor, F. J., Gill, R., Joseph, J., and Radke, J. (1988). A 20 bit logarithmic number system processor. *IEEE Transactions on Computers*, 37(2):190–200.

Teranishi, N., Kohno, A., Ishihara, Y., Oda, E., and Arai, K. (1984). An interline CCD image sensor with reduced image lag. *IEEE Transactions on Electron Devices*, ED-31(12):1829–1833.

Thanachayanont, A., Payne, A., and Pookaiyaudom, S. (1997). A current-mode phase-locked loop using a log-domain oscillator. In *Proceedings of the 1997 IEEE International Symposium on Circuits and Systems*, volume 1, pages 277–280. ISCAS '97: Hong Kong, 9–12 June.

Thanachayanont, A., Pookaiyaudom, S., and Toumazou, C. (1995). State-space synthesis of log-domain oscillators. *Electronics Letters*, 31(21):1797–1799.

Theuwissen, A. J. P. (1995). *Solid-state imaging with charge-coupled devices*. Kluwer, Dordrecht, The Netherlands.

Thompson, S. (1999). Sub 100 nm CMOS: Technology performances, trends and challenges. In *International Electron Devices Meeting (IEDM) short course, Washington D.C.*, pages 23, 26, 28 & 29.

Thompson, S., Packan, P., Ghani, T., Stettler, M., Alavi, M., Post, I., Tyagi, S., Ahmed, S., Yang, S., and Bohr, M. (1998). Source/drain extension scaling for 0.1 μm and below channel length MOSFETs. In *1998 Symposium on VLSI Technology: digest of technical papers*, pages 132–133.

Tomazou, C., Lidgey, F. J., and Haigh, D. G., editors (1990). *Analogue IC design: the current-mode approach*. Peregrinus, Stevenage, Herts., UK.

Toumazou, C., Lidgey, F. J., and Yang, M. (1989). Translinear Class AB Current Amplifier. *Electronics Letters*, 25(13):873–874.

Troutman, R. R. (1979). VLSI limitations from drain-induced barrier lowering. *IEEE Transactions on Electron Devices*, ED-26(4):461–469.

Tsividis, Y. (1996). *Mixed analog-digital VLSI devices and technology*. McGraw-Hill, New York.

Tsividis, Y. (1998). *Operation and modeling of the MOS transistor*. McGraw-Hill, New York.

Tuinhout, H. P., Elzinga, H., Brugman, J. T. H., and Postma, F. (1996). The floating gate measurement technique for characterization of capacitor matching. *IEEE Transactions on Semiconductor Manufacturing*, 9(1):2–8.

Vainio, O. and Neuvo, Y. (1986). Logarithmic arithmetic in FIR filters. *IEEE Transactions on Circuits and Systems*, CAS-33(8):826–828.

van der Gevel, M. and Kuenen, J. C. (1994). $\sqrt{x}$ circuit based on a novel, back-gate-using multiplier. *Electronics Letters*, 30(3):183–184.

Van der Tol, M. J. and Chamberlain, S. G. (1993). Drain-induced barrier lowering in buried-channel MOSFET's. *IEEE Transactions on Electron Devices*, 40(4):741–749.

van der Ziel, A. (1970). *Noise: Sources, Characterization, Measurement*, pages 171–173. Prentice-Hall.

Van Valkenburg, M. E. and Kinariwala, B. K. (1982). *Linear Circuits*, pages 68–72. Prentice-Hall, Englewood Cliffs, NJ.

Vittoz, E. (1996). Analog VLSI implementation of neural networks. In Fiesler, E. and Beale, R., editors, *Handbook of Neural Computation*, chapter E1.3. Oxford University Press and Institute of

Physics Publishing, Oxford and Bristol.

Vittoz, E. A. (1983). MOS transistors operated in the lateral bipolar mode and their application in CMOS technology. *IEEE Journal of Solid-State Circuits*, SC-18(3):273–279.

Vittoz, E. A. (1985). The design of high-performance analog circuits on digital CMOS chips. *IEEE Journal of Solid-State Circuits*, SC-20(3):657–665.

Vittoz, E. A. (1990). MOS transistor. Intensive Summer Course on CMOS VLSI Design, pages 532–539. Ecole polytechnique fédérale de Lausanne, Switzerland.

Vittoz, E. A. (1994). Micropower techniques. In Franca, J. E. and Tsividis, Y., editors, *Design of Analog-Digital VLSI Circuits for Telecommunications and Signal Processing*, chapter 3, pages 53–96. Prentice Hall, Englewood Cliffs, NJ, 2nd edition.

Vittoz, E. A. and Arreguit, X. (1993). Linear networks based on transistors. *Electronics Letters*, 29(3):297–299.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14:85–100.

Walden, R. H., Krambeck, R. H., Strain, R. J., McKenna, J., Schryer, N. L., and Smith, G. E. (1972). The buried channel charge coupled device. *The Bell System Technical Journal*, 51(7):1635–1640.

Wassenaar, R. F., Seevinck, E., van Leeuwen, M. G., Speelman, C. J., and Holle, E. (1988). New techniques for high-frequency RMS-to-DC conversion based on a multifunctional V-to-I convertor. *IEEE Journal of Solid-State Circuits*, 23(3):802–815.

Weckler, G. P. (1967). Operation of p-n junction photodetectors in a photon flux integrating mode. *IEEE Journal of Solid-State Circuits*, SC-2(3):65–73.

Weste, N. H. E. and Eshraghian, K. (1994). *Principles of CMOS VLSI design*. Addison-Wesley, Reading, MA, 2nd edition.

Wilson, B. (1990). Recent developments in current conveyors and current-mode circuits. *IEE Proceedings G: Circuits, Devices and Systems*, 137(2):63–77.

Wilson, C. S., Morris, T. G., and DeWeerth, S. P. (1999). A two-dimensional, object-based analog VLSI visual attention system. In DeWeerth, S. P., Wills, S. M., and Ishii, A. T., editors, *Proceedings of the 20th Anniversary Conference on Advanced Research in VLSI*, pages 291–308. IEEE Computer Society Press. Los Alamitos, CA.

Wolf, S. (1995). *Silicon processing for the VLSI era*, volume 3 - The submicron MOSFET. Lattice Press, Sunset Beach, CA. ISBN 0-9616721-5-3.

Wong, C. K., Wassenaar, R. F., and Seevinck, E. (1983). A wideband accurate vector-sum circuit. In *ESSCIRC'83, Ninth European Solid-State Circuits Conference Digest of Technical Papers*, pages 135–138, Lausanne, Switzerland. Presses Polytechniques Romandes.

Yamaguchi, Y., Ishibashi, A., Shimizu, M., Nishimura, T., Tsukamoto, K., Horie, K., and Akasaka, Y. (1993). A high-speed 0.6-μm 16K CMOS gate array on a thin SIMOX film. *IEEE Transactions on Electron Devices*, 40(1):179–186.

Yan, R. H., Lee, K. F., Jeon, D. Y., Kim, Y. O., Tennant, D. M., Westerwick, E. H., Chin, G. M., Morris, M. D., Early, K., and Mulgrew, P. (1992). High-performance deep-submicrometer Si MOSFET's using vertical doping engineering. *IEEE Transactions on Electron Devices*, 39(11):2639.

Yang, H., Sheu, B. J., and Lee, J.-C. (1992). A nonvolatile analog neural memory using floating-gate MOS transistors. *Analog Integrated Circuits and Signal Processing*, 2(1):19–25.

Yu, L. K. and Lewis, D. M. (1991). A 30-b integrated logarithmic number system processor. *IEEE Journal of Solid-State Circuits*, 26(10):1433–1440.

Yuille, A. L. and Geiger, D. (1995). Winner-take-all mechanisms. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 1056–1060. MIT Press, Cambridge, MA.

Index