FUNDAMENTALS OF MODERN VLSI DEVICES

YUAN TAUR TAK H. NING

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1998

Printed in the United States of America

Typeset in Times Roman 11/14 pt. and Eurostile in $\LaTeX 2_\varepsilon$ [TB]

A catalog record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Taur, Yuan, 1946–
Fundamentals of modern VLSI devices / Yuan Taur, Tak H. Ning.
p. cm.
ISBN 0-521-55056-4. – ISBN 0-521-55959-6 (pbk.)
1. Metal oxide semiconductors, Complementary. 2. Bipolar
transistors. 3. Integrated circutis – Very large scale integration.
I. Ning, Tak H.. 1943– . II. Title.
TK7871.99.M44T38 1998
621.39′5 – dc21 98-16162
CIP

ISBN 0 521 55056 4 hardback
ISBN 0 521 55959 6 paperback

1 INTRODUCTION

Since the invention of the bipolar transistor in 1947, there has been an unprecedented growth of the semiconductor industry, with an enormous impact on the way people work and live. In the last twenty years or so, by far, the strongest growth area of the semiconductor industry has been in silicon very-large-scale-integration (VLSI) technology. The sustained growth in VLSI technology is fueled by the continued shrinking of transistors to ever smaller dimensions. The benefits of miniaturization – higher packing densities, higher circuit speeds, and lower power dissipation – have been key in the evolutionary progress leading to today's computers and communication systems that offer superior performance, dramatically reduced cost per function, and much reduced physical size, in comparison with their predecessors. On the economic side, the integrated-circuit (IC) business has grown worldwide in sales from $1 billion in 1970 to $20 billion in 1984 and is projected to reach $185 billion in 1997. The electronics industry is now among the largest industries in terms of output as well as employment in many nations. The importance of microelectronics in the economic, social, and even political development throughout the world will no doubt continue to ascend. The large worldwide investment in VLSI technology constitutes a formidable driving force that will all but guarantee the continued progress in IC integration density and speed, for as long as physical principles will allow.

1.1 EVOLUTION OF VLSI DEVICE TECHNOLOGY

An excellent account of the evolution of the metal–oxide–semiconductor field-effect transistor (MOSFET), from its initial concept to VLSI applications in the mid-1980s, can be found in the paper by Sah (Sah, 1988). Figure 1.1 gives a chronology of the major milestone events in the development of VLSI technology. The bipolar transistor technology was developed early on and was applied to the first integrated-circuit memory in mainframe computers in the 1960s. Bipolar transistors have been used all along where raw circuit speed is most important, for bipolar circuits remain the fastest at the individual-circuit level. However, the large power dissipation of

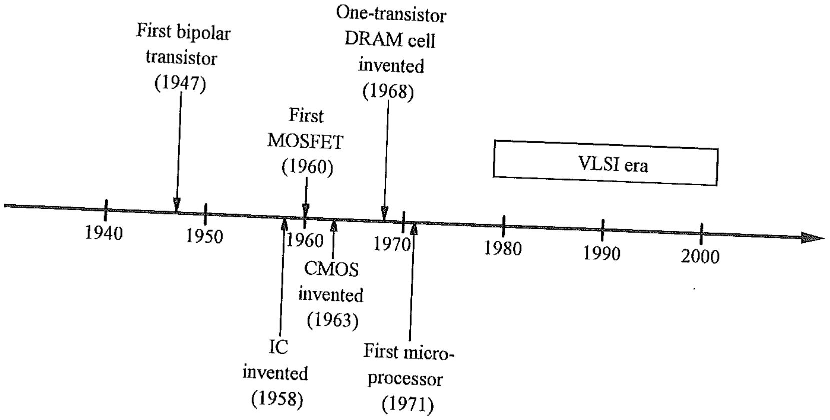

First bipolar transistor (1947)
One-transistor DRAM cell invented (1968)
First MOSFET (1960)
VLSI era
1940
1950
1960
1970
1980
1990
2000
CMOS invented (1963)
IC invented (1958)
First micro-processor (1971)

FIGURE 1.1. A brief chronology of the major milestones in the development of VLSI.

bipolar circuits has severely limited their integration level, to about 10^4 circuits per chip. This integration level is quite low by today's VLSI standard.

The idea of modulating the surface conductance of a semiconductor by the application of an electric field was first reported in 1930. However, early attempts to fabricate a surface-field-controlled device were not successful because of the presence of large densities of surface states which effectively shielded the surface potential from the influence of an external field. The first MOSFET on a silicon substrate using SiO_2 as the gate insulator was fabricated in 1960 (Kahng and Atalla, 1960). During the 1960s and 1970s, n-channel and p-channel MOSFETs were widely used, along with bipolar transistors, for implementing circuit functions on a silicon chip. Although the MOSFET devices were slow compared to the bipolar devices, they had a higher layout density and were relatively simple to fabricate; the simplest MOSFET chip could be made using only four masks and a single doping step. However, just like bipolar circuits, single-polarity MOSFET circuits suffered from large standby power dissipation, and hence were limited in the level of integration on a chip.

The major breakthrough in the level of integration came in 1963 with the invention of CMOS (complementary MOS) (Wanlass and Sah, 1963), in which both n-channel and p-channel MOSFETs are constructed simultaneously on the same substrate. A CMOS circuit typically consists of an n-channel MOSFET and a p-channel MOSFET connected in series between the power-supply terminals, so that there is negligible standby power dissipation. Significant power is dissipated only during switching of the circuit (i.e., only when the circuits are active.) By cleverly designing the "switch activities" of the circuits on a chip to minimize active power dissipation, engineers have been able to integrate hundreds of millions of CMOS transistors on a single chip and still have the chip readily air-coolable. Until recently, the integration level of CMOS was not limited by chip-level power dissipation, but by chip fabrication technology. Another advantage of CMOS circuits comes from the ratioless, full rail-to-rail logic swing, which improves the noise margin and makes a CMOS chip easier to design.

As linear dimensions reached the 0.5-μm level in the early 1990s, the performance advantage of bipolar transistors was outweighed by the significantly greater circuit density of CMOS devices. The system performance benefit of integrated functionality superseded that of raw transistor performance, and practically all the VLSI chips in production today are based on CMOS technology. Bipolar transistors are used only where raw circuit speed makes an important difference. Consequently, bipolar transistors are usually used in small-size bipolar-only chips, or in so-called BiCMOS chips where most of the functions are implemented using CMOS transistors and only a relatively small number are implemented using bipolar transistors.

Advances in lithography and etching technologies have enabled the industry to scale down transistors in physical dimensions, and to pack more transistors in the same chip area. Such progress, combined with a steady growth in chip size, resulted in an exponential growth in the number of transistors and memory bits per chip. The recent trends and future projections in these areas are illustrated in Fig. 1.2. Dynamic random-access memories (DRAMs) have characteristically contained the highest component count of any IC chips. This has been so because of the small size of the one-transistor memory cell (Dennard, 1968) and because of the large and often insatiable demand for more memory in computing systems. It is interesting to note that the entire content of this book can be stored in one 64-Mb DRAM chip, which is in volume production in 1997 and has an area equivalent to a square of about 1.2 × 1.2 cm^2.

One remarkable feature of silicon devices that fuels the rapid growth of the information technology industry is that their speed increases and their cost decreases

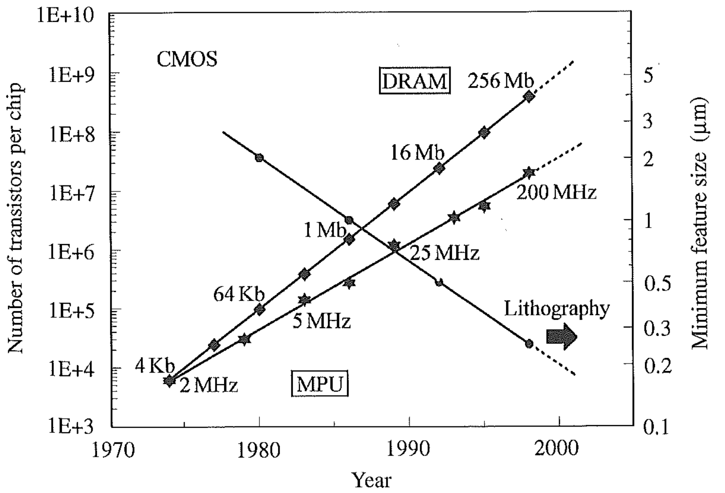


FIGURE 1.2. Trends in lithographic feature size, and number of transistors per chip for DRAM and microprocessor chips.

as their size is reduced. The transistors manufactured today are 20 times faster and occupy less than 1% of the area of those built 20 years ago. This is illustrated in the trend of microprocessor units (MPUs) in Fig. 1.2. The increase in the clock frequency of microprocessors is the result of a combination of improvements in microprocessor architecture and improvements in transistor speed.

1.2 MODERN VLSI DEVICES

It is clear from Fig. 1.2 that modern transistors of practical interest have feature sizes of 0.5 μm and smaller. Although the basic operation principles of large and small transistors are the same, the relative importance of the various device parameters and performance factors for the small-dimension modern transistors is quite different from that for the transistors of the early 1980s or earlier. It is our intention to focus our discussion in this book on the fundamentals of silicon devices of sub-0.5-μm generations.

1.2.1 MODERN CMOS TRANSISTORS

A schematic cross section of modern CMOS transistors, consisting of an n-channel MOSFET and a p-channel MOSFET integrated on the same chip, is shown in Fig. 1.3. A generic process flow for fabricating the CMOS transistors is outlined in Appendix 1. The key physical features of the modern CMOS technology, as illustrated in Fig. 1.3, include: p-type polysilicon gate for the p-channel MOSFET and n-type polysilicon gate for the n-channel MOSFET, refractory metal silicide on the polysilicon gate as well as on the source and drain diffusion regions, and shallow-trench oxide isolation.

In the electrical design of the modern CMOS transistor, the power-supply voltage is reduced with the physical dimensions in some coordinated manner. A great deal of design detail goes into determining the channel length, or separation between the source and drain, accurately, maximizing the on current of the transistor while

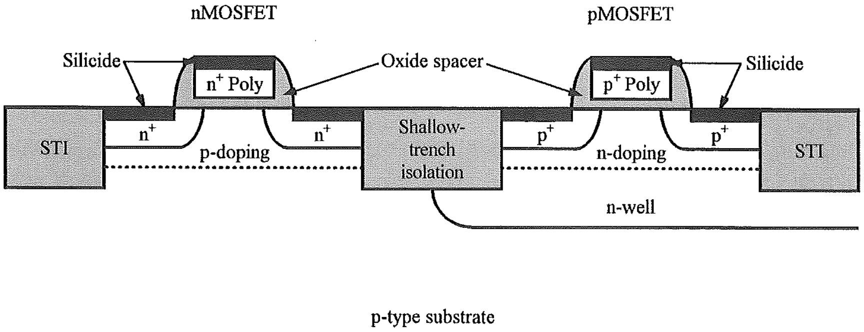


FIGURE 1.3. Schematic device cross section for an advanced CMOS technology.

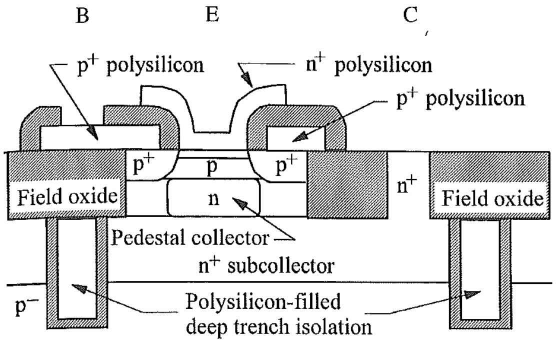


FIGURE 1.4. Schematic cross section of a modern n–p–n bipolar transistor.

maintaining an adequately low off current, minimizing variation of the transistor characteristics with process tolerances, and minimizing the parasitic resistances and parasitic capacitances.

1.2.2 MODERN BIPOLAR TRANSISTORS

The schematic cross section of a modern silicon bipolar transistor is shown in Fig. 1.4. The process outline for fabricating this transistor is shown in Appendix 2. The salient features of this transistor include: shallow-trench field oxide and deep-trench isolation, polysilicon emitter, polysilicon base contact which is self-aligned to the emitter contact, and a pedestal collector which is doped to the desired level only directly underneath the emitter.

Unlike CMOS, the power-supply voltage for a bipolar transistor is usually kept constant as the transistor physical dimensions are reduced. Without the ability to reduce the operating voltage, electrical breakdown is a severe concern in the design of modern bipolar transistors. In designing a modern bipolar transistor, a lot of effort is spent tailoring the doping profile of the various device regions in order to maintain adequate breakdown-voltage margins while maximizing the device performance. At the same time, unlike the bipolar transistors before the early 1980s when the device performance was mostly limited by the device physical dimensions practical at the time, a modern bipolar transistor often has its performance limited by its current-density capability and not by its physical dimensions. Attempts to improve the current-density capability of a transistor usually lead to reduced breakdown voltages.

1.3 SCOPE AND BRIEF DESCRIPTION OF THE BOOK

In writing this book, it is our goal to address the factors governing the performance of modern VLSI devices in depth. This is carried out by first discussing the design of the various device parameters individually, and then discussing the relative importance of the individual device parameters in determining the performance of small-dimension modern transistors. A substantial part of the book is devoted to in-depth discussions on the subtle tradeoffs in the design of modern CMOS and bipolar transistors.

This book contains sufficient background tutorials to be used as a textbook for students taking a graduate or advanced undergraduate course in microelectronics. The prerequisite will be one semester of either solid-state physics or semiconductor physics. For the practicing engineer, this book provides an extensive source of reference material that covers the fundamentals of CMOS and bipolar technology, devices, and circuits. It should be useful to VLSI process engineers interested in learning basic device principles, and to device design or characterization engineers who desire more in-depth knowledge in their specialized areas. Below is a brief description of each chapter.

CHAPTER 2: BASIC DEVICE PHYSICS

Chapter 2 is devoted to introducing just the right level of basic device physics to make the book self-contained, and to prepare the reader with the necessary background on device operation and material physics to follow the discussion in the rest of the book. It is assumed that the reader has the basic knowledge equivalent to one semester of upper-level undergraduate or graduate-level solid-state physics or semiconductor physics. From there, the device concepts or device physics needed to understand the subsequent chapters are introduced.

Starting with the energy bands in silicon, Chapter 2 first introduces the basic concepts of Fermi level, carrier concentration, drift and diffusion current transport, and Poisson's equation. The next two sections focus on the most elementary building blocks of silicon devices: the p–n junction and the MOS capacitor. Basic knowledge of their characteristics is a prerequisite to further understand the operation of the VLSI devices they lead into: bipolar and MOSFET transistors. The rest of Chapter 2 covers high-field effects, Si–SiO_2 systems, hot carriers, and the physics of tunneling and breakdown relevant to VLSI device reliability. The purpose here is to introduce to the reader the basic physical concepts needed to follow the discussion in this book. The details on the recent advances in each of these areas are beyond the scope of this book.

CHAPTER 3: MOSFET DEVICES

Chapter 3 describes the basic characteristics of MOSFET devices, using the n-channel MOSFET as an example for most of the discussions. It is divided into two parts. The first part deals with the more elementary long-channel MOSFETs, including subsections on drain current models, I–V characteristics, subthreshold currents, channel mobility, and intrinsic capacitances. These serve as a foundation for understanding the more important but more complex short-channel MOSFETs, which have lower capacitances and carry higher currents per gate voltage swing. The second part of Chapter 3 covers the specific features of short-channel MOSFETs important for device design purposes. The subsections include short-channel effects, velocity saturation and overshoot, channel-length modulation, and source–drain series resistance.

CHAPTER 4: CMOS DEVICE DESIGN

Chapter 4 considers the major device design issues in a CMOS technology. It begins with the concept of MOSFET scaling – the most important guiding principle for achieving density, speed, and power improvements in VLSI evolution. Several non-scaling factors are addressed, notably, the thermal voltage and the silicon bandgap, which have significant implications on the deviation of the CMOS evolution path from ideal scaling. Two key CMOS device design parameters – threshold voltage and channel length – are then discussed in detail. Subsections on threshold voltage include off-current requirement, nonuniform channel doping, gate work-function effects, channel profile design, and quantum-mechanical and discrete dopant effects on threshold voltage. Subsections on channel length include the definition of effective channel length, its extraction by the conventional method and the shift-and-ratio method, and the physical interpretation of effective channel length.

CHAPTER 5: CMOS PERFORMANCE FACTORS

Chapter 5 examines the key factors that govern the switching performance and power dissipation of basic digital CMOS circuits which form the building blocks of a VLSI chip. Starting with a brief description of static CMOS logic gates and their layout, we examine the parasitic resistances and capacitances that may adversely affect the delay of a CMOS circuit. These include source and drain series resistance, junction capacitance, overlap capacitance, gate resistance, and interconnect capacitance and resistance. Next, we formulate a delay equation and use it to study the sensitivity of CMOS delay performance to a variety of device and circuit parameters such as wire loading, device width and length, gate oxide thickness, power-supply voltage, threshold voltage, parasitic components, and substrate sensitivity in stacked circuits. The last section of Chapter 5 further extends the delay equation to project the performance factors of several advanced CMOS materials and device structures. These include SOI CMOS, high-mobility Si–SiGe CMOS, and low-temperature CMOS. The unique advantage of each approach is discussed in depth.

CHAPTER 6: BIPOLAR DEVICES

The basic components of a bipolar transistor are described in Chapter 6. The discussion is based entirely on the vertical n–p–n transistor, since practically all high-speed bipolar transistors used in digital circuits are of the vertical n–p–n type. However, the basic device operation concept and device physics can be readily extended to other types of bipolar transistors, such as p–n–p bipolar transistors and lateral bipolar transistors.

The basic operation of a bipolar transistor is described in terms of two p–n diodes connected back to back. The basic theory of a p–n diode is modified and applied to derive the current equations for a bipolar transistor. From these current equations, other important device parameters and phenomena, such as current gain, Early voltage, base–collector junction avalanche, emitter–collector punch-through, base widening, and diffusion capacitance, are examined. Finally, the basic equivalent-circuit models relating the device parameters to circuit parameters are developed. These equivalent-circuit models form the starting point for discussing the performance of a bipolar transistor in circuit applications.

CHAPTER 7: BIPOLAR DEVICE DESIGN

Chapter 7 covers the basic design of a bipolar transistor. The design of the individual device regions, namely the emitter, the base, and the collector, are discussed separately. Since the detailed characteristics of a bipolar transistor depend on its operating point, the focus of this chapter is on optimizing the device design according to its intended operating condition and environment, and on the trade-offs that must be made in the optimization process. The sections include an examination of the effect of grading the base doping profile to enhance the drift field in the intrinsic base, and a derivation of the collector-current equations when there is significant heavy doping effect in the base and when Ge is used to engineer the intrinsic-base bandgap. The chapter concludes with a discussion of the salient features of the most commonly used modern bipolar device structure.

CHAPTER 8: BIPOLAR PERFORMANCE FACTORS

The major factors governing the performance of bipolar transistors in circuit applications are discussed in Chapter 8. Several of the commonly used figures of merit, namely, cutoff frequency, maximum oscillation frequency, and logic gate delay, are examined, and how a bipolar transistor can be optimized for a given figure of merit is discussed. Sections are devoted to examining the important delay components of a logic gate, and how these components can be minimized. The power–delay trade-offs in the design of a bipolar transistor under various circuit loading conditions are also examined. Finally, the scaling properties of bipolar transistors, and how the large standby power dissipation of bipolar circuits limits the integration level of bipolar circuit chips, are discussed.

3 MOSFET DEVICES

The metal–oxide–semiconductor field-effect transistor (MOSFET) is the building block of VLSI circuits in microprocessors and dynamic memories. Because the current in a MOSFET is transported predominantly by carriers of one polarity only (e.g., electrons in an n-channel device), the MOSFET is usually referred to as a unipolar or majority-carrier device. Throughout this chapter, n-channel MOSFETs are used as an example to illustrate device operation and derive drain-current equations. The results can easily be extended to p-channel MOSFETs by exchanging the dopant types and reversing the voltage polarities.

The basic structure of a MOSFET is shown in Fig. 3.1. It is a four-terminal device with the terminals designated as *gate* (subscript g), *source* (subscript s), *drain* (subscript d), and *substrate* or *body* (subscript b). An n-channel MOSFET, or nMOSFET, consists of a p-type silicon substrate into which two n^+ regions, the source and the drain, are formed (e.g., by ion implantation). The gate electrode is usually made of metal or heavily doped polysilicon and is separated from the substrate by a thin silicon dioxide film, the *gate oxide*. The gate oxide is usually formed by thermal oxidation of silicon. In VLSI circuits, a MOSFET is surrounded by a thick oxide called the *field oxide* to isolate it from the adjacent devices. The surface region under the gate oxide between the source and drain is called the *channel* region and is critical for current conduction in a MOSFET. The basic operation of a MOSFET device can be easily understood from the MOS capacitor discussed in Section 2.3. When there is no voltage applied to the gate or when the gate voltage is zero, the p-type silicon surface is either in accumulation or in depletion and there is no current flow between the source and drain. The MOSFET device acts like two back-to-back p–n junction diodes with only low-level leakage currents present. When a sufficiently large positive voltage is applied to the gate, the silicon surface is inverted to n-type, which forms a conducting channel between the n^+ source and drain. If there is a voltage difference between them, an electron current will flow from the source to the drain. A MOSFET device therefore operates like a switch ideally suited for digital circuits. ***Since the gate electrode is electrically insulated from the substrate, there is effectively no dc gate current, and the channel is capacitively***

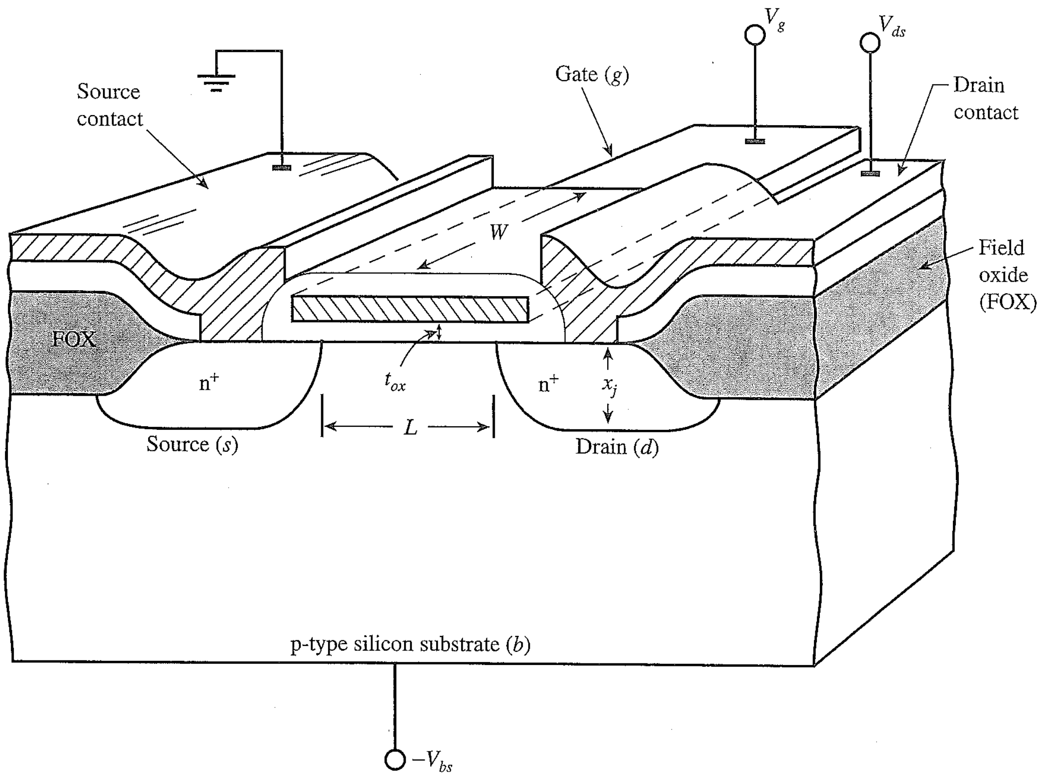


FIGURE 3.1. Three-dimensional view of basic MOSFET device structure. (After Arora, 1993.)

coupled to the gate via the electric field in the oxide (hence the name *field-effect transistor*).

3.1 LONG-CHANNEL MOSFETs

This section describes the basic characteristics of a long-channel MOSFET, which will serve as the foundation for understanding the more important but more complex short-channel MOSFETs in Section 3.2. First, a general MOSFET current model based on the *gradual channel approximation* (GCA) is formulated in Section 3.1.1. The GCA is valid for most regions of MOSFET operation except beyond the pinch-off or saturation point. A *charge-sheet approximation* is then introduced in Section 3.1.2 to obtain analytical expressions for the source–drain current in the linear and saturation regions. Current characteristics in the subthreshold region are discussed in Section 3.1.3. Section 3.1.4 addresses the threshold-voltage dependence on substrate bias and temperature. Section 3.1.5 presents an empirical model for electron and hole mobilities in a MOSFET channel. Lastly, intrinsic MOSFET capacitances and inversion-layer capacitance effects (neglected in the charge sheet approximation) are covered in Section 3.1.6.

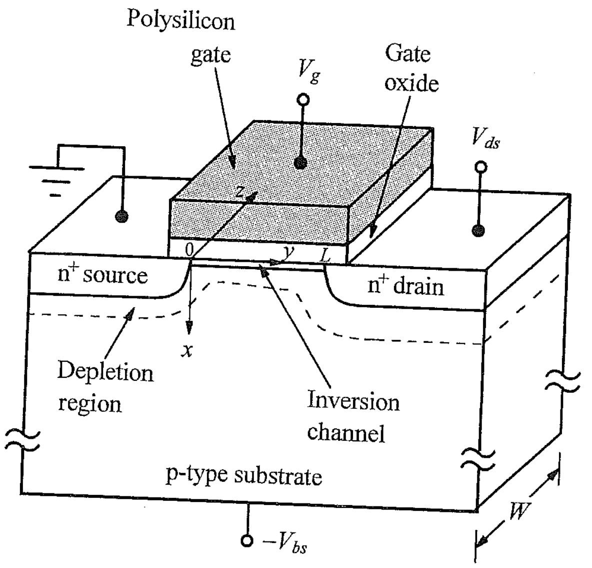


FIGURE 3.2. A schematic MOSFET cross section, showing the axes of coordinates and the bias voltages at the four terminals for the drain-current model.

3.1.1 DRAIN-CURRENT MODEL

In this subsection, we formulate a general drain-current model for a long-channel MOSFET. The model will then be simplified using a *charge-sheet approximation* in the next subsection, leading to an analytical expression for the source–drain current. Figure 3.2 shows the schematic cross section of an n-channel MOSFET in which the source is the n^+ region on the left, and the drain is the n^+ region on the right. A thin oxide film separates the gate from the channel region between the source and drain. We choose an x–y coordinate system consistent with Section 2.3 on MOS capacitors, namely, the x-axis is perpendicular to the gate electrode and is pointing into the p-type substrate with $x=0$ at the silicon surface. The y-axis is parallel to the channel or the current flow direction, with $y=0$ at the source and $y=L$ at the drain. L is called the *channel length* and is a key parameter in a MOSFET device. The MOSFET is assumed to be uniform along the z-axis over a distance called the *channel width*, W, determined by the boundaries of the thick field oxide.

Conventionally, the source voltage is defined as the ground potential. The drain voltage is V_{ds}, the gate voltage is V_g, and the p-type substrate is biased at $-V_{bs}$. Initially, we assume $V_{bs}=0$, i.e., the substrate contact is grounded to the source potential. Later on, we will discuss the effect of substrate bias on MOSFET characteristics. The p-type substrate is assumed to be uniformly doped with an acceptor concentration N_a.

As defined in Section 2.3, $\psi(x, y)$ is the band bending, or intrinsic potential, at (x, y) with respect to the bulk intrinsic potential. We further assume that $V(y)$ is the electron quasi-Fermi potential at a point y along the channel with respect to the Fermi potential of the n^+ source. Since there is no bias between the source

3.1
and
subs
sam
non
esse
x in
$V(y$
3.1
OF (
Fro
give
Foll
elec
The
whi
whi
3.1
On
app
y-d
the
to r
The

and the substrate, the Fermi potential of the source is the same as that of the bulk substrate. Therefore, ***in terms of the channel-to-substrate diode, V(y) plays the same role as the reverse bias V_R in Section 2.3.5 on MOS capacitors under nonequilibrium***. As was discussed in Section 2.2.3, the quasi-Fermi potential stays essentially constant across the depletion region, i.e.,$V(y)$ does not change with x in the direction perpendicular to the surface. At the drain end of the channel, $V(y = L) = V_{ds}$.

3.1.1.1 INVERSION CHARGE DENSITY AS A FUNCTION OF QUASI-FERMI POTENTIAL

From Eq. (2.150) and Eq. (2.187), the electron concentration at any point (x, y) is given by

$$n(x, y) = \frac{n_i^2}{N_a} e^{q(\psi - V)/kT}. \tag{3.1}$$

Following the same approach as in Section 2.3.2, one obtains an expression for the electric field similar to that of Eq. (2.153):

$$\mathscr{E}^2(x, y) = \left(\frac{d\psi}{dx}\right)^2 = \frac{2kTN_a}{\varepsilon_{si}}\left[\left(e^{-q\psi/kT} + \frac{q\psi}{kT} - 1\right) + \frac{n_i^2}{N_a^2}\left(e^{-qV/kT}\left(e^{q\psi/kT} - 1\right) - \frac{q\psi}{kT}\right)\right]. \tag{3.2}$$

The condition for surface inversion, Eq. (2.190), becomes

$$\psi(0, y) = V(y) + 2\psi_B, \tag{3.3}$$

which is a function of y. From Eq. (2.191), the maximum depletion layer width is

$$W_{dm}(y) = \sqrt{\frac{2\varepsilon_{si}[V(y) + 2\psi_B]}{qN_a}}, \tag{3.4}$$

which is also a function of y.

3.1.1.2 GRADUAL-CHANNEL APPROXIMATION

One of the key assumptions in any 1-D MOSFET model is the ***gradual channel approximation (GCA), which assumes that the variation of the electric field in the y-direction (along the channel) is much less than the corresponding variation in the x-direction (perpendicular to the channel)*** (Pao and Sah, 1966). This allows us to reduce Poisson's equation to the 1-D form (x-component only) as in Eq. (2.147). The GCA is valid for most of the channel regions except beyond the *pinch-off* point, which will be discussed later. One further assumes that both the hole current and the generation and recombination current are negligible, so that the current continuity equation can be applied to the electron current in the y-direction. In other words, the total drain-to-source current I_{ds} is the same at any point along the channel. From Eq. (2.45), the electron current density at a point (x, y) is

$$J_n(x, y) = -q\mu_n n(x, y)\frac{dV(y)}{dy}, \tag{3.5}$$

where $n(x, y)$ is the electron density, and μ_n is the electron mobility in the channel. The carrier mobility in the channel is generally much lower than the mobility in the bulk, due to additional surface scattering mechanisms, as will be addressed in Section 3.1.5. ***With V(y) defined as the quasi-Fermi potential, i.e., playing the role of ϕ_n in Eq. (2.45), Eq. (3.5) includes both the drift and diffusion currents***. The total current at a point y along the channel is obtained by multiplying Eq. (3.5) with the channel width W and integrating over the depth of the inversion layer. The integration is carried out from $x = 0$ to x_i, the bottom of the inversion layer where $\psi = \psi_B$:

$$I_{ds}(y) = qW\int_0^{x_i} \mu_n n(x, y)\frac{dV}{dy}\,dx. \tag{3.6}$$

There is a sign change, as we define $I_{ds} > 0$ to be the drain-to-source current in the $-y$ direction. Since V is a function of y only, dV/dy can be taken outside the integral. We also assume that μ_n can be taken outside the integral by defining an *effective mobility*, μ_{eff}, at some average gate and drain fields. What remains in the integral is the electron concentration, $n(x, y)$. Its integration over the inversion layer gives the inversion charge per unit gate area, Q_i:

$$Q_i(y) = -q\int_0^{x_i} n(x, y)\,dx. \tag{3.7}$$

Equation (3.6) then becomes

$$I_{ds}(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(y) = -\mu_{eff}W\frac{dV}{dy}Q_i(V). \tag{3.8}$$

In the last step, Q_i is expressed as a function of V; V is interchangeable with y, since V is a function of y only. Multiplying both sides of Eq. (3.8) by dy and integrating from 0 to L (source to drain) yield

$$\int_0^L I_{ds}\,dy = \mu_{eff}W\int_0^{V_{ds}} [-Q_i(V)]\,dV. \tag{3.9}$$

Current continuity requires that I_{ds} be a constant, independent of y. Therefore, the drain-to-source current is

$$I_{ds} = \mu_{eff}\frac{W}{L}\int_0^{V_{ds}}[-Q_i(V)]\,dV. \tag{3.10}$$

3.1.1.3 PAO AND SAH'S DOUBLE INTEGRAL

An alternative form of $Q_i(V)$ can be derived if $n(x, y)$ is expressed as a function of (ψ, V) using Eq. (3.1), i.e.,

$$n(x, y) = n(\psi, V) = \frac{n_i^2}{N_a}e^{q(\psi-V)/kT}, \tag{3.11}$$

and substituted into Eq. (3.7):

$$\begin{aligned} Q_i(V) &= -q\int_{\psi_s}^{\psi_B} n(\psi, V)\frac{dx}{d\psi}\,d\psi \\ &= -q\int_{\psi_B}^{\psi_s}\frac{(n_i^2/N_a)e^{q(\psi-V)/kT}}{\mathscr{E}(\psi, V)}\,d\psi. \end{aligned} \tag{3.12}$$

Here, ψ_s is the surface potential at $x = 0$ and $\mathscr{E}(\psi, V) = -d\psi/dx$ is given by the square root of Eq. (3.2). Substituting Eq. (3.12) into Eq. (3.10) yields

$$I_{ds} = q\mu_{eff}\frac{W}{L}\int_0^{V_{ds}}\left(\int_{\psi_B}^{\psi_s}\frac{(n_i^2/N_a)e^{q(\psi-V)/kT}}{\mathscr{E}(\psi, V)}\,d\psi\right)dV. \tag{3.13}$$

This is referred to as *Pao and Sah's double integral* (Pao and Sah, 1966). The boundary value ψ_s is determined by two coupled equations: Eq. (2.180) and $Q_s = -\varepsilon_{si}\mathscr{E}_s(\psi_s)$ or Gauss's law, where $\mathscr{E}_s(\psi_s)$ is obtained by letting $\psi = \psi_s$ in Eq. (3.2). In inversion, only two of the terms in Eq. (3.2) are significant and need to be kept. The merged equation is then

$$\begin{aligned} V_g &= V_{fb} + \psi_s - \frac{Q_s}{C_{ox}} \\ &= V_{fb} + \psi_s + \frac{\sqrt{2\varepsilon_{si}kTN_a}}{C_{ox}}\left[\frac{q\psi_s}{kT} + \frac{n_i^2}{N_a^2}e^{q(\psi_s-V)/kT}\right]^{1/2}, \end{aligned} \tag{3.14}$$

which is an implicit equation for $\psi_s(V)$. Equations (3.14) and (3.13) can only be solved numerically.

3.1.2 MOSFET *I–V* CHARACTERISTICS

In this subsection, we derive the basic expressions for long-channel current in the *linear* and *saturation* regions.

3.1.2.1 CHARGE-SHEET APPROXIMATION

In order to derive an analytical solution for the drain current, we simplify the general model using the charge-sheet approximation (Brews, 1978) in which the inversion-layer thickness is treated as zero. ***It assumes that all the inversion charges are located at the silicon surface like a sheet of charge and that there is no potential drop or band bending across the inversion layer***. Furthermore, the depletion approximation is applied to the bulk depletion region. After the onset of inversion, the surface potential is pinned at $\psi_s = 2\psi_B + V(y)$, as indicated by Eq. (3.3). From Eq. (3.4), the bulk depletion charge density is

$$Q_d = -qN_aW_{dm} = -\sqrt{2\varepsilon_{si}qN_a(2\psi_B + V)}. \tag{3.15}$$

The total charge density in the silicon is given by Eq. (2.180),

$$Q_s = -C_{ox}(V_g - V_{fb} - \psi_s) = -C_{ox}(V_g - V_{fb} - 2\psi_B - V). \tag{3.16}$$

The inversion charge density is then the difference of the above two equations,

$$\begin{aligned} Q_i &= Q_s - Q_d \\ &= -C_{ox}(V_g - V_{fb} - 2\psi_B - V) + \sqrt{2\varepsilon_{si}qN_a(2\psi_B + V)}. \end{aligned} \tag{3.17}$$

Substituting Eq. (3.17) into Eq. (3.10) and carrying out the integration, we obtain the drain current as a function of the gate and drain voltages:

$$\begin{aligned} I_{ds} = \mu_{eff}C_{ox}\frac{W}{L}\Bigg[&\left(V_g - V_{fb} - 2\psi_B - \frac{V_{ds}}{2}\right)V_{ds} \\ &- \frac{2\sqrt{2\varepsilon_{si}qN_a}}{3C_{ox}}\left[(2\psi_B + V_{ds})^{3/2} - (2\psi_B)^{3/2}\right]\Bigg]. \end{aligned} \tag{3.18}$$

Equation (3.18) represents the basic I–V characteristics of a MOSFET device based on the charge-sheet model. It indicates that, for a given V_g, the drain current I_{ds} first increases linearly with the drain voltage V_{ds} (called the *linear* or *triode* region), then gradually levels off to a saturated value (*saturation* region). These two distinct regions are further examined below.

3.1.2.2 CHARACTERISTICS IN THE LINEAR (TRIODE) REGION

When V_{ds} is small, one can expand Eq. (3.18) into a power series in V_{ds} and keep only the lowest-order (first-order) terms:

$$\begin{aligned} I_{ds} &= \mu_{eff}C_{ox}\frac{W}{L}\left(V_g - V_{fb} - 2\psi_B - \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}}\right)V_{ds} \\ &= \mu_{eff}C_{ox}\frac{W}{L}(V_g - V_t)V_{ds}, \end{aligned} \tag{3.19}$$

where V_t is the *threshold voltage* given by

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si} q N_a \psi_B}}{C_{ox}}. \tag{3.20}$$

Comparing this equation with Eq. (2.175) and Eq. (2.180), one can see that ***V_t is simply the gate voltage when the surface potential or band bending reaches $2\psi_B$ and the silicon charge (the square root) is equal to the bulk depletion charge for that potential***. As a reminder, $2\psi_B = (2kT/q)\ln(N_a/n_i)$, which is typically 0.6–0.9 V. When V_g is below V_t, there is very little current flow and the MOSFET is said to be in the *subthreshold* region, to be discussed in Section 3.1.3. Equation (3.19) indicates that, ***in the linear region, the MOSFET simply acts like a resistor with a sheet resistivity, $\rho_{sh} = 1/[\mu_{eff} C_{ox}(V_g - V_t)]$, modulated by the gate voltage***. The threshold voltage V_t can be determined by plotting I_{ds} versus V_g at low drain voltages, as shown in Fig. 3.3. The extrapolated intercept of the linear portion of the $I_{ds}(V_g)$ curve with the V_g-axis gives the approximate value of V_t. In reality, such a *linearly extrapolated threshold voltage* (V_{on}) is slightly higher than the "$2\psi_B$" V_t due to inversion-layer capacitance and other effects, as will be addressed in Section 3.1.6. Notice that the $I_{ds}(V_g)$ curve is not linear near the threshold voltage. This is because the charge-sheet approximation, on which Eq. (3.19) is based, is no longer valid in that regime. Low-drain $I_{ds}(V_g)$ curves are also used to extract the *effective channel length* of a MOSFET, as will be discussed in Chapter 4.

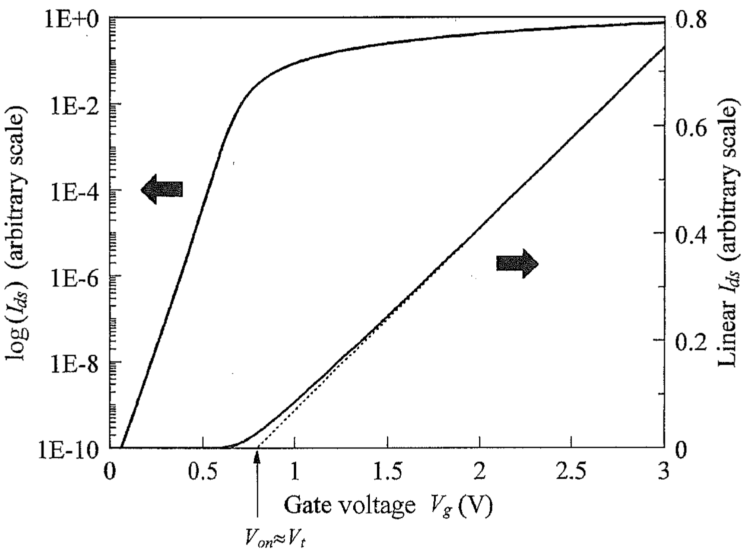


FIGURE 3.3. Typical MOSFET I_{ds}–V_g characteristics at low drain bias voltages. The same current is plotted on both linear and logarithmic scales. The dotted line illustrates the determination of the linearly extrapolated threshold voltage, V_{on}.

3.1.2.3 CHARACTERISTICS IN THE SATURATION REGION

For larger values of V_{ds}, the second-order terms in the power series expansion of Eq. (3.18) are also important and must be kept. A good approximation to the drain current is then

$$I_{ds} = \mu_{eff} C_{ox}\frac{W}{L}\left((V_g - V_t)V_{ds} - \frac{m}{2}V_{ds}^2\right), \qquad (3.21)$$

where

$$m = 1 + \frac{\sqrt{\varepsilon_{si} q N_a/4\psi_B}}{C_{ox}} = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}} \qquad (3.22)$$

is the *body-effect coefficient.* Here m typically lies between 1.1 and 1.4 and is related to the *body effect* to be discussed in Section 3.1.4. Equation (3.22) shows several alternative expressions for m. The one in terms of the capacitance ratio follows from Eq. (2.174), where C_{dm} is the bulk depletion capacitance at $\psi_s = 2\psi_B$. Alternatively, m can be expressed in terms of a thickness ratio, since $C_{dm} = \varepsilon_{si}/W_{dm}$, $C_{ox} = \varepsilon_{ox}/t_{ox}$, and $\varepsilon_{si}/\varepsilon_{ox} \approx 3$. The threshold voltage, given by Eq. (3.20), can be expressed in terms of m as $V_t = V_{fb} + (2m - 1)2\psi_B$. ***As V_{ds} increases, I_{ds} follows a parabolic curve, as shown in Fig. 3.4, until a maximum or saturation value is reached.*** This occurs when $V_{ds} = V_{dsat} = (V_g - V_t)/m$, at which

$$I_{ds} = I_{dsat} = \mu_{eff}\, C_{ox}\frac{W}{L}\frac{(V_g - V_t)^2}{2m}. \qquad (3.23)$$

Equation (3.23) reduces to the well-known expression for the MOSFET

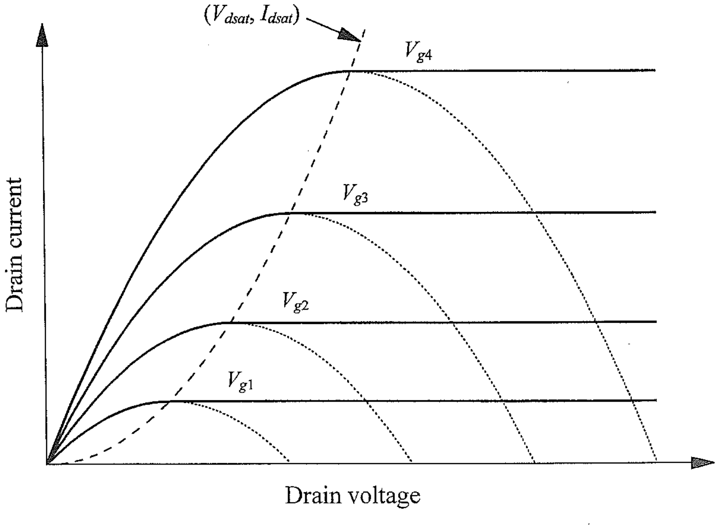


FIGURE 3.4. Long-channel MOSFET I_{ds}–V_{ds} characteristics (solid curves) for several different values of V_g. The dashed curve shows the trajectory of drain voltage beyond which the current saturates. The dotted curves help to illustrate the parabolic behavior of the characteristics before saturation.

nsion of
he drain

(3.21)

(3.22)

4 and is
!) shows
acitance
tance at
io, since
given by
. *As V_{ds}*
imum or
at which

(3.23)

IOSFET

saturation current when the bulk depletion charge is neglected (valid for low substrate doping) so $m = 1$. The dashed curve in Fig. 3.4 shows the trajectory of V_{dsat} through the various I_{ds}–V_{ds} curves for different V_g. Equation (3.21), or Eq. (3.18), is valid only for $V_{ds} \le V_{dsat}$. ***Beyond V_{dsat}, I_{ds} stays constant at I_{dsat}, independent of V_{ds}.***

3.1.2.4 THE ONSET OF PINCH-OFF AND CURRENT SATURATION

The saturation of drain current can be understood from the inversion charge density, Eq. (3.17). For $V \le 2\psi_B$, one can expand the square-root term of Eq. (3.17) into a power series in V and keep only the two lowest terms,

$$Q_i(V) = -C_{ox}(V_g - V_t - mV). \tag{3.24}$$

$Q_i(V)$ is plotted in Fig. 3.5. Equation (3.10) states that the drain current is proportional to the area under the $-Q_i(V)$ curve between $V = 0$ and V_{ds}. When V_{ds} is small (linear region), the inversion charge density at the drain end of the channel is only slightly lower than that at the source end. As the drain voltage increases (for a fixed gate voltage), the current increases, but the inversion charge density at the drain decreases until finally it goes to zero when $V_{ds} = V_{dsat} = (V_g - V_t)/m$. At this point, I_{ds} reaches its maximum value. In other words, ***the surface channel vanishes at the drain end of the channel when saturation occurs***. This is called *pinch-off* and is illustrated in Fig. 3.6. When V_{ds} increases beyond saturation, the pinch-off point moves toward the source, but the drain current remains essentially the same. This is because for $V_{ds} > V_{dsat}$, the voltage at the pinch-off point remains

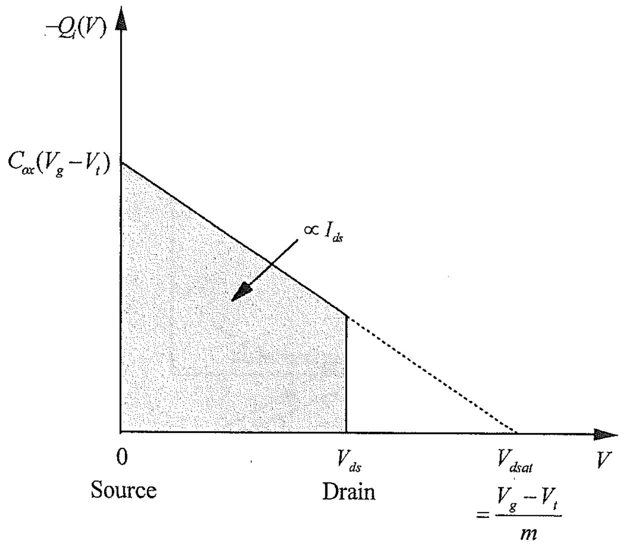


FIGURE 3.5. Inversion charge density as a function of the quasi-Fermi potential of a point in the channel. Before saturation, the drain current is proportional to the shaded area integrated from zero to the drain voltage.

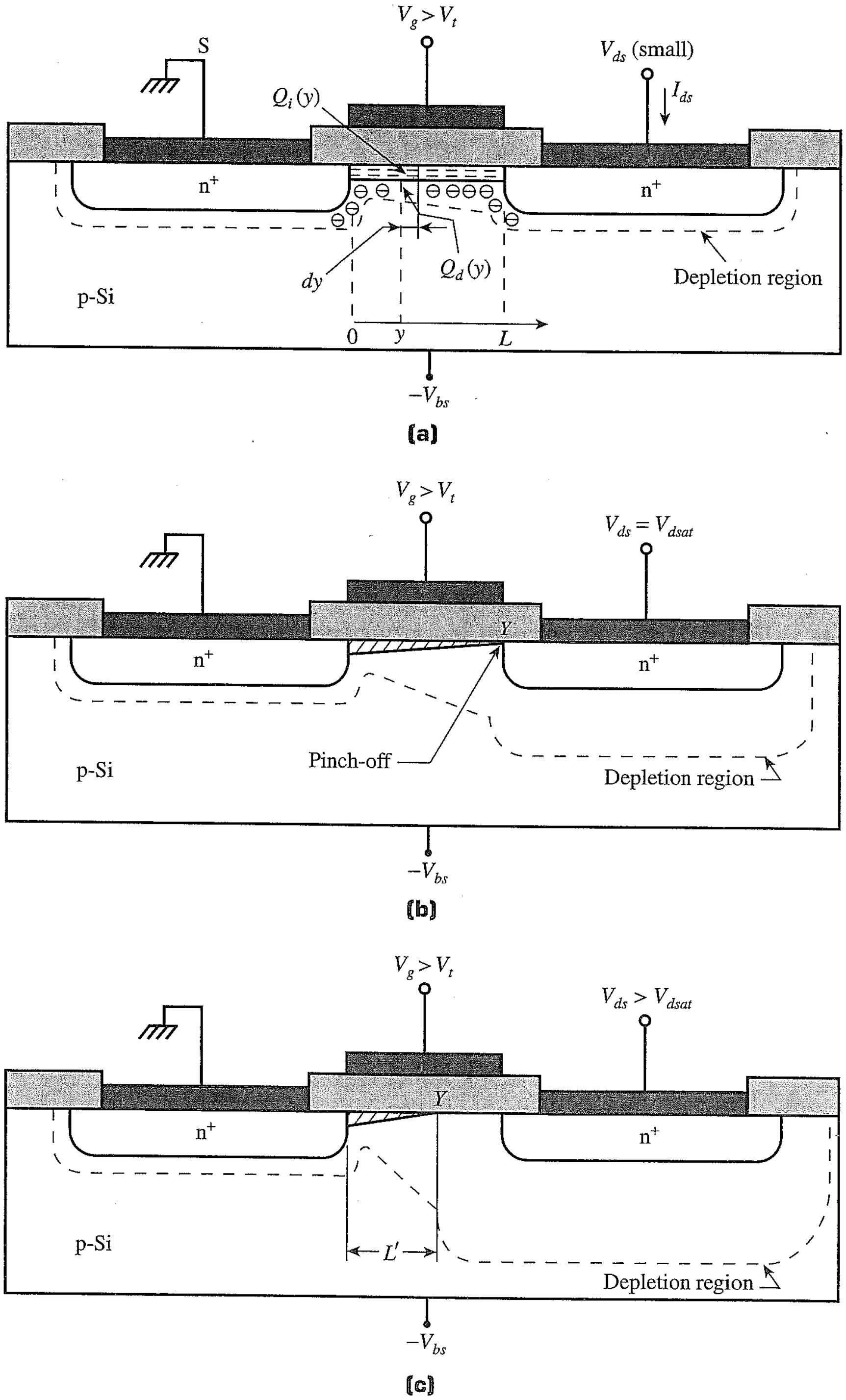


FIGURE 3.6. (a) MOSFET operated in the linear region (low drain voltage). (b) MOSFET operated at the onset of saturation. The pinch-off point is indicated by Y. (c) MOSFET operated beyond saturation where the channel length is reduced to L'. (After Sze, 1981.)

at V_{ds}

stays
This
assoc
Fu
amin
Eq. (

Su

Both
value
As V

$V($

$\frac{V_g - }{m}$

FIGU
and th
ration
charg
parab

at V_{dsat} and the current, given by

$$\int_0^{L'} I_{ds}\,dy = \mu_{eff} W \int_0^{V_{dsat}} [-Q_i(V)]\,dV, \tag{3.25}$$

stays the same apart from a slight decrease in L (to L'), as shown in Fig. 3.6. This phenomenon is called *channel length modulation* and will be discussed in association with short-channel MOSFETs in Section 3.2.

Further insight into the MOSFET behavior at pinch-off can be gained by examining the function $V(y)$. Integrating from 0 to y after multiplying both sides of Eq. (3.8) by dy yields

$$\begin{aligned} I_{ds}y &= \mu_{eff} W \int_0^{V} [-Q_i(V)]\,dV \\ &= \mu_{eff} C_{ox} W \left((V_g - V_t)V - \frac{m}{2}V^2 \right). \end{aligned} \tag{3.26}$$

Substituting I_{ds} from Eq. (3.21) into Eq. (3.26), one can solve for $V(y)$:

$$V(y) = \frac{V_g - V_t}{m} - \sqrt{\left(\frac{V_g - V_t}{m}\right)^2 - 2\frac{y}{L}\left(\frac{V_g - V_t}{m}\right)V_{ds} + \frac{y}{L}V_{ds}^2}. \tag{3.27}$$

Both $V(y)$ and $-Q_i/mC_{ox} = (V_g - V_t)/m - V(y)$ are plotted in Fig. 3.7 for several values of V_{ds}. At low V_{ds}, $V(y)$ varies almost linearly between the source and drain. As V_{ds} increases, the inversion charge density at the drain decreases due to the

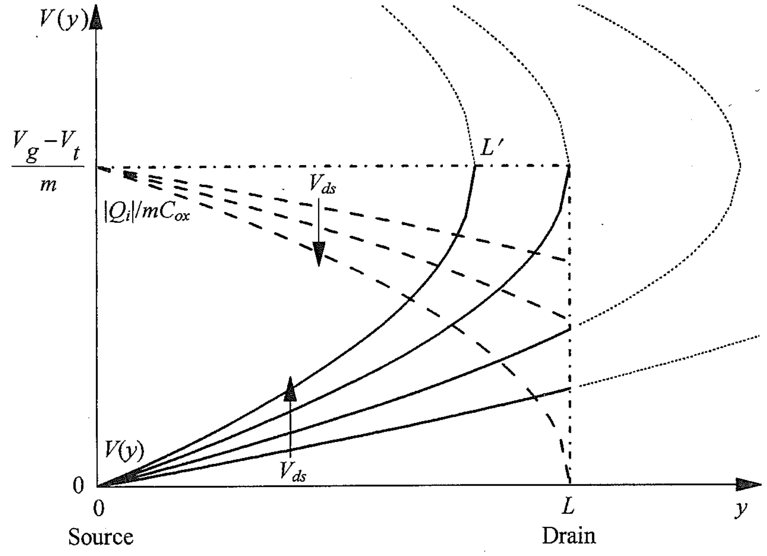


FIGURE 3.7. Quasi-Fermi potential versus distance between the source and the drain for several V_{ds}-values from the linear region to beyond saturation. The dashed curves show the corresponding variation of inversion charge density along the channel. The dotted curves help visualize the parabolic behavior of the characteristics.

lowering of the electron quasi-Fermi level. This is accompanied by a corresponding increase of dV/dy to maintain current continuity. When V_{ds} reaches $V_{dsat} = (V_g - V_t)/m$, we have $Q_i(y = L) = 0$ and $V(y)$ exhibits a singularity at the drain, where $dV/dy = \infty$. This implies that ***the electric field in the y-direction changes more rapidly than the field in the x-direction and the gradual channel approximation breaks down***. In other words, beyond the pinch-off point, carriers are no longer confined to the surface channel, and a 2-D Poisson's equation must be solved for carrier injection from the pinch-off point into the drain depletion region (El-Mansy and Boothroyd, 1977).

Strictly speaking, if $V_{ds} > 2\psi_B$, neither Eq. (3.17) nor Eq. (3.18) can be expanded into a power series in V_{ds}. A more general form of the saturation voltage is obtained by letting $Q_i = 0$ in Eq. (3.17) and solving for $V = V_{dsat}$ [equivalent to solving $dI_{ds}/dV_{ds} = 0$ by differentiating Eq. (3.18)]:

$$V_{dsat} = V_g - V_{fb} - 2\psi_B + \frac{\varepsilon_{si}qN_a}{C_{ox}^2} - \sqrt{\frac{2\varepsilon_{si}qN_a}{C_{ox}^2}\left(V_g - V_{fb} + \frac{\varepsilon_{si}qN_a}{2C_{ox}^2}\right)}. \tag{3.28}$$

The corresponding saturation current can be found by substituting Eq. (3.28) for V_{ds} in Eq. (3.18). The mathematics is rather tedious (Brews, 1981). A few selected curves are plotted in Fig. 3.8 and compared with those calculated from Eq. (3.21). It turns out that Eq. (3.21) serves as a good approximation to the drain current over a much wider range of voltages than expected. Even for a drain voltage several times greater than $2\psi_B$, the current is only slightly ($\approx$5%) underestimated.

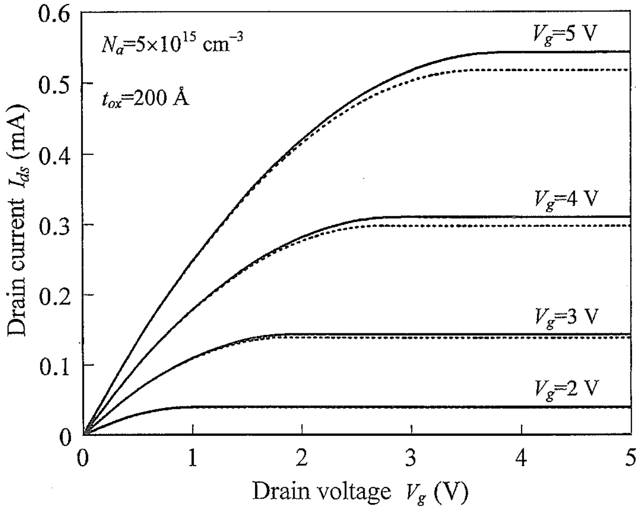
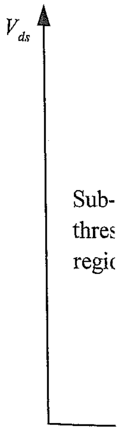


FIGURE 3.8. I_{ds}–V_{ds} curves calculated from the full equation (3.18) (solid curves), compared with the parabolic approximation (3.21) (dotted curves).

3.1.2.5
So far we
eration ai
that it is
(cf. Fig. :
For exam
gate and
from the
Since
other terr
that all th
voltage i
voltage c
V_{dd}, in c
ground p
required.
configura

3.1.3

Dependi
in one of
istics hav
the chara
In Fig. 3
ately bel
drain cur
This is b



sponding
$= (V_g -$
in, where
ges more
ximation
ιο longer
olved for
El-Mansy

ιn be ex-
voltage is
ivalent to

(3.28)

3.28) for
selected
q. (3.21).
rent over
e several
d.

3.1.2.5 pMOSFET *I–V* CHARACTERISTICS

So far we have used an n-channel device as an example to discuss MOSFET operation and I–V characteristics. A p-channel MOSFET operates similarly, except that it is fabricated inside an n-well with implanted p^+ source and drain regions (cf. Fig. 3.2), and that the polarities of all the voltages and currents are reversed. For example, I_{ds}–V_{ds} characteristics for a pMOSFET (cf. Fig. 3.8) have negative gate and drain voltages with respect to the source terminal for a hole current to flow from the source to the drain.

Since the source of a pMOSFET is at the highest potential compared with the other terminals, it is usually connected to the power supply V_{dd} in a CMOS circuit so that all the voltages are positive (or zero). In that case, the device conducts if the gate voltage is lower than $V_{dd} - V_t$, where V_t (>0) is the magnitude of the threshold voltage of the pMOSFET. The ohmic contact to the n-well is also connected to V_{dd}, in contrast to an nMOSFET, where the p-type substrate is usually tied to the ground potential. This leaves the n-well-to-p-substrate junction reverse biased as required. More about nMOSFET and pMOSFET bias conditions in a CMOS circuit configuration will be given in Section 5.1.

3.1.3 SUBTHRESHOLD CHARACTERISTICS

Depending on the gate and source–drain voltages, a MOSFET device can be biased in one of the three regions shown in Fig. 3.9. Linear and saturation region characteristics have been described in the previous subsection. In this subsection, we discuss the characteristics of a MOSFET device in the subthreshold region where $V_g < V_t$. In Fig. 3.3, the drain current on a linear scale appears to approach zero immediately below the threshold voltage. On a logarithmic scale, however, the descending drain current remains at nonnegligible levels for several tenths of a volt below V_t. This is because the inversion charge density does not drop to zero abruptly. Rather,

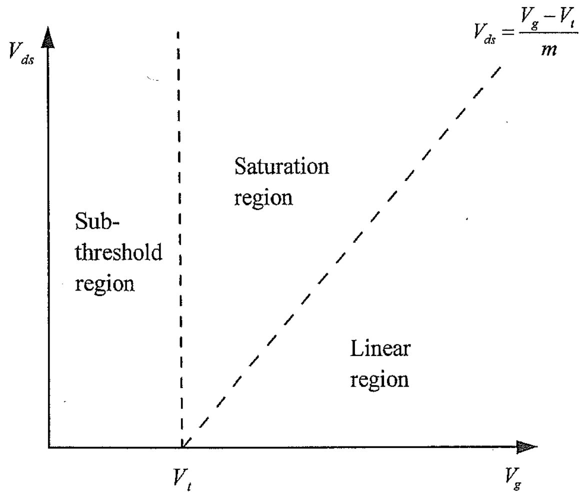


FIGURE 3.9. Three regions of MOSFET operation in the V_{ds}–V_g plane.

it follows an exponential dependence on ψ_s or V_g, as is evident from Eq. (3.11). Subthreshold behavior is of particular importance in low-voltage, low-power applications, such as in digital logic and memory circuits, because it describes how a MOSFET device switches off. The subthreshold region immediately below V_t, in which $\psi_B \le \psi_s \le 2\psi_B$, is also called the *weak inversion* region.

3.1.3.1 DRIFT AND DIFFUSION COMPONENTS OF DRAIN CURRENT

Unlike the strong inversion region, in which the drift current dominates, subthreshold conduction is dominated by the diffusion current. Both current components are included in Pao and Sah's double integral, Eq. (3.13). In general, current continuity only applies to the total current, not to its individual components. In other words, the fractional ratio between the drift and the diffusion components may vary from one point of the channel to another. At low drain bias voltages, however, it is possible to separate the drift and diffusion components using the implicit $\psi_s(V)$ relation, Eq. (3.14). When $qV/kT \ll 1$, only the first-order terms of V need to be kept. In Eq. (3.8), $Q_i(V)$ can be replaced by its zeroth-order value, $Q_i(V=0)$; hence V must vary linearly from the source to the drain, as required by current continuity. Since the total current is proportional to dV/dy and the drift current is proportional to the electric field or $d\psi_s/dy$, the drift fraction of the current is given by the change of surface potential (band bending) with respect to the quasi-Fermi potential, i.e., $d\psi_s/dV$. This can be evaluated from Eq. (3.14) in the limit of $V \to 0$:

$$\frac{d\psi_s}{dV} = \frac{(n_i^2/N_a^2)e^{q\psi_s/kT}}{1+(n_i^2/N_a^2)e^{q\psi_s/kT}+(C_{ox}^2/\varepsilon_{si}qN_a)(|Q_s|/C_{ox})}, \tag{3.29}$$

where $|Q_s|/C_{ox}$ is the voltage drop across the oxide given by the last term of Eq. (3.14). It is clear that in weak inversion where $\psi_B < \psi_s < 2\psi_B$, the numerator is much less than unity and the diffusion component dominates. Conversely, beyond strong inversion, $d\psi_s/dV \approx 1$ and the drift current dominates. These kinds of behavior are further illustrated in Fig. 3.10.

3.1.3.2 SUBTHRESHOLD CURRENT EXPRESSION

To find an expression for the subthreshold current, we start with Eq. (3.2) and apply Gauss's law to obtain the total charge density in silicon,

$$-Q_s = \varepsilon_{si}\mathscr{E}_s = \sqrt{2\varepsilon_{si}kTN_a}\left[\frac{q\psi_s}{kT}+\frac{n_i^2}{N_a^2}e^{q(\psi_s-V)/kT}\right]^{1/2}, \tag{3.30}$$

where only the two significant terms are kept in the square bracket. In weak inversion, the second term in the bracket arising from the inversion charge density is

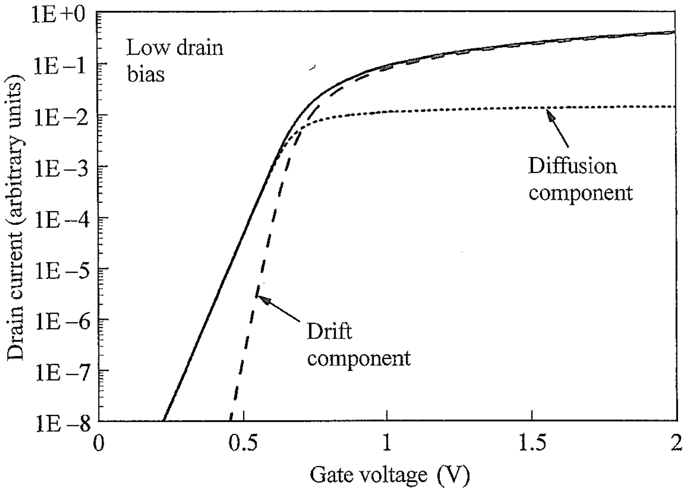


FIGURE 3.10. Drift and diffusion components of current in an I_{ds}–V_g plot. Their sum is the total current represented by the solid curve.

much less than the first term from the depletion charge density. Equation (3.30) can then be expanded into a power series: the zeroth-order term is the depletion charge density $-Q_d$, and the first-order term gives the inversion charge density,

$$-Q_i = \sqrt{\frac{\varepsilon_{si} q N_a}{2\psi_s}} \left(\frac{kT}{q}\right)\left(\frac{n_i}{N_a}\right)^2 e^{q(\psi_s - V)/kT}. \tag{3.31}$$

The surface potential ψ_s is related to the gate voltage through Eq. (3.14). Since the inversion charge density is small, ψ_s can be considered as a function of V_g only, independent of V. This also means that the electric field along the channel direction is small; hence the drift current is negligible.

Substituting Q_i into Eq. (3.10) and carrying out the integration, we obtain the drain current in the subthreshold region:

$$I_{ds} = \mu_{eff} \frac{W}{L} \sqrt{\frac{\varepsilon_{si} q N_a}{2\psi_s}} \left(\frac{kT}{q}\right)^2 \left(\frac{n_i}{N_a}\right)^2 e^{q\psi_s/kT}(1 - e^{-qV_{ds}/kT}). \tag{3.32}$$

ψ_s can be expressed in terms of V_g using Eq. (3.14), where only the depletion charge term needs to be kept:

$$V_g = V_{fb} + \psi_s + \frac{\sqrt{2\varepsilon_{si} q N_a \psi_s}}{C_{ox}}. \tag{3.33}$$

It is straightforward to solve a quadratic equation for ψ_s. To further simplify the result, we consider ψ_s as only slightly deviated from the threshold value, $2\psi_B$ (Swanson and Meindl, 1972). In other words, we assume that $|\psi_s - 2\psi_B| \ll 2\psi_B$ and expand the square-root term in Eq. (3.33) around $\psi_s = 2\psi_B$:

$$V_g = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}} + \left(1 + \frac{\sqrt{\varepsilon_{si}qN_a/4\psi_B}}{C_{ox}}\right)(\psi_s - 2\psi_B). \tag{3.34}$$

Using Eq. (3.20) and Eq. (3.22) for V_t and m, one can rewrite Eq. (3.34) as $V_g = V_t + m(\psi_s - 2\psi_B)$. Solving for ψ_s and substituting it into Eq. (3.32) yield the subthreshold current as a function of V_g:

$$I_{ds} = \mu_{eff}\frac{W}{L}\sqrt{\frac{\varepsilon_{si}qN_a}{4\psi_B}}\left(\frac{kT}{q}\right)^2 e^{q(V_g-V_t)/mkT}(1 - e^{-qV_{ds}/kT}), \tag{3.35}$$

or

$$I_{ds} = \mu_{eff}\, C_{ox}\frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_g-V_t)/mkT}(1 - e^{-qV_{ds}/kT}). \tag{3.36}$$

3.1.3.3 SUBTHRESHOLD SLOPE

The subthreshold current is independent of the drain voltage once V_{ds} is larger than a few kT/q, as would be expected for diffusion-dominated current transport. The dependence on gate voltage, on the other hand, is exponential with a *subthreshold slope* (Fig. 3.10),

$$S = \left(\frac{d(\log_{10} I_{ds})}{dV_g}\right)^{-1} = 2.3\frac{mkT}{q} = 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{ox}}\right), \tag{3.37}$$

of typically 70–100 mV/decade. Here $m = 1 + (C_{dm}/C_{ox})$ from Eq. (3.22). If the Si–SiO$_2$ interface trap density is high, the subthreshold slope may be more graded than that given by Eq. (3.37), since the capacitance associated with the interface trap is in parallel with the depletion-layer capacitance C_{dm}. It should be noted that for ψ_s substantially below $2\psi_B$, e.g., when V_g is a few tenths of a volt below V_t, Eq. (3.37) tends to underestimate the subthreshold slope by 5–10%. As a result, the subthreshold current can be 2 to 4 times higher than that given by Eq. (3.36). For VLSI circuits, a steep subthreshold slope is desirable for the ease of switching the transistor current off. However, ***except for a slight dependence on bulk doping concentration through C_{dm}, the subthreshold slope is rather insensitive to device***

parameters. It is only a function of temperature. This has significant implications on device scaling as will be discussed in Chapter 4.

3.1.4 SUBSTRATE BIAS AND TEMPERATURE DEPENDENCE OF THRESHOLD VOLTAGE

The threshold voltage is one of the key parameters of a MOSFET device. In this subsection, we examine the dependence of threshold voltage on substrate bias and temperature.

3.1.4.1 SUBSTRATE SENSITIVITY (BODY EFFECT)

The drain-current equation in Section 3.1.2 was derived assuming zero substrate bias (V_{bs}). If $V_{bs} \neq 0$, one can modify the previously discussed MOSFET equations by considering that applying $-V_{bs}$ to the substrate is equivalent to raising all other voltages (namely, gate, source, and drain voltages) by $+V_{bs}$ while keeping the substrate grounded. This is shown in Fig. 3.11.

Using the charge-sheet model as before, Eq. (3.17) becomes

$$Q_i = -C_{ox}(V_g + V_{bs} - V_{fb} - 2\psi_B - V) + \sqrt{2\varepsilon_{si} q N_a (2\psi_B + V)}, \qquad (3.38)$$

where V is the reverse bias voltage between a point in the channel and the substrate.

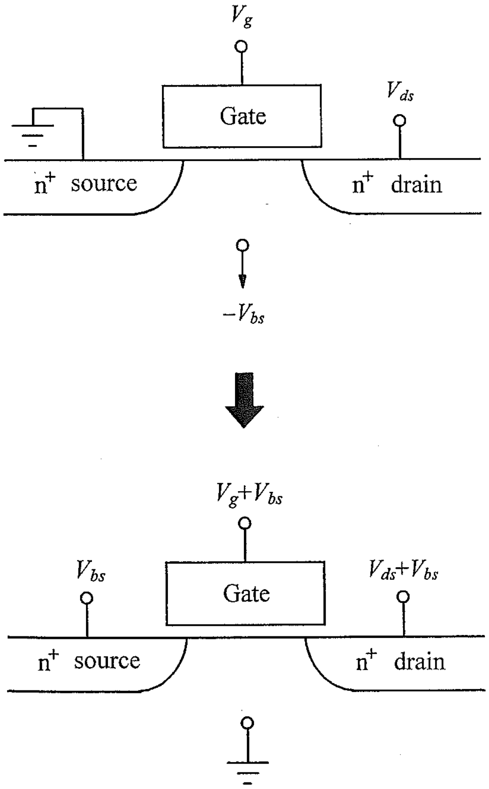


FIGURE 3.11. Equivalent circuits used to evaluate the effect of substrate bias on MOSFET I–V characteristics.

The current is obtained by integrating Q_i from V_{bs} (source) to $V_{bs} + V_{ds}$ (drain):

$$I_{ds} = \mu_{eff} C_{ox} \frac{W}{L} \left[\left(V_g - V_{fb} - 2\psi_B - \frac{V_{ds}}{2} \right) V_{ds} - \frac{2\sqrt{2\varepsilon_{si} q N_a}}{3C_{ox}} \left[(2\psi_B + V_{bs} + V_{ds})^{3/2} - (2\psi_B + V_{bs})^{3/2} \right] \right]. \tag{3.39}$$

At low drain voltages (linear region), the current is still given by Eq. (3.19), except that the threshold voltage is now

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B + V_{bs})}}{C_{ox}}. \tag{3.40}$$

It can be seen from Eq. (3.40) that ***the effect of (reverse) substrate bias is to widen the bulk depletion region and raise the threshold voltage***. Figure 3.12 plots V_t as a function of V_{bs}. The slope of the curve,

$$\frac{dV_t}{dV_{bs}} = \frac{\sqrt{\varepsilon_{si} q N_a / 2(2\psi_B + V_{bs})}}{C_{ox}}, \tag{3.41}$$

is referred to as the *substrate sensitivity*. At $V_{bs} = 0$, the slope equals C_{dm}/C_{ox}, or $m - 1$ [Eq. (3.22)]. The substrate sensitivity is higher for a higher bulk doping concentration. It is clear from Fig. 3.12 that the substrate sensitivity decreases as the substrate (reverse) bias voltage increases. From Eq. (3.37), a (reverse) substrate bias also makes the subthreshold slope slightly steeper, since it widens the depletion region and lowers C_{dm}.

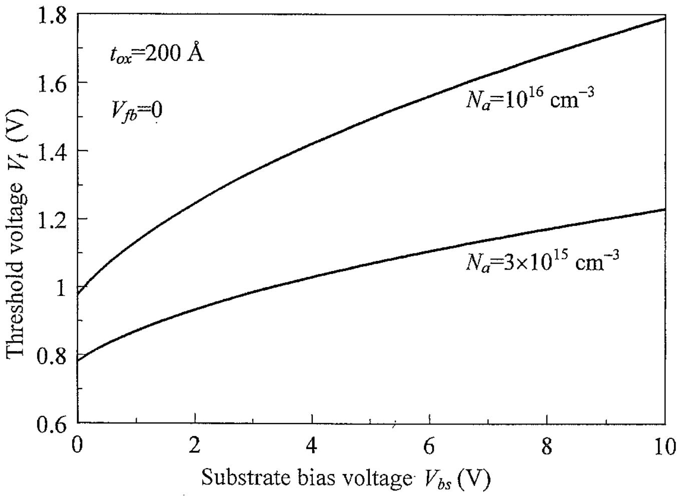


FIGURE 3.12. Threshold-voltage variation with reverse substrate bias for two uniform substrate doping concentrations.

3.1.4.2 TEMPERATURE DEPENDENCE OF THRESHOLD VOLTAGE

Next, we examine the temperature dependence of the threshold voltage. The flat-band voltage of an nMOSFET with n$^+$ polysilicon gate is $V_{fb} = -E_g/2q - \psi_B$ [Eq. (2.181)], assuming there is no oxide charge. Substituting it into Eq. (3.20) yields the threshold voltage,

$$V_t = -\frac{E_g}{2q} + \psi_B + \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}}, \tag{3.42}$$

at zero substrate bias. The temperature dependence of V_t is related to the temperature dependence of E_g and ψ_B:

$$\begin{aligned}\frac{dV_t}{dT} &= -\frac{1}{2q}\frac{dE_g}{dT} + \left(1 + \frac{\sqrt{\varepsilon_{si}qN_a/\psi_B}}{C_{ox}}\right)\frac{d\psi_B}{dT} \\ &= -\frac{1}{2q}\frac{dE_g}{dT} + (2m-1)\frac{d\psi_B}{dT}.\end{aligned} \tag{3.43}$$

$d\psi_B/dT$ stems from the temperature dependence of the intrinsic carrier concentration, which can be evaluated using Eq. (2.37) and Eq. (2.7):

$$\begin{aligned}\frac{d\psi_B}{dT} &= \frac{d}{dT}\left[\frac{kT}{q}\ln\left(\frac{N_a}{\sqrt{N_cN_v}\,e^{-E_g/2kT}}\right)\right] \\ &= -\frac{k}{q}\ln\left(\frac{\sqrt{N_cN_v}}{N_a}\right) - \frac{kT}{q\sqrt{N_cN_v}}\frac{d\sqrt{N_cN_v}}{dT} + \frac{1}{2q}\frac{dE_g}{dT}.\end{aligned} \tag{3.44}$$

Since both N_c and N_v are proportional to $T^{3/2}$, we have $d(N_cN_v)^{1/2}/dT = \frac{3}{2}(N_cN_v)^{1/2}/T$. Substituting Eq. (3.44) into Eq. (3.43) yields

$$\frac{dV_t}{dT} = -(2m-1)\frac{k}{q}\left[\ln\left(\frac{\sqrt{N_cN_v}}{N_a}\right) + \frac{3}{2}\right] + \frac{m-1}{q}\frac{dE_g}{dT}. \tag{3.45}$$

From Section 2.1.1 and Table 2.1, $dE_g/dT \approx -2.7 \times 10^{-4}$ eV/K and $(N_cN_v)^{1/2} \approx 2.4 \times 10^{19}$ cm^{-3}. For $N_a \sim 10^{16}$ cm^{-3} and $m \approx 1.1$, ***dV_t/dT is typically −1 mV/K.*** Note that the temperature coefficient decreases slightly as N_a increases: for $N_a \sim 10^{18}$ cm^{-3} and $m \approx 1.3$, dV_t/dT is about −0.7 mV/K. These numbers imply that, at an elevated temperature of, for example, 100°C, the threshold voltage is 55–75 mV lower than at room temperature. Since digital VLSI circuits often operate at elevated temperatures due to heat generation, this effect, plus the degradation of subthreshold slope with temperature, causes the leakage current at $V_g = 0$ to increase considerably over its room-temperature value. Typically, the off-state leakage current of a MOSFET at 100°C is 30–50 times larger than the leakage current at 25°C. These are important design considerations, to be addressed in detail in Chapter 4.

3.1.5 MOSFET CHANNEL MOBILITY

The carrier mobility in a MOSFET channel is significantly lower than that in bulk silicon, due to additional scattering mechanisms. Lattice or phonon scattering is aggravated by the presence of crystalline discontinuity at the surface boundary, and surface roughness scattering severely degrades mobility at high normal fields. Channel mobility is also affected by processing conditions that alter the Si–SiO$_2$ interface properties (e.g., oxide charge and interface traps, as discussed in Section 2.3.6).

3.1.5.1 EFFECTIVE MOBILITY AND EFFECTIVE NORMAL FIELD

In Section 3.1.1, the channel mobility was treated as a constant by defining an effective mobility as

$$\mu_{eff} = \frac{\int_0^{x_i} \mu_n n(x)\,dx}{\int_0^{x_i} n(x)\,dx}, \tag{3.46}$$

which is essentially an average value weighted by the carrier concentration in the inversion layer. Empirically, it has been found that ***when μ_{eff} is plotted against an effective normal field $\mathscr{E}_{eff}$, there exists a universal relationship independent of the substrate bias, doping concentration, and gate oxide thickness*** (Sabnis and Clemens, 1979). The effective normal field is defined as the average electric field perpendicular to the Si–SiO$_2$ interface experienced by the carriers in the channel. Using Gauss's law, one can express $\mathscr{E}_{eff}$ in terms of the depletion and inversion charge densities:

$$\mathscr{E}_{eff} = \frac{1}{\varepsilon_{si}}\left(|Q_d| + \frac{1}{2}|Q_i|\right), \tag{3.47}$$

where $|Q_d| + \frac{1}{2}|Q_i|$ is the total silicon charge inside a Gaussian surface through the middle of the inversion layer. Using Eq. (2.161) and Eq. (3.20), the depletion charge can be expressed as

$$|Q_d| = \sqrt{4\varepsilon_{si} q N_a \psi_B} = C_{ox}(V_t - V_{fb} - 2\psi_B). \tag{3.48}$$

Substituting this expression and $|Q_i| \approx C_{ox}(V_g - V_t)$ into Eq. (3.47) yields

$$\mathscr{E}_{eff} = \frac{V_t - V_{fb} - 2\psi_B}{3t_{ox}} + \frac{V_g - V_t}{6t_{ox}}, \tag{3.49}$$

where $C_{ox} = \varepsilon_{ox}/t_{ox}$ and $\varepsilon_{si} \approx 3\varepsilon_{ox}$ were used. Equation (3.49) can further be simplified if the gate electrode is n$^+$ polysilicon (for nMOSFETs) such that $V_{fb} = -E_g/2q - \psi_B$. For submicron CMOS technologies, $\psi_B = 0.30$–0.42 V. Therefore, the effective normal field can be expressed in terms of explicit device parameters as

$$\mathscr{E}_{eff} = \frac{V_t + 0.2}{3t_{ox}} + \frac{V_g - V_t}{6t_{ox}}. \tag{3.50}$$

Equation (3.50) is valid for low drain voltages. At high drain voltages, Q_i decreases toward the drain end of the channel. To estimate the average effective field in that case, the second term in Eq. (3.50) should be reduced accordingly.

3.1.5.2 ELECTRON MOBILITY DATA

A typical set of data on mobility versus effective normal field for nMOSFETs is shown in Fig. 3.13 (Takagi *et al.*, 1988). At room temperature, the mobility follows a $\mathscr{E}_{eff}^{-1/3}$ dependence below 5 × 10^5 V/cm. A simple, approximate expression for this case is (Baccarani and Wordeman, 1983)

$$\mu_{eff} \approx 32500 \times \mathscr{E}_{eff}^{-1/3}. \tag{3.51}$$

Beyond $\mathscr{E}_{eff} = 5 \times 10^5$ V/cm, μ_{eff} decreases much more rapidly with increasing $\mathscr{E}_{eff}$ because of increased surface roughness scattering as carriers are distributed closer to the surface under high normal fields. For each doping concentration, there exists an effective field below which the mobility falls off the universal curve. This is

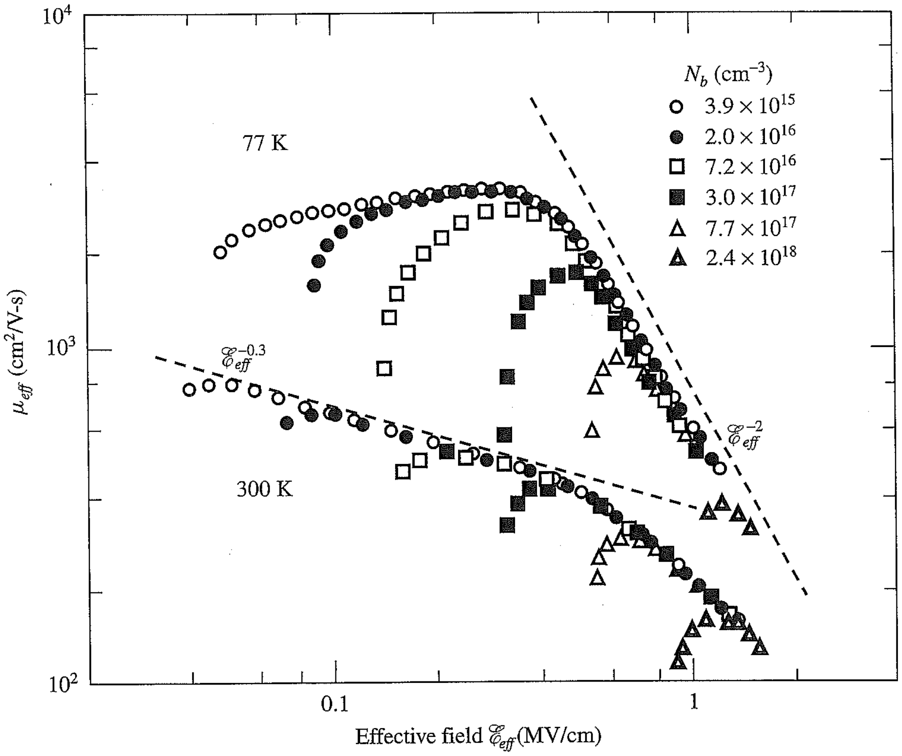


FIGURE 3.13. Measured electron mobility at 300 and 77 K versus effective normal field for several substrate doping concentrations. (After Takagi *et al.*, 1988).

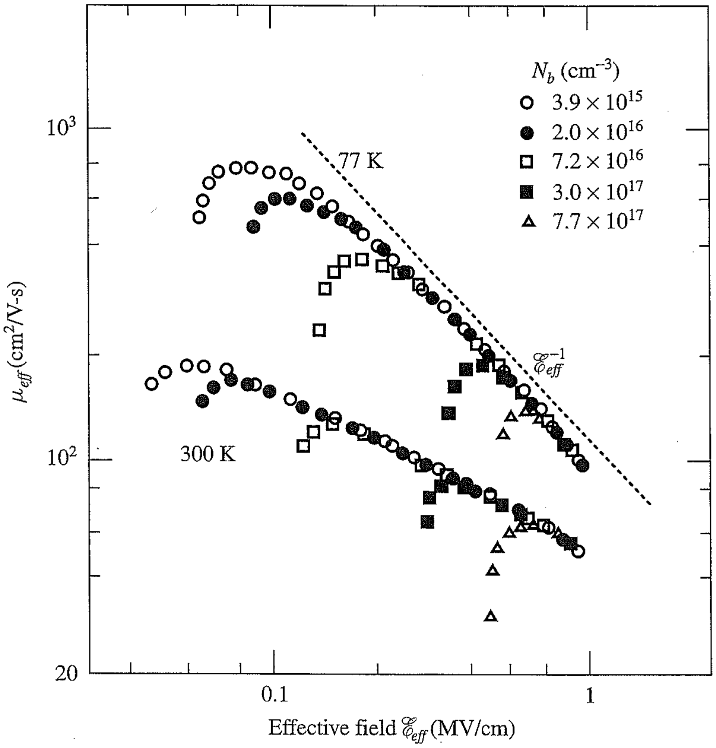


FIGURE 3.14. Measured hole mobility at 300 K and 77 K versus effective normal field (with a factor $\frac{1}{3}$) for several substrate doping concentrations. (After Takagi *et al.*, 1988).

believed to be due to Coulomb (or impurity) scattering, which becomes more important when the doping concentration is high and the gate voltage or the normal field is low. There is less effect of Coulomb scattering on mobility when the inversion charge density is high because of charge screening effects. At 77 K, μ_{eff} is an even stronger function of $\mathscr{E}_{eff}$ and N_a. At low temperatures, surface scattering is the dominant mechanism at high fields, while Coulomb scattering dominates at low fields.

3.1.5.3 HOLE MOBILITY DATA

Similar mobility–field data for pMOSFETs are shown in Fig. 3.14. In this case, however, the effective normal field is defined by

$$\mathscr{E}_{eff} = \frac{1}{\varepsilon_{si}}\left(|Q_d| + \frac{1}{3}|Q_i|\right), \tag{3.52}$$

which has been found necessary in order for the measured hole mobilities to fall on a universal curve when plotted against $\mathscr{E}_{eff}$ (Arora and Gildenblat, 1987). Note that the factor $\frac{1}{3}$ is entirely empirical with no physical reasoning behind it. Like electron mobility, hole mobility is also influenced by Coulomb scattering at low fields, depending on the doping concentration. The field dependence is also stronger at 77 K, but not quite as strong as in the electron case. It should be noted that the hole mobility data were taken from surface-channel pMOSFETs with p$^+$ polysilicon gate. Buried-channel pMOSFETs with n$^+$ polysilicon gate have a higher mobility (by about 30%) for the same threshold voltage. This is because the normal field in a buried-channel device, given by Eq. (3.49) with $V_{fb} = +E_g/2q - \psi_B$, is much lower. However, buried-channel MOSFETs cannot be scaled to as short a channel length as surface-channel MOSFETs, as will be discussed in Chapter 4.

At higher temperatures, the MOSFET channel mobility decreases because of increased phonon scattering. The temperature dependence is similar to that of bulk mobility discussed in Section 2.1.3, i.e., $\mu_{eff} \propto T^{-3/2}$.

3.1.6 MOSFET CAPACITANCES AND INVERSION-LAYER CAPACITANCE EFFECT

In this subsection, we discuss the intrinsic capacitances of a MOSFET device in different regions of operation and the effect of finite inversion-layer capacitance, which has been neglected in the charge-sheet model, on linear I_{ds}–V_g characteristics.

3.1.6.1 INTRINSIC MOSFET CAPACITANCES

The capacitance of a MOSFET device plays a key role in the switching delay of a logic gate, since for a given current the capacitance determines how fast the gate can be charged (or discharged) to a certain potential which turns on (or off) the source-to-drain current. MOSFET capacitances can be divided into two main categories: intrinsic capacitances and parasitic capacitances. This sub-subsection focuses on the intrinsic MOSFET capacitances arising from the inversion and depletion charges in the channel region. Parasitic capacitances are discussed in Section 5.2. As in the earlier drain-current discussions, gate capacitances are also considered separately in the three regions of MOSFET operation: subthreshold region, linear region, and saturation region, as shown in Fig. 3.9.

- *Subthreshold region.* In the subthreshold region, the inversion charge is negligible. Only the depletion charge needs to be supplied when the gate potential is changed. Therefore, the intrinsic gate-to-source–drain capacitance is essentially zero (the extrinsic gate-to-source–drain overlap capacitance is discussed in Section 5.2.2), while the gate-to-body capacitance is given by the serial combination of C_{ox} and C_d (Fig. 2.28), i.e.,

$$C_g = WL\left(\frac{1}{C_{ox}} + \frac{1}{C_d}\right)^{-1} \approx WLC_d, \qquad (3.53)$$

where C_d is the depletion capacitance per unit area given by Eq. (2.174). For high drain biases, the surface potential and therefore the depletion width at the drain end of the channel become larger, according to Eq. (3.4). The average

C_d to be used in Eq. (3.53) should then be slightly lower than that evaluated at the source end.

- *Linear region.* Once the surface channel forms, there is no more capacitive coupling between the gate and the body due to screening by the inversion charge. All the gate capacitances are to the channel, i.e., to the source and drain terminals. Within the framework of the charge-sheet model, Eq. (3.24), the inversion charge density Q_i at low drain biases varies linearly from $-C_{ox}(V_g - V_t)$ at the source end to $-C_{ox}(V_g - V_t - mV_{ds})$ at the drain end. The total inversion charge under the gate is then $-WLC_{ox}(V_g - V_t - mV_{ds}/2)$, and the gate to channel capacitance is simply given by the oxide capacitance,

$$C_g = WLC_{ox}. \tag{3.54}$$

- *Saturation region.* When V_{ds} is appreciable, the inversion charge density $Q_i(y)$ varies parabolically along the channel, as shown in Fig. 3.7. At the pinch-off condition, $V_{ds} = V_{dsat} = (V_g - V_t)/m$, and $Q_i = 0$ at the drain. In this case,

$$Q_i(y) = -C_{ox}(V_g - V_t)\sqrt{1 - \frac{y}{L}}, \tag{3.55}$$

from Eqs. (3.24) and (3.27). The total inversion charge obtained by integrating Eq. (3.55) in both the channel length (y) and the channel width directions is then $-\frac{2}{3}WLC_{ox}(V_g - V_t)$, and the gate-to-channel capacitance in the saturation region is

$$C_g = \frac{2}{3}WLC_{ox}. \tag{3.56}$$

3.1.6.2 INVERSION-LAYER CAPACITANCE

In Section 3.1.2 and in the discussions above, MOSFET I–V relations and capacitances were derived based on the charge-sheet approximation that all the inversion charge is located at the silicon surface and the surface potential is pinned at $\psi_s(\text{inv}) = 2\psi_B$ once the inversion layer forms. In reality, the inversion layer has a finite thickness (Fig. 2.26), and the surface potential still increases slightly with V_g even beyond $2\psi_B$ (Fig. 2.25). In other words, there is a finite inversion-layer capacitance, $C_i = -dQ_i/d\psi_s$, in series with the oxide capacitance. As a result, the inversion charge density is less than that given by Eq. (3.17). The error is illustrated in the Q_i–V_g curves in Fig. 3.15. The dashed line represents $|Q_i| = C_{ox}(V_g - V_t)$ from the charge-sheet model. The solid curve is a more exact solution calculated numerically from Eq. (3.12) and Eq. (3.14). The discrepancy extends to high gate voltages but is more serious for low-voltage operation. An approximate expression for the inversion charge density taking this effect into account can be derived by

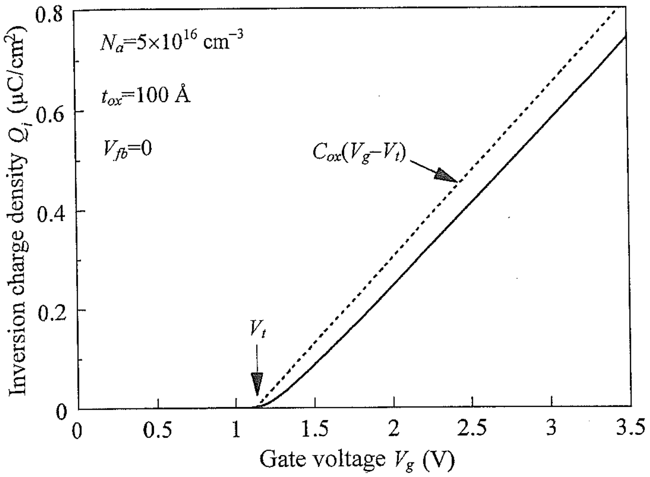


FIGURE 3.15. Q_i–V_g curve (solid line) calculated from Pao and Sah's model for zero drain voltage, compared with that of the charge-sheet approximation (dotted line). V_t indicates the $2\psi_B$ threshold.

considering the small-signal capacitances in Fig. 2.28(b) (Wordeman, 1986),

$$\frac{d(-Q_i)}{dV_g} = \frac{C_{ox}C_i}{C_{ox}+C_i+C_d} \approx C_{ox}\left(1 - \frac{1}{1+C_i/C_{ox}}\right). \tag{3.57}$$

Here $C_d \approx 0$ after the onset of strong inversion because of screening by the inversion charge. From Eq. (2.178), $C_i \approx |Q_i|/(2kT/q)$. Since $|Q_i| \approx C_{ox}(V_g - V_t)$, one can write $C_i/C_{ox} = (V_g - V_t)/(2kT/q)$. Substituting it into Eq. (3.57) and integrating with respect to V_g, one obtains

$$-Q_i = C_{ox}\left[(V_g - V_t) - \frac{2kT}{q}\ln\left(1 + \frac{q(V_g - V_t)}{2kT}\right)\right], \tag{3.58}$$

which agrees well with the numerically calculated curve in Fig. 3.15.

3.1.6.3 EFFECT OF POLYSILICON-GATE DEPLETION ON INVERSION CHARGE

Depletion of polysilicon gates, discussed in Section 2.3.4, can also have an effect on the Q_i–V_g curve if the gate is not doped highly enough. To first order, the depletion region in polysilicon acts like a large capacitor in series with the oxide capacitor, which further degrades inversion charge density for a given applied gate voltage. In contrast to the inversion-layer capacitance effect, however, the gate depletion effect becomes more severe at high gate voltages. Assuming an n$^+$ polysilicon gate on nMOSFET (and vice versa for pMOSFET) and following a similar approach to that in Eq. (2.185), one can add an additional term to Eq. (3.58) for polysilicon depletion effects:

$$-Q_i = C_{ox}\left[(V_g - V_t) - \frac{2kT}{q}\ln\left(1 + \frac{q(V_g - V_t)}{2kT}\right) - \frac{C_{ox}^2(V_g - V_t)^2}{2\varepsilon_{si} q N_p}\right]. \tag{3.59}$$

Here N_p is the electrically active doping concentration of the polysilicon gate, and the gate charge density Q_p has been approximated by $C_{ox}(V_g - V_t)$ (ignoring the depletion charge in bulk silicon). Note that the factor of $\frac{1}{2}$ in the polysilicon depletion term arises from integrating a gate-voltage-dependent capacitance with respect to the gate voltage. In order to keep the last degradation term negligible, N_p should be in the range of 10^{20} cm^{-3}, especially for thin-oxide MOSFETs.

3.1.6.4 LINEAR I_{ds}–V_g CHARACTERISTICS

Given $Q_i(V_g)$ and $\mu_{eff}(V_g)$ [Eqs. (3.51) and (3.50)], the low-drain-bias (linear) I_{ds}–V_g curve is simply

$$I_{ds}(V_g) = \mu_{eff}(V_g)\frac{W}{L}Q_i(V_g)V_{ds}. \tag{3.60}$$

An example is shown in Fig. 3.16, where I_{ds} is calculated assuming no polysilicon depletion with $Q_i(V_g)$ from Fig. 3.15 (solid curve). Both the drain current and the *linear transconductance*, defined by $g_m \equiv dI_{ds}/dV_g$, are degraded significantly at high gate voltages because of the decrease of mobility with increasing normal field. There is a point of maximum slope or linear transconductance about 0.5 V above the threshold voltage. It is conventional to define the linearly extrapolated threshold

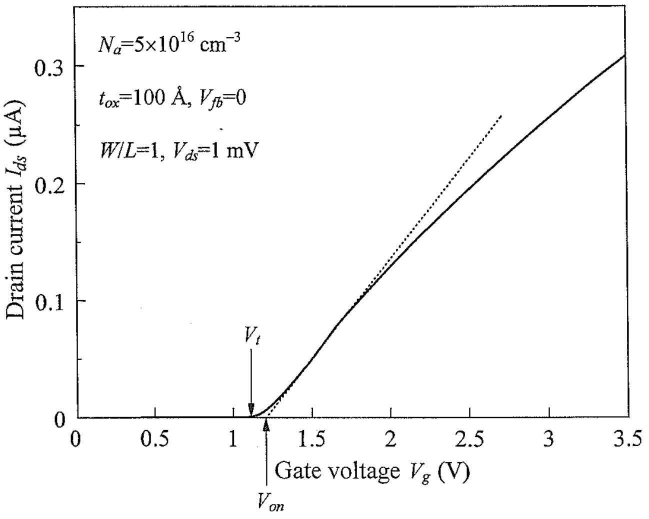


FIGURE 3.16. Calculated low-drain I_{ds}–V_g curve with inversion-layer capacitance and mobility degradation effects. The dotted line shows the linearly extrapolated threshold voltage V_{on}.

voltage, V_{on}, by the intercept of a tangent through this point. For a second-order correction in V_{ds} based on Eq. (3.21), V_{on} is obtained by subtracting $mV_{ds}/2$ from the intercept. Because of the combined inversion-layer capacitance and mobility degradation effects, the linearly extrapolated threshold voltage, V_{on}, is typically (2–4)kT/q higher than the threshold voltage V_t at $\psi_s(\text{inv}) = 2\psi_B$. One should be careful not to mix up V_{on} with V_t, which is used in Eq. (3.36) for estimating subthreshold currents. At $V_g = V_{on}$, the extrapolated subthreshold current (along the same subthreshold slope in a semilog plot) is about 10× of that of Eq. (3.36) for $V_g = V_t$. This current is rather insensitive to temperature but does depend on the technology generation.

A commonly used expression for the low-drain, linear I_{ds}–V_g characteristics takes the form

$$I_{ds}(V_g) = \mu'_{eff}(V_g)C_{ox}\frac{W}{L}\left(V_g - V_{on} - \frac{m}{2}V_{ds}\right)V_{ds}. \tag{3.61}$$

Note that the inversion-layer capacitance effect is lumped into $\mu'_{eff}(V_g)$, which in general is different from $\mu_{eff}(V_g)$ in Eq. (3.60) except at high gate voltages.

3.2 SHORT-CHANNEL MOSFETs

It is clear from Section 3.1 that for a given supply voltage, the MOSFET current increases with decreasing channel length. The intrinsic capacitance of a short-channel MOSFET is also lower, which makes it easier to switch. However, for a given process, the channel length cannot be arbitrarily reduced even if allowed by lithography. Short-channel MOSFETs differ in many important aspects from long-channel devices discussed in Section 3.1. This section covers the basic features of short-channel devices that are important for device design consideration. These features are: (a) short-channel effect, (b) velocity saturation, (c) channel length modulation, (d) source–drain series resistance, and (e) MOSFET breakdown.

3.2.1 SHORT-CHANNEL EFFECT

The *short-channel effect* (SCE) is the decrease of the MOSFET threshold voltage as the channel length is reduced. An example is shown in Fig. 3.17 (Taur *et al.*, 1985). The short-channel effect is especially pronounced when the drain is biased at a voltage equal to that of the power supply (high drain bias). In a CMOS VLSI technology, channel length varies statistically from chip to chip, wafer to wafer, and lot to lot due to process tolerances. The short-channel effect is therefore an important consideration in device design; one must ensure that the threshold voltage does not become too low for the minimum-channel-length device on the chip.

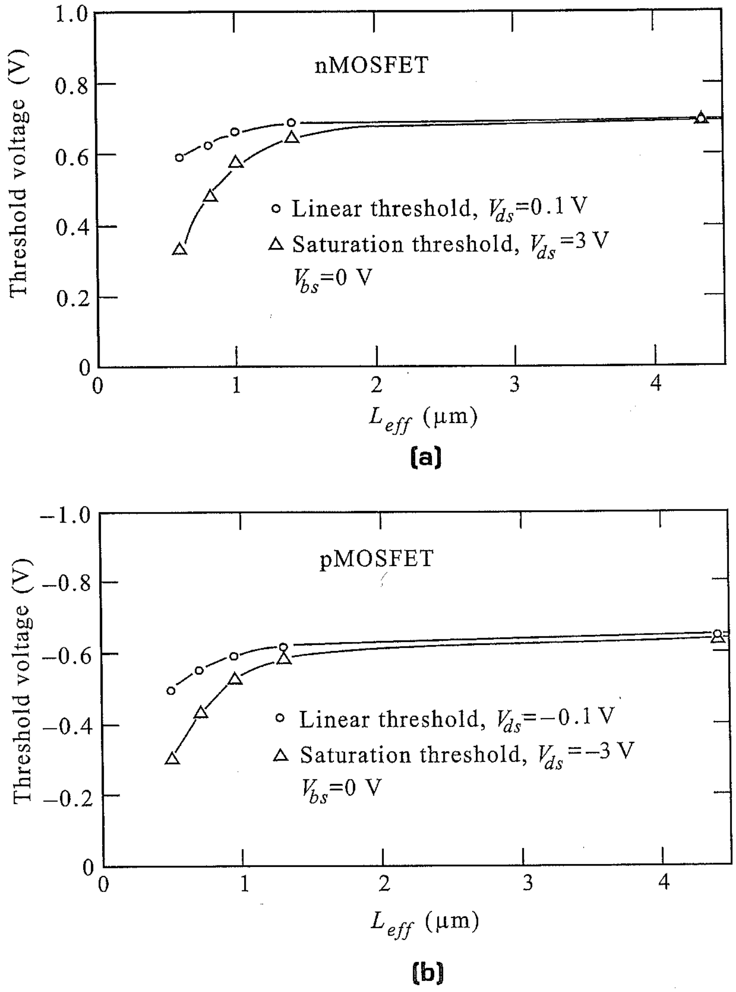


FIGURE 3.17. Short-channel threshold roll-off: Measured low- and high-drain threshold voltages of n- and p-MOSFETs versus channel length. (After Taur *et al.*, 1985.)

3.2.1.1 2-D POTENTIAL CONTOURS AND THE CHARGE-SHARING MODEL

The key difference between a short-channel and a long-channel MOSFET is that the field pattern in the depletion region of a short-channel MOSFET is two-dimensional, as shown in Fig. 3.18. The constant-potential contours in a long-channel device in Fig. 3.18(a) are largely parallel to the oxide–silicon interface or along the channel length direction (y-axis), so that the electric field is one-dimensional (along the vertical direction or x-axis) over the most part of the device. The constant-potential contours in a short-channel device in Fig. 3.18(b), however, are more curvilinear, and the resulting electric field pattern is of a two-dimensional nature. In other words, both the x- and y-components of the electric field are appreciable in a short-channel MOSFET. It is also important to note that, for a given gate voltage, there is more band bending (higher ψ) at the silicon–oxide interface in a short-channel device than in a long-channel device. Specifically, the maximum surface potential

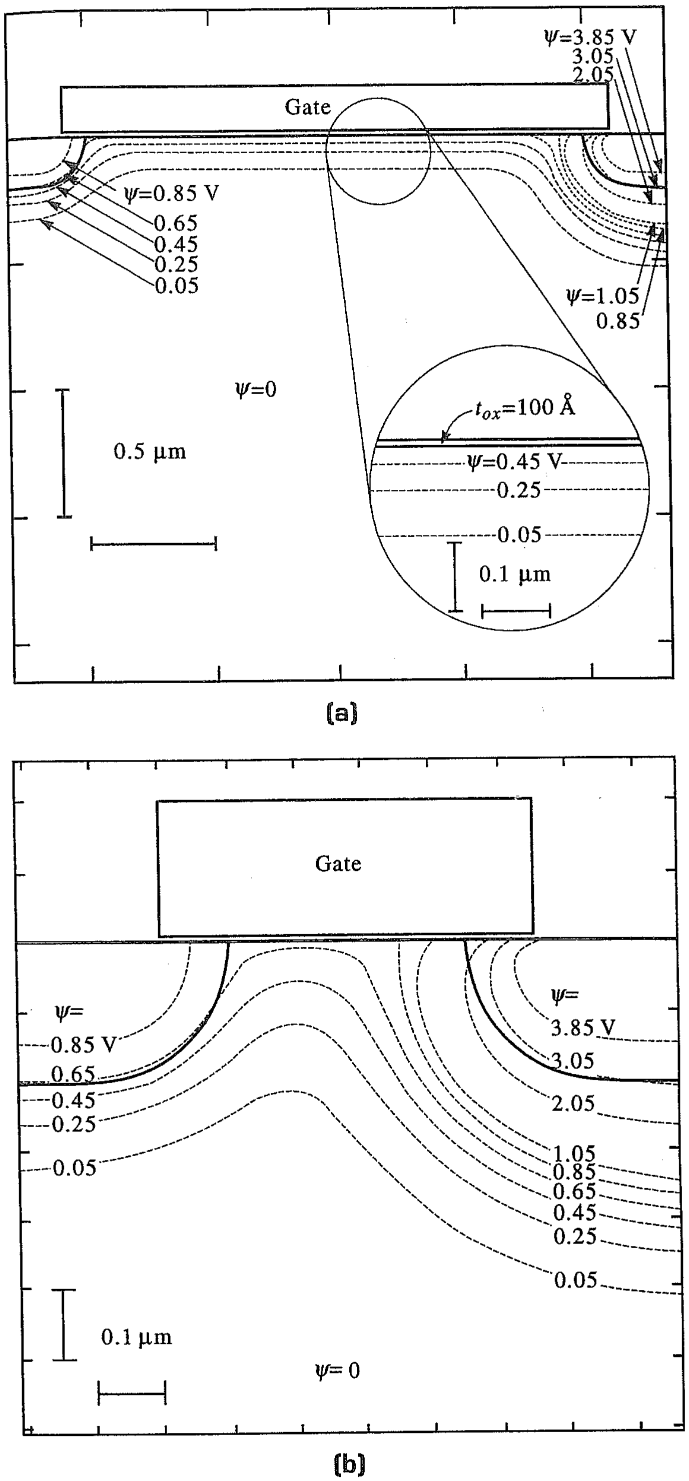


FIGURE 3.18. Simulated constant potential contours of (a) a long-channel and (b) a short-channel nMOSFET. The contours are labeled by the band bending with respect to the neutral p-type region. The solid lines indicate the location of the source and drain junctions (metallurgical). The drain is biased at 3.0 V. Both devices are biased at the same gate voltage slightly below the threshold.

at the
ional,
ice in
annel
g the
ential
inear,
other
short-
there
annel
ential

is slightly over 0.65 V (the fourth contour from the bottom) in Fig. 3.18(b), but below 0.65 V in Fig. 3.18(a). The depletion region width, as indicated by the depth of the first contour ($\psi = 0.05$ V) from the bottom, is also wider in the short-channel case. These all point to a lower threshold voltage in the short-channel MOSFET.

The two-dimensional field pattern in a short-channel device arises from the proximity of source and drain regions. Just like the depletion region under an MOS gate (Section 2.3), there are also depletion regions surrounding the source and drain junctions (Section 2.2 and Fig. 2.34). In a long-channel device, the source and drain are far enough separated that their depletion regions have no effect on the potential or field pattern in most part of the device. ***In a short-channel device, however, the source–drain distance is comparable to the MOS depletion width in the vertical direction, and the source–drain potential has a strong effect on the band bending over a significant portion of the device.*** One way to describe it is to consider the net charge (ionized acceptors or donors) in the depletion region of the device. The field lines terminating on these fixed charges originate either from the gate or from the source and drain. This is referred to as the *charge-sharing model* (Yau, 1974), as shown in Fig. 3.19. At a low drain voltage, only the field lines terminating on the depletion charges within the trapezoidal region are assumed to originate from the gate. The rest of the field lines originate either from the source or from the drain. The total charge within the trapezoidal region, $Q'_B \propto W_{dm} \times (L + L')/2$, is proportionally less than the total gate depletion charge, $Q_B \propto W_{dm} \times L$, in the long-channel case. As a result, it takes a lower gate voltage to reach the threshold condition of a short-channel device,

$$V_t = V_{fb} + 2\psi_B + \frac{Q'_B}{WLC_{ox}}, \tag{3.62}$$

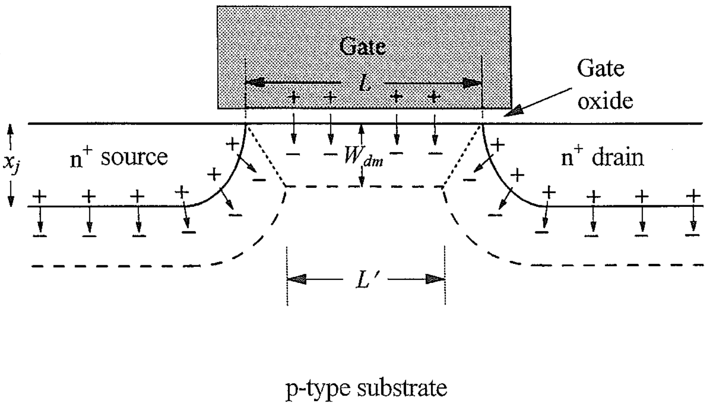
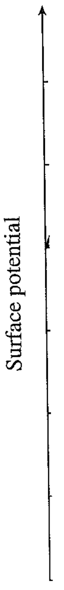


FIGURE 3.19. Schematic diagram of the charge-sharing model. The dashed lines indicate the boundary of the gate and source–drain depletion regions. The arrows represent electric field lines that originate from a positive charge and terminate on a negative charge. The dotted lines partition the depletion charge and form the two sides of the trapezoid discussed in the text. (After Yau, 1974.)



FIGU
L) fr
MOS
volta

base
volt
char
mor
in th

3.2
The
by c
the
Und
fron
volt
limi
by t
flat
the
pen
betv

), but
lepth
annel
ET.
n the
MOS
drain
drain
ential
r, the
rtical
ıding
ısider
vice.
ıte or
(Yau,
ating
inate
from
')/2,
n the
shold

3.62)

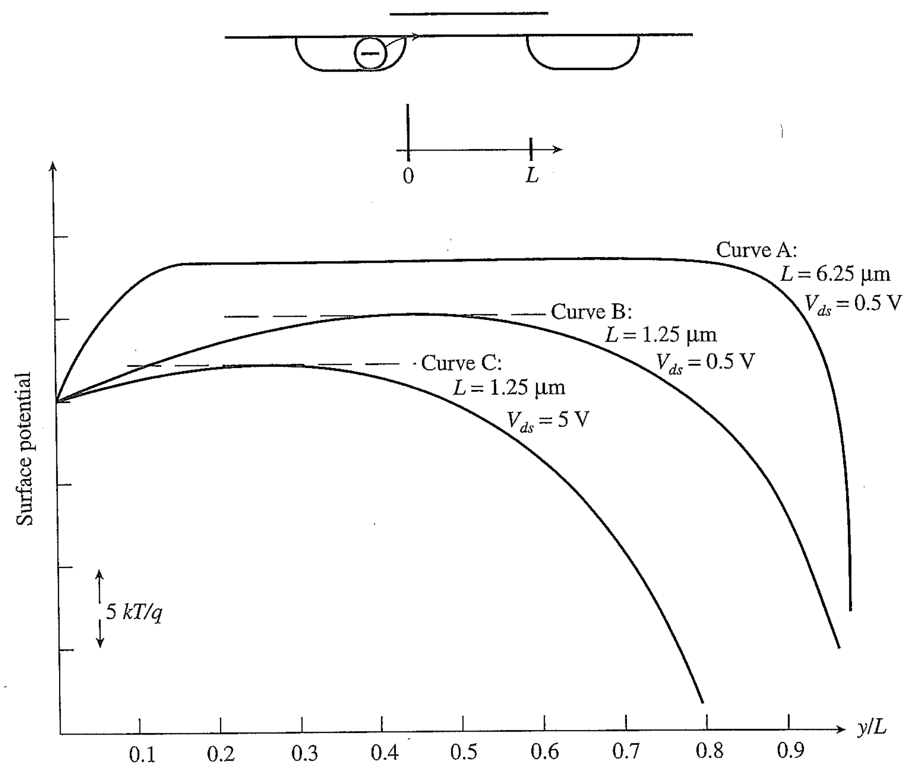


FIGURE 3.20. Surface potential versus lateral distance (normalized to the channel length L) from the source to the drain for (a) a long-channel MOSFET, (b) a short-channel MOSFET at low drain bias, and (c) a short-channel MOSFET at high drain bias. The gate voltage is the same for all three cases. (After Troutman, 1979.)

based on Eq. (3.20). Even though a simple, analytical expression for threshold voltage can be obtained from the charge-sharing model, the division of depletion charge between the gate and the source and drain is somewhat arbitrary. Furthermore, there is no simple way of dealing with high-drain-bias conditions, especially in the subthreshold region.

3.2.1.2 DRAIN-INDUCED BARRIER LOWERING

The physics of the short-channel effect can be understood from a different angle by considering the potential barrier (to electrons for an n-channel MOSFET) at the surface between the source and drain, as shown in Fig. 3.20 (Troutman, 1979). Under off conditions, this potential barrier (p-type region) prevents electron current from flowing to the drain. The surface potential is mainly controlled by the gate voltage. When the gate voltage is below the threshold voltage, there are only a limited number of electrons injected from the source over the barrier and collected by the drain (subthreshold current). In the long-channel case, the potential barrier is flat over most part of the device. Source and drain fields only affect the very ends of the channel. As the channel length is shortened, however, the source and drain fields penetrate deeply into the middle of the channel, which lowers the potential barrier between the source and drain. This causes a substantial increase of the subthreshold

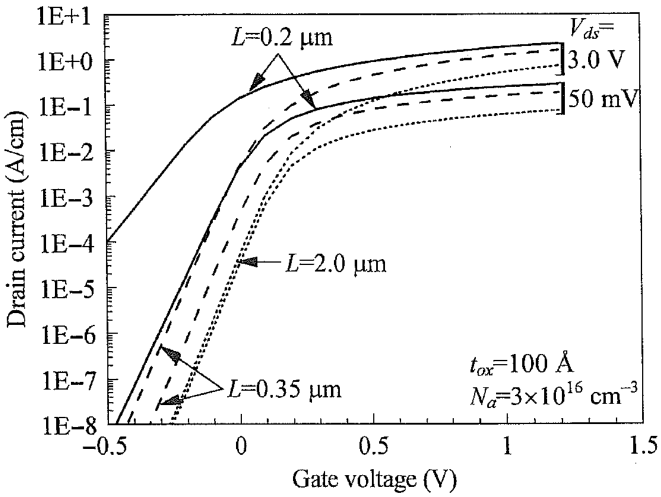


FIGURE 3.21. Subthreshold characteristics of long- and short-channel devices at low and high drain bias.

current. In other words, the threshold voltage becomes lower than the long-channel value. The region of maximum potential barrier also shrinks to a single point near the center of the device.

When a high drain voltage is applied to a short-channel device, the barrier height is lowered even more, resulting in further decrease of the threshold voltage. The point of maximum barrier also shifts toward the source end as shown in Fig. 3.20. This effect is referred to as *drain-induced barrier lowering* (DIBL). It explains the experimentally observed increase of subthreshold current with drain voltage in a short-channel MOSFET. Figure 3.21 shows the subthreshold characteristics of long- and short-channel devices at different drain bias voltages. For long-channel devices, the subthreshold current is independent of drain voltage ($\geq 2kT/q$), as expected from Eq. (3.36). For short-channel devices, however, there is a parallel shift of the curve to a lower threshold voltage for high drain bias conditions. At even shorter channel lengths, the subthreshold slope starts to degrade as the surface potential is more controlled by the drain than by the gate. Eventually, the device reaches the *punch-through* condition when the gate totally loses control of the channel and high drain current persists independent of gate voltage.

3.2.1.3 2-D POISSON'S EQUATION AND LATERAL FIELD PENETRATION

Further insight into the role of the lateral electric field, $\mathscr{E}_y = -\partial\psi_i/\partial y$, in a short-channel MOSFET can be gained by examining the two-dimensional Poisson's equation,

$$\frac{\partial^2\psi_i}{\partial x^2} + \frac{\partial^2\psi_i}{\partial y^2} = -\frac{\rho}{\varepsilon_{si}}. \tag{3.63}$$

In the depletion region of an nMOSFET, mobile carrier densities are negligible.

Only ionized acceptors need to be considered. For a uniformly doped background concentration N_a, Poisson's equation can be written in terms of the electric fields as

$$\frac{\partial \mathscr{E}_x}{\partial x}+\frac{\partial \mathscr{E}_y}{\partial y}=\frac{\rho}{\varepsilon_{si}}=-\frac{qN_a}{\varepsilon_{si}}, \tag{3.64}$$

where $\mathscr{E}_x = -\partial\psi_i/\partial x$ is the electric field in the vertical direction. The depletion charge density, $\rho = -qN_a$, from ionized acceptors can be considered as being split into two parts: the first part, $\varepsilon_{si}\partial\mathscr{E}_x/\partial x$, is controlled by the gate field in the vertical direction; the second part, $\varepsilon_{si}\partial\mathscr{E}_y/\partial y$, is controlled by the source–drain field in the lateral direction (Nguyen and Plummer, 1981). In a long-channel device, the lateral field is negligible over most of the channel, and almost all of the depletion charge is controlled by the gate field. In a short-channel device, the lateral field becomes appreciable. Figure 3.22(a) shows an example of the magnitude of the lateral field

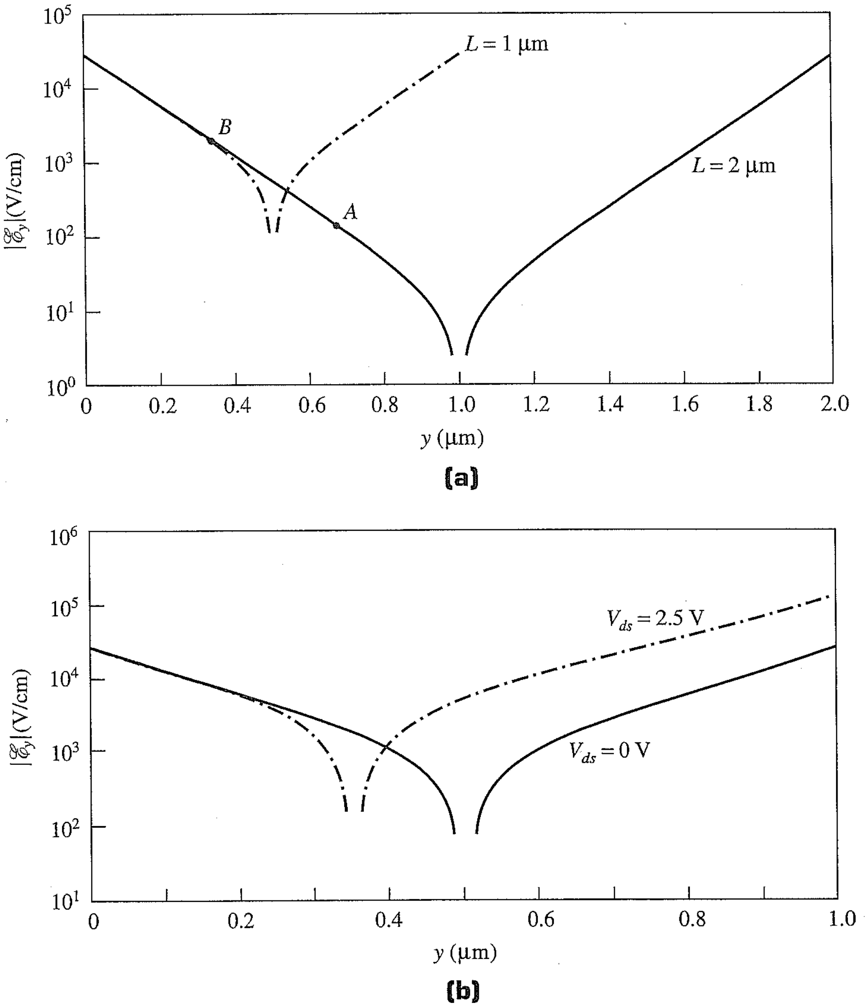


FIGURE 3.22. Simulated lateral field as a function of lateral distance along a horizontal cut through the gate-depletion layer for (a) long- and short-channel devices and (b) low and high drain bias voltages. (After Nguyen, 1984.)

along the channel length direction as obtained from a 2-D numerical simulation. The lateral field is highest at the source and drain junctions, decreasing exponentially toward the middle of the channel. At low drain voltages, the source and drain fields cancel each other exactly at the center of the device. ***When the channel length becomes shorter, the characteristic length of the exponential decay remains unchanged, while the magnitude of the lateral field near the middle of the device increases significantly. This depicts the penetration of source and drain fields into the channel region of a short-channel MOSFET.*** Application of a high drain voltage [Fig. 3.22(b)] does not change the source field but does increase the drain field. This shifts the zero-field point toward the source, thus making it asymmetric, and at the same time raises the lateral field intensity even further. The zero-field point corresponds to the point of the least band bending in Fig. 3.18, as well as the point of maximum potential barrier in Fig. 3.20.

With the increase of the lateral field strength, the source–drain controlled depletion charge density, $\varepsilon_{si}\,\partial\mathscr{E}_y/\partial y$, increases as the gate-controlled depletion charge density,

$$\varepsilon_{si}\frac{\partial\mathscr{E}_x}{\partial x} = \rho - \varepsilon_{si}\frac{\partial\mathscr{E}_y}{\partial y}, \tag{3.65}$$

decreases and becomes appreciably less than the ionized impurity charge concentration, ρ (Nguyen and Plummer, 1981). (For most typical doping profiles, $\partial\mathscr{E}_x/\partial x$ and $\partial\mathscr{E}_y/\partial y$ have the same sign as ρ.) Just as in the one-dimensional MOS capacitor discussed in Section 2.3.2, an effectively lower depletion charge concentration results in reduced surface field and wider depletion width for a given surface potential, which means a lower threshold voltage for short-channel devices. Figure 3.23

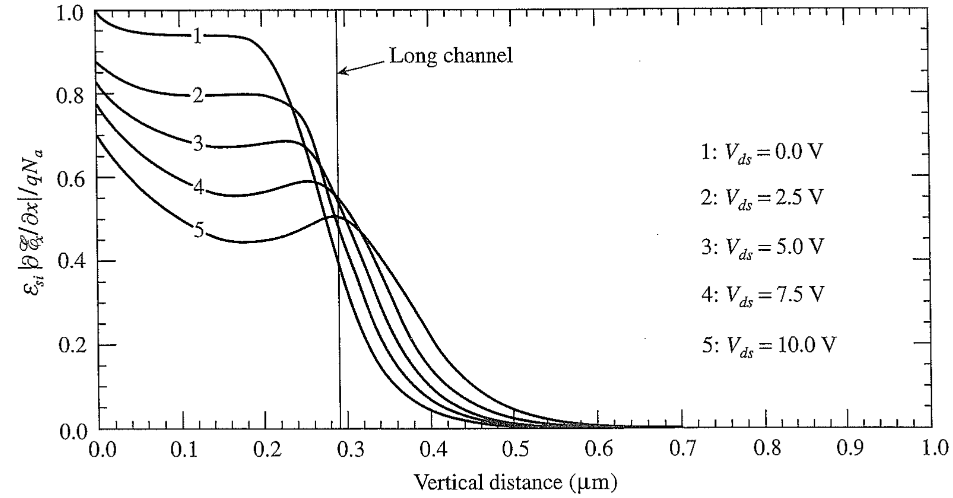


FIGURE 3.23. Fraction of the gate-controlled depletion charge density versus vertical distance for several different drain voltages. Under the depletion approximation, the long-channel curve is a step function that goes from 1 to 0 at the edge of the depletion region. (After Nguyen, 1984.)

shows the fraction of depletion charge controlled by the gate field versus the vertical distance from the surface. As the drain voltage increases, the effective gate-controlled charge density decreases significantly below the long-channel value. Even though the depletion region becomes slightly wider, the integrated charge density (area under the curve), and therefore the threshold voltage, decreases.

3.2.1.4 AN ANALYTICAL EXPRESSION FOR SHORT-CHANNEL THRESHOLD VOLTAGE

With a few approximations, an analytical solution to the two-dimensional Poisson's equation can be obtained using a simplified short-channel MOSFET geometry in Appendix 6 (Nguyen, 1984). The region of interest is a rectangular box of length equal to the channel length L defined as the distance between the source and the drain (Fig. 3.19). In the vertical direction, the box consists of an oxide region of thickness t_{ox} and a silicon region of depth given by the depletion-layer width W_d [Eq. (2.160)]. To eliminate the discontinuity of $\partial\psi_i/\partial x$ across the silicon–oxide boundary, the oxide is replaced by an equivalent region of the same dielectric constant as silicon, but with a thickness equal to $(\varepsilon_{si}/\varepsilon_{ox})t_{ox} = 3t_{ox}$. The entire rectangular region can then be treated as a homogeneous material of height $W_d + 3t_{ox}$ and dielectric constant ε_{si}. This is a good approximation when the oxide is thin compared with the depletion depth W_d, as is the case with most practical CMOS technologies.

The boundary conditions of the electrostatic potential at the source and drain boundaries are ψ_{bi} and $\psi_{bi} + V_{ds}$, respectively, with the potential in the neutral p-type region defined as zero. Here ψ_{bi} is the built-in potential of the source- or drain-to-substrate junction, and V_{ds} is the drain voltage. For an abrupt n$^+$–p junction, $\psi_{bi} = E_g/2q + \psi_B$, where ψ_B is given by Eq. (2.37). Typically, $\psi_{bi} \approx$ 0.8–0.9 V.

Under subthreshold conditions, current conduction is dominated by diffusion and is mainly controlled by the point of highest barrier for electrons along the channel, as shown in Fig. 3.18(b) and Fig. 3.20. The threshold voltage of a short-channel device is defined as the gate voltage at which the minimum electrostatic potential (maximum barrier for electrons) at the surface equals $2\psi_B$. It is shown in Appendix 6 that this occurs at a gate voltage lower than the long-channel threshold voltage by an amount

$$\Delta V_t = \frac{24t_{ox}}{W_{dm}}\sqrt{\psi_{bi}(\psi_{bi} + V_{ds})}e^{-\pi L/2(W_{dm}+3t_{ox})}. \tag{3.66}$$

Here W_{dm} is the minimum depletion width at the threshold condition in a short-channel device, as indicated in Fig. 3.19. In order to distinguish it from W_{dm}, the depletion width at the threshold condition in a long-channel device will be designated as W^0_{dm}, as in Appendix 6. If L is not too short, the body-effect coefficient m

can be approximated by

$$m \equiv 1 + \frac{\varepsilon_{si}/W_{dm}^0}{C_{ox}} \approx 1 + \frac{3t_{ox}}{W_{dm}}, \tag{3.67}$$

and Eq. (3.66) can be expressed as

$$\Delta V_t = 8(m-1)\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})}e^{-\pi L/2mW_{dm}}. \tag{3.68}$$

Typically, $m \approx 1.1$–1.4. These analytical short-channel threshold roll-off expressions are good approximations if the source and drain junctions are deeper than the maximum gate depletion region, i.e., if $x_j \geq W_{dm}$.

Because of the exponential facor, the threshold voltage roll-off with channel length is very sensitive to the gate depletion width W_{dm}. For very short channel lengths, the minimum depletion width is larger than the long-channel value, as can be seen from Figs. 3.18 and 3.23. If the short-channel effect is not too severe, however, W_{dm} can be approximated to the first order by its long-channel value,

$$W_{dm}^0 = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2N_a}} \tag{3.69}$$

from Eq. (2.162). W_{dm}^0 is plotted in Fig. 3.24 versus N_a. ***To avoid excessive short-channel effects, the substrate (or well) doping concentration in a CMOS device design should be chosen such that the minimum channel length, L_{min}, is about 2–3 times W_{dm}.*** This is similar to the well-known criterion that the minimum channel length should be greater than the sum of the source and the drain depletion widths, i.e., $L_{\min} \geq W_S + W_D$, where W_S and W_D are given by Eqs. (A6.7) and (A6.8). Another observation of Eq. (3.66) is that a reverse substrate bias, $-V_{bs}$, aggravates short-channel effects, since it changes ψ_{bi} to a higher value, $\psi_{bi} + V_{bs}$. By the same

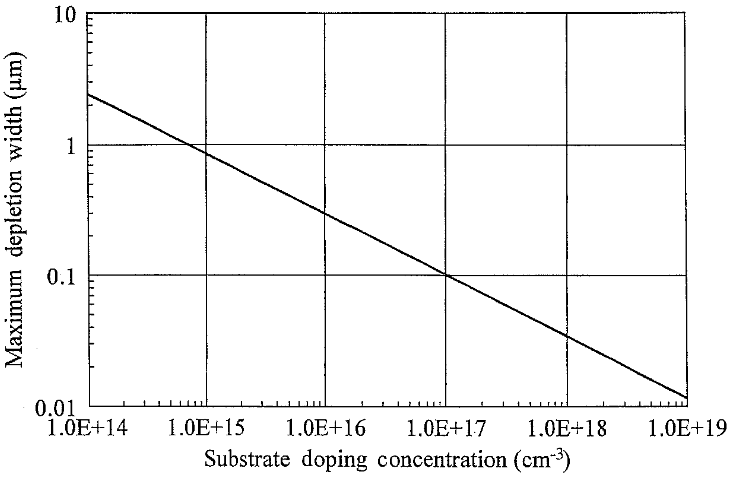


FIGURE 3.24. Depletion region width at $2\psi_B$ threshold condition versus doping concentration for uniformly doped substrates.

token, the substrate sensitivity [Eq. (3.41)] of a short-channel MOSFET is less than that of a long-channel MOSFET. This can be understood from the charge-sharing model (Fig. 3.19) that as W_{dm} increases due to a reverse substrate bias, the gate-controlled depletion charge within the trapezoid region increases in proportion to W_{dm} in a long-channel device, but less than in proportion to W_{dm} in a short-channel device. More on the short-channel substrate sensitivity can be found in Appendix 6.

It should be noted that Eq. (3.66) is, of course, just an approximation. In general, short-channel device design should be carried out with a two-dimensional device simulator for more accurate results. Further details on channel profile and threshold design are discussed in Section 4.2.

3.2.2 VELOCITY SATURATION

As discussed in Section 3.1.2, when the drain voltage increases in a long-channel MOSFET, the drain current first increases, then becomes saturated at a voltage equal to $V_{dsat} = (V_g - V_t)/m$ with the onset of pinch-off at the drain. ***In a short-channel device, the saturation of drain current may occur at a much lower voltage due to velocity saturation.*** This causes the saturation current I_{dsat} to deviate from the $1/L$ dependence depicted in Eq. (3.23) for long-channel devices. Velocity–field relationships in bulk silicon are plotted in Fig. 2.9. Saturation velocities of electrons and holes in a MOSFET channel are slightly lower than their bulk values. $v_{sat} \approx 7\text{–}8 \times 10^6$ cm/s for electrons and $v_{sat} \approx 6\text{–}7 \times 10^6$ cm/s for holes have been reported in the literature (Coen and Muller, 1980; Taur *et al.*, 1993a). Figure 3.25

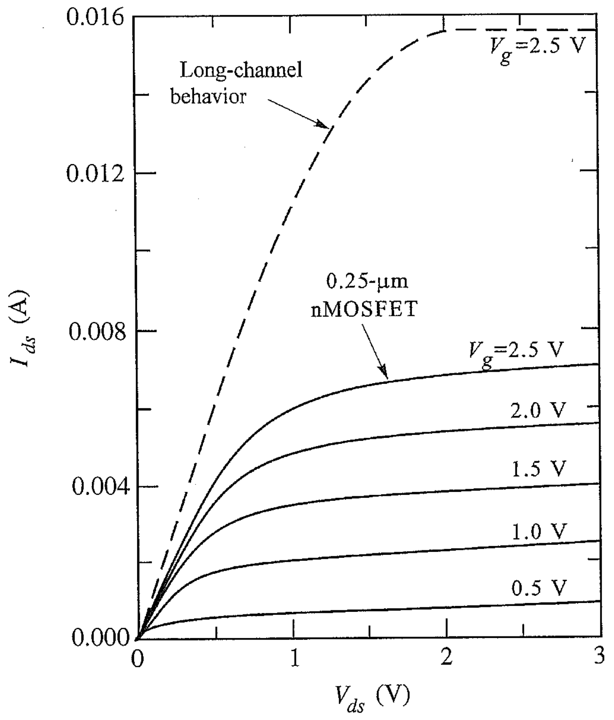


FIGURE 3.25. Experimental I–V curves of a 0.25-μm nMOSFET (solid lines). The device width is 9.5 μm. The dashed curve shows the long-channel-like drain current expected for this channel length if there were no velocity saturation. (After Taur *et al.*, 1993a.)

shows the experimentally measured I_{ds}–V_{ds} curves of a 0.25-μm nMOSFET. The dashed curve represents the long-channel-like current given by Eq. (3.23) for $V_g = 2.5$ V. Due to velocity saturation, the drain current saturates at a drain voltage much lower than $(V_g - V_t)/m$, thus severely limiting the saturation current of a short-channel device.

3.2.2.1 VELOCITY-FIELD RELATIONSHIP

Experimental measurements show that the velocity–field relationship for electrons and holes takes the empirical form (Caughey and Thomas, 1967)

$$v = \frac{\mu_{eff}\mathscr{E}}{[1 + (\mathscr{E}/\mathscr{E}_c)^n]^{1/n}}, \tag{3.70}$$

where $n = 2$ for electrons and $n = 1$ for holes. n (≥ 1) is a measure of how rapidly the carriers approach saturation. The parameter $\mathscr{E}_c$ is called the *critical field.* When the field strength is comparable to or greater than $\mathscr{E}_c$, velocity saturation becomes important. At low fields, $v = \mu_{eff}\mathscr{E}$, which is simply Ohm's law. As $\mathscr{E} \to \infty$, $v = v_{sat} = \mu_{eff}\mathscr{E}_c$. Therefore,

$$\mathscr{E}_c = \frac{v_{sat}}{\mu_{eff}}. \tag{3.71}$$

It was discussed in Section 3.1.5 that the effective mobility μ_{eff} is a function of the vertical (or normal) field $\mathscr{E}_{eff}$. Since v_{sat} is a constant independent of $\mathscr{E}_{eff}$, the critical field $\mathscr{E}_c$ is a function of $\mathscr{E}_{eff}$ as well. More specifically, ***for a higher vertical field, the effective mobility is lower, but the critical field for velocity saturation becomes higher*** (Sodini *et al.*, 1984). Similarly, holes have a critical field higher than that of electrons, since hole mobilities are lower.

3.2.2.2 AN ANALYTICAL SOLUTION FOR $n = 1$

It is more important to treat velocity saturation for electrons. However, the mathematics in solving the $n = 2$ case are rather tedious (Taylor, 1984). To gain an insight into the velocity saturation phenomenon in a MOSFET, we will analyze the $n = 1$ case instead, which has the same basic characteristics although with a different rate of approaching saturation. Following similar steps to those in Section 3.1.1, one replaces the low-field drift velocity, $-\mu_{eff}\, dV/dy$, in Eq. (3.8) with Eq. (3.70) to allow for high-field velocity saturation effects ($n = 1$):

$$I_{ds} = -WQ_i(V)\frac{\mu_{eff}\, dV/dy}{1 + (\mu_{eff}/v_{sat})\, dV/dy}. \tag{3.72}$$

Here V is the quasi-Fermi potential at a point y in the channel, and $Q_i(V)$ is the integrated (vertically) inversion charge density at that point. Note that $dV/dy = -\mathscr{E} > 0$. Current continuity requires that I_{ds} be a constant, independent of y.

FET. The
for $V_g =$
n voltage
rrent of a

electrons

(3.70)

w rapidly
ld. When
becomes
∞, $v =$

(3.71)

ınction of
f $\mathscr{E}_{eff}$, the
r vertical
aturation
ıld higher

ıe mathe-
an insight
the $n = 1$
ferent rate
3.1.1, one
(3.70) to

(3.72)

(V) is the
$dV/dy=$
ent of y.

Rearranging Eq. (3.72), one obtains

$$I_{ds} = -\left(\mu_{eff} W Q_i(V) + \frac{\mu_{eff} I_{ds}}{v_{sat}}\right)\frac{dV}{dy}. \tag{3.73}$$

Multiplying by dy on both sides and integrating from $y = 0$ to L and from $V = 0$ to V_{ds}, one solves for I_{ds}:

$$I_{ds} = \frac{-\mu_{eff}(W/L)\int_0^{V_{ds}} Q_i(V)\,dV}{1 + (\mu_{eff} V_{ds}/v_{sat}L)}. \tag{3.74}$$

The numerator is simply the long-channel current, Eq. (3.10), without velocity saturation. It is clear that if the "average" field along the channel, V_{ds}/L, is much less than the critical field $\mathscr{E}_c = v_{sat}/\mu_{eff}$, the drain current is hardly affected by velocity saturation. When V_{ds}/L becomes comparable to or greater than $\mathscr{E}_c$, however, the drain current is significantly reduced. If one uses the approximate expression (3.24) in the charge-sheet model for $Q_i(V)$,

$$Q_i(V) = -C_{ox}(V_g - V_t - mV), \tag{3.75}$$

the integration in Eq. (3.74) can be carried out to yield

$$I_{ds} = \frac{\mu_{eff} C_{ox}(W/L)\left[(V_g - V_t)V_{ds} - (m/2)V_{ds}^2\right]}{1 + (\mu_{eff} V_{ds}/v_{sat}L)}. \tag{3.76}$$

3.2.2.3 SATURATION DRAIN VOLTAGE AND CURRENT

For a given V_g, I_{ds} increases with V_{ds} until a maximum current is reached. Beyond this point, the drain current is saturated. The saturation voltage, V_{dsat}, can be found by solving $dI_{ds}/dV_{ds} = 0$:

$$V_{dsat} = \frac{2(V_g - V_t)/m}{1 + \sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)}}. \tag{3.77}$$

This expression is always less than the long-channel saturation voltage, $(V_g - V_t)/m$. Substituting Eq. (3.77) into Eq. (3.76), one finds the saturation current,

$$I_{dsat} = C_{ox} W v_{sat}(V_g - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} - 1}{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} + 1}. \tag{3.78}$$

Example curves of I_{dsat} versus $V_g - V_t$ are plotted in Fig. 3.26 for several different channel lengths. In the long-channel case, the solid curve calculated from Eq. (3.78) is not too different from the dashed curve representing the drain current without velocity saturation. In fact, it can be shown that Eq. (3.78) reduces to the long-channel saturation current [Eq. (3.23)],

$$I_{dsat} = \mu_{eff} C_{ox}\frac{W}{L}\frac{(V_g - V_t)^2}{2m}, \tag{3.79}$$

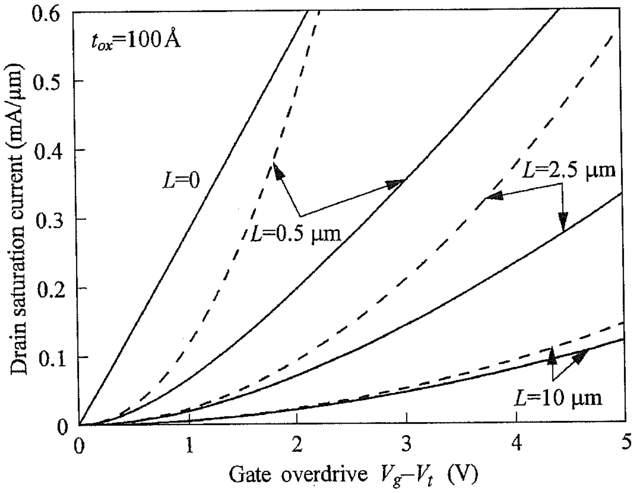


FIGURE 3.26. Saturation current calculated from Eq. (3.78) versus $V_g - V_t$ for several different channel lengths (solid curves). The dashed curves are the corresponding "long-channel-like" saturation currents calculated from Eq. (3.79), i.e., by letting $v_{sat} \to \infty$ in Eq. (3.78). The $L = 0$ line represents the limiting case imposed by velocity saturation, Eq. (3.80).

when $V_g - V_t \ll m v_{sat} L / 2\mu_{eff}$. As the channel length becomes shorter, the velocity-saturated current (solid curves) is significantly less than that of Eq. (3.79) (dashed curves) over an increasing range of gate voltage. In the limit of $L \to 0$, Eq. (3.78) becomes the velocity-saturation-limited current,

$$I_{dsat} = C_{ox} W v_{sat} (V_g - V_t), \tag{3.80}$$

as indicated by the straight line labeled $L = 0$ in Fig. 3.26. ***Note that Eq. (3.80) is independent of channel length L and varies linearly with $V_g - V_t$ instead of quadratically as in the long-channel case.*** This is consistent with observations of the experimental curves in Fig. 3.25. For very short channel lengths, the saturation voltage, Eq. (3.77), can be approximated by

$$V_{dsat} = \sqrt{2 v_{sat} L (V_g - V_t) / m \mu_{eff}}, \tag{3.81}$$

which decreases with channel length.

3.2.2.4 PINCH-OFF POINT AT VELOCITY SATURATION

It is instructive to examine the charge and field behavior at the drain end of the channel when $V_{ds} = V_{dsat}$. From Eq. (3.75),

$$Q_i(y = L) = -C_{ox}(V_g - V_t - m V_{dsat}). \tag{3.82}$$

3.2

Sub

Con
velo
Eq.
infin
Sub
appr
with
are i
Thei
in S
velo
char

3.2.
All t
defi
imat
equi
ultr
tial i
fract
of th
libri
these
velo
A
forn
the l
Fisc
nMC
et al
char
over
duct

from

Substituting V_{dsat} from Eq. (3.77), one finds

$$Q_i(y = L) = -C_{ox}(V_g - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} - 1}{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} + 1}. \tag{3.83}$$

Comparison with Eq. (3.78) yields $I_{dsat} = -Wv_{sat}Q_i(y = L)$, i.e., the carrier drift velocity at the drain end of the channel is equal to the saturation velocity. From Eq. (3.72), this means that the lateral field along the channel, dV/dy, approaches infinity at the drain. Just as in the long-channel pinch-off situation discussed in Subsection 3.1.2, such a singularity leads to the breakdown of the gradual-channel approximation which assumed that the lateral field changes slowly in comparison with the vertical field. In other words, ***beyond the satuation point, carriers which are traveling at satuation velocity are no longer confined to the surface channel.*** Their transport must then be described by a 2-D Poisson's equation to be elaborated in Section 3.2.3. A key difference between pinch-off in long-channel devices and velocity saturation in short-channel devices is that in the latter case, the inversion charge density at the drain, Eq. (3.83), does not vanish.

3.2.2.5 VELOCITY OVERSHOOT

All the MOSFET current formulations discussed thus far, including the mobility definition and velocity saturation, are under the realm of the drift–diffusion approximation, which treats carrier transport in some average fashion always in thermal equilibrium with the silicon lattice. ***The drift–diffusion model breaks down in ultrashort-channel devices where high field or rapid spatial variation of potential is present.*** In such cases, the scattering events are no longer localized, and some fraction of the carriers may acquire much higher than thermal energy over a portion of the device, for example, near the drain. These carriers are not in thermal equilibrium with the silicon lattice and are generally referred to as *hot carriers*. Under these circumstances, it is possible for the carrier velocity to exceed the saturation velocity. This phenomenon is called *velocity overshoot.*

A more rigorous treatment of the carrier transport under spatially nonuniform high-field conditions has been carried out by a Monte Carlo solution of the Boltzmann transport equation for the electron distribution function (Laux and Fischetti, 1988). Figure 3.27 shows the calculated saturation transconductance of nMOSFETs versus channel length, together with experimental results (Sai-Halasz *et al.*, 1988). Local velocity overshoot near the drain starts to occur below 0.2-μm channel length. At channel lengths near 0.05 μm, velocity overshoot takes place over a substantial portion of the device such that the terminal saturation transconductance exceeds the velocity saturation limited value,

$$g_{msat} \equiv \frac{dI_{dsat}}{dV_g} = C_{ox}Wv_{sat}, \tag{3.84}$$

from Eq. (3.80).

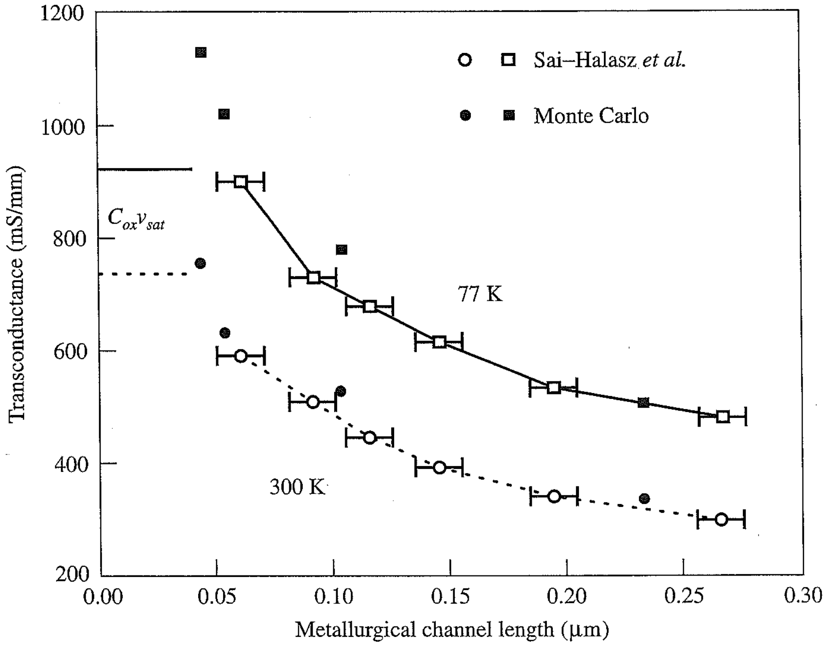


FIGURE 3.27. Measured (open symbols) and calculated (solid symbols) saturation transconductance versus channel length. The gate oxide is 45 Å thick. Absolute upper bounds for the transconductance in the absence of velocity overshoot at 300 and 77 K are indicated by the lines labeled $C_{ox}v_{sat}$. (After Laux and Fischetti, 1988.)

It should be noted that while the carrier velocity can reach rather high values in the high-field region near the drain, MOSFET currents are mainly controlled by the average carrier velocity near the source end of the channel, where the inversion charge density Q_i is $C_{ox}(V_g - V_t)$, independent of drain voltage. Carrier velocity near the source, in turn, is determined by both the thermal velocity as the carriers are injected from the source and the field and scattering rates (mobility) in the channel region near the source (Lundstrom, 1997). Velocity overshoot near the drain helps raise MOSFET currents only to the extent that it increases the field near the source. Once the device approaches the *ballistic* limit, the current depends only on the injection velocity from the source, independent of the field and scattering parameters. In other words, there is an upper limit on the MOSFET current set by thermal injection from the source, and velocity overshoot in the channel does not extend this limit. Such a limiting current takes the same form as Eq. (3.80), except that the parameter v_{sat} should be interpreted as the source thermal velocity v_T (Lundstrom, 1997), which for electrons can be 1.5–2 times v_{sat} in a heavily doped degenerate n$^+$ source.

3.2.3 CHANNEL LENGTH MODULATION

In this subsection, we discuss the characteristics of short-channel MOSFETs biased beyond saturation. In a long-channel device, the drain current stays constant

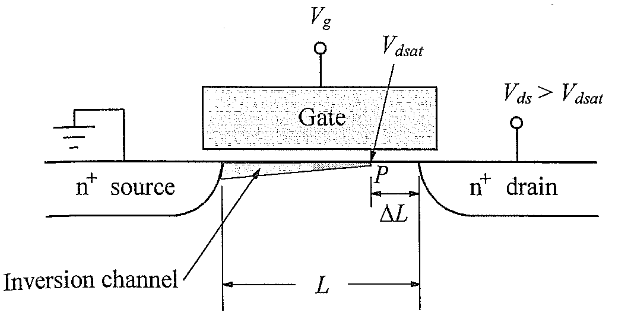


FIGURE 3.28. Schematic diagram showing channel length modulation when a MOSFET is biased beyond saturation. The surface channel collapses at point P, where carriers reach saturation velocity.

when the drain voltage exceeds V_{dsat}, as shown in Fig. 3.8. The output conductance, dI_{ds}/dV_{ds}, is zero in the saturation region. In contrast, the drain current of a short-channel MOSFET can still increase slightly beyond the pinch-off or the velocity saturation point with a nonzero output conductance, as is evident from the experimental curves in Fig. 3.25. This arises because of two factors: the short-channel effect and channel length modulation. The short-channel effect was discussed in Section 3.2.1; when the drain voltage increases beyond saturation in a short-channel device, the threshold voltage decreases, and therefore the drain current increases. In this subsection, we focus on channel length modulation.

3.2.3.1 DRAIN CURRENT BEYOND THE SATURATION POINT

According to the one-dimensional model in the preceding subsection, the electric field along the channel approaches infinity at the saturation point. In practice, the field remains finite. However, its magnitude becomes comparable to the vertical field, so that the gradual-channel approximation breaks down and carriers are no longer confined to the surface channel. ***As the drain voltage increases beyond the saturation voltage V_{dsat}, the saturation point where the surface channel collapses begins to move slightly toward the source***, as shown in Fig. 3.28. The voltage at the saturation point remains constant at V_{dsat}, independent of V_{ds}. The voltage difference $V_{ds} - V_{dsat}$ is dropped across the region between the saturation point and the drain. Carriers injected from the surface channel into this region travel at saturation velocity until collected by the drain junction. The distance between the saturation point and the drain, ΔL, is referred to as the amount of channel length modulation by the drain voltage. Since the one-dimensional model is still valid between the source and the saturation point where the voltage remains at V_{dsat}, the device acts as if its channel length were shortened by ΔL. The drain current is then obtained simply by replacing L with $L - \Delta L$ in Eq. (3.78). In the long-channel limit, this increases the drain current by a factor of $(1 - \Delta L/L)^{-1}$, i.e.,

$$I_{ds} = \frac{I_{dsat}}{1 - (\Delta L/L)}. \tag{3.85}$$

Since ΔL increases with increasing drain voltage, the drain current continues to increase in the saturation region.

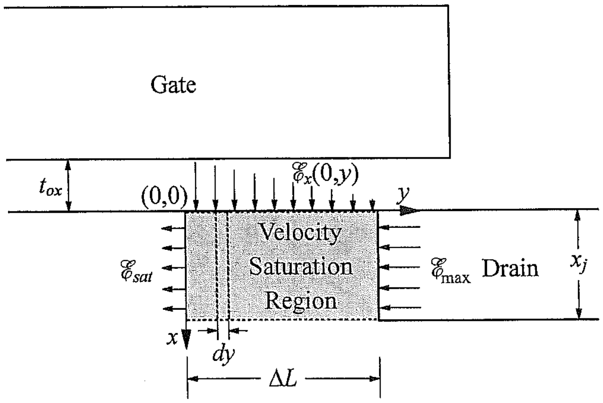


FIGURE 3.29. Schematic diagram of the velocity saturation region for illustrating the pseudo-2-D model. The 2-D Gauss's law is applied to a vertical stripe of width dy bounded by two parallel dotted lines. (After Ko, 1989.)

3.2.3.2 A PSEUDO-2-D MODEL FOR THE VELOCITY SATURATION REGION

To find out ΔL as a function of $V_{ds} - V_{dsat}$, we adopt a pseudo-2-D model (Ko *et al.*, 1981) that yields simple results yet captures the essential physics taking place beyond the saturation point. Figure 3.29 shows a schematic cross section of the region of interest. The velocity saturation region is bounded by $y = 0$ (saturation point) to $y = \Delta L$ (drain), and $x = 0$ (surface) to $x = x_j$ (drain junction depth). The mobile carriers are assumed to spread to a depth equal to the drain junction depth x_j. It is also assumed that the heavily doped drain junction is infinitely abrupt with a square corner.

Along the surface, the quasi-Fermi level $V(y)$ increases from V_{dsat} at $y = 0$ to V_{ds} at $y = \Delta L$. This results in a reduction of the potential drop V_{ox} across the oxide, since the total band offset,

$$V_g - V_{fb} = V_{ox}(y) + \psi_s(y) = V_{ox}(y) + 2\psi_B + V(y), \qquad (3.86)$$

is constant for a fixed gate voltage. Here the surface potential ψ_s is assumed to be pinned at $2\psi_B + V$ as given by Eq. (3.3) for strong inversion. This is valid as long as $V(y) \leq (V_g - V_t)/m$, the long-channel pinch-off voltage. It then follows that the vertical field at the silicon surface,

$$\mathscr{E}_x(0, y) = \frac{\varepsilon_{ox}}{\varepsilon_{si}} \mathscr{E}_{ox}(y) = \frac{\varepsilon_{ox}}{\varepsilon_{si}} \frac{V_{ox}(y)}{t_{ox}}, \qquad (3.87)$$

also decreases toward the drain, as depicted in Fig. 3.29. The silicon–oxide boundary condition, Eq. (2.146), was applied here with $\mathscr{E}_{ox}$ being the oxide field. At $y = 0$, all the silicon charges are still controlled by the gate, so that the one-dimensional Gauss's law is applicable:

$$\mathscr{E}_x(0, 0) = \frac{qN_a x_j + Q_i(y = 0)}{\varepsilon_{si}}, \qquad (3.88)$$

where Q_i (> 0) is the mobile (electron) charge density per unit area. It is assumed here that the junction depth x_j is comparable to the depletion width W_{dm}. Since carriers are already traveling at saturation velocity such that $I_{ds} = WQ_i v_{sat}$, the mobile charge density,

$$Q_i(y) = q \int_0^{x_j} n(x, y)\, dx, \tag{3.89}$$

has to remain constant, i.e., independent of y, toward the drain in order to maintain current continuity. Therefore, ***as the vertical field $\mathscr{E}_x(0, y)$ and the gate-controlled charge decrease toward the drain, some of the mobile charge spreads deep and becomes controlled by the drain***. The physics is similar to that of the 2-D fields discussed in Section 3.2.1. The difference is that fixed depletion charges are involved in the short-channel effect, while mobile charges are involved in the saturation region. As a result of the drain gradually taking control of the mobile charge, the electric field, $\mathscr{E}_y$, originating from the drain increases toward the drain.

Assuming that $\mathscr{E}_y$ is uniform in the x-direction and neglecting the vertical field at the bottom boundary ($x = x_j$), one can apply the two-dimensional Gauss's law to a thin slice of width dy and length x_j located at y (Fig. 3.29):

$$\mathscr{E}_x(0, y)\, dy - \mathscr{E}_y(y + dy)x_j + \mathscr{E}_y(y)x_j = \frac{qN_a x_j\, dy + Q_i\, dy}{\varepsilon_{si}} \tag{3.90}$$

Expanding $\mathscr{E}_y(y + dy)$ into $\mathscr{E}_y(y) + (d\,\mathscr{E}_y/dy)\, dy$ and making use of Eq. (3.88), we obtain

$$-x_j \frac{d\mathscr{E}_y}{dy} = \mathscr{E}_x(0, 0) - \mathscr{E}_x(0, y). \tag{3.91}$$

From Eqs. (3.87) and (3.86), the vertical field difference can be expressed as

$$\begin{aligned} \mathscr{E}_x(0, 0) - \mathscr{E}_x(0, y) &= \frac{\varepsilon_{ox}}{\varepsilon_{si} t_{ox}}[V_{ox}(0) - V_{ox}(y)] \\ &= \frac{\varepsilon_{ox}}{\varepsilon_{si} t_{ox}}[V(y) - V(0)]. \end{aligned} \tag{3.92}$$

Since $V(0) = V_{dsat}$ and $\mathscr{E}_y = -dV/dy$, substituting Eq. (3.92) into Eq. (3.91) yields

$$\frac{d^2V}{dy^2} = \frac{\varepsilon_{ox}}{\varepsilon_{si} t_{ox} x_j}[V(y) - V_{dsat}], \tag{3.93}$$

or

$$\frac{d^2V}{dy^2} = \frac{V(y) - V_{dsat}}{l^2}, \tag{3.94}$$

where the characteristic length l is given by

$$l = \sqrt{\frac{\varepsilon_{si}}{\varepsilon_{ox}} t_{ox} x_j} \approx \sqrt{3 t_{ox} x_j}. \tag{3.95}$$

Equation (3.94) is a linear, second-order differential equation which can be solved with the boundary conditions $V(0) = V_{dsat}$ and $\mathscr{E}_y(0) = -[dV/dy]_{y=0} = -\mathscr{E}_{sat}$:

$$V(y) = V_{dsat} + l\mathscr{E}_{sat}\sinh\left(\frac{y}{l}\right). \tag{3.96}$$

Mathematically, there is no unambiguous definition for $\mathscr{E}_{sat}$, the lateral field at the saturation point, since carriers do not reach saturation velocity until $\mathscr{E}_y = \infty$. In practice, ***carriers traveling close to the saturation velocity start moving away from the surface when the lateral field becomes appreciable compared to the vertical field***. A good choice for $\mathscr{E}_{sat}$ is a field strength on the order of or several times the critical field $\mathscr{E}_c$ defined by Eq. (3.71). For example, $\mathscr{E}_{sat} = 2\mathscr{E}_c = 2v_{sat}/\mu_{eff}$, which is on the order of 5×10^4 V/cm for electrons, has been used in the literature (Ko, 1982). This is a reasonable value, since the vertical field in a MOSFET device typically lies in the range of 10^5–10^6 V/cm.

3.2.3.3 PEAK FIELD AT THE DRAIN

Once $V(y)$ is known, ΔL can be found by solving $V(y = \Delta L) = V_{ds}$:

$$\Delta L = l\ln\left[\frac{V_{ds} - V_{dsat}}{l\mathscr{E}_{sat}} + \sqrt{\left(\frac{V_{ds} - V_{dsat}}{l\mathscr{E}_{sat}}\right)^2 + 1}\,\right]. \tag{3.97}$$

It is then straightforward to substitute ΔL into Eq. (3.85) or, more accurately, replace L with $L - \Delta L$ in Eq. (3.78), to obtain the source–drain current beyond saturation. From Eq. (3.96), the electric field along the channel is given by

$$\mathscr{E}_y(y) = -\frac{dV}{dy} = -\mathscr{E}_{sat}\cosh\left(\frac{y}{l}\right), \tag{3.98}$$

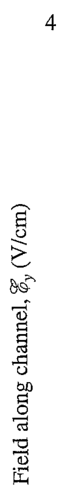
which increases exponentially toward the drain. An example is shown in Fig. 3.30. The peak field is reached at the drain, where

$$\mathscr{E}_{\max} \equiv \mathscr{E}_y(y = \Delta L) = -\sqrt{\left(\frac{V_{ds} - V_{dsat}}{l}\right)^2 + \mathscr{E}_{sat}^2}. \tag{3.99}$$

This field can be as high as mid-10^5 to 10^6 V/cm and is responsible for a variety of hot-carrier effects such as impact ionization, substrate current, and oxide degradation.

3.2.4 SOURCE–DRAIN SERIES RESISTANCE

In the discussion of MOSFET current thus far, it was assumed that the source and drain regions were perfectly conducting. In reality, as the current flows from the



FIGU
and d
5 × 1

chan
drair
long
with
serie
there
Tl
regic
chan

from
the s

Sinc

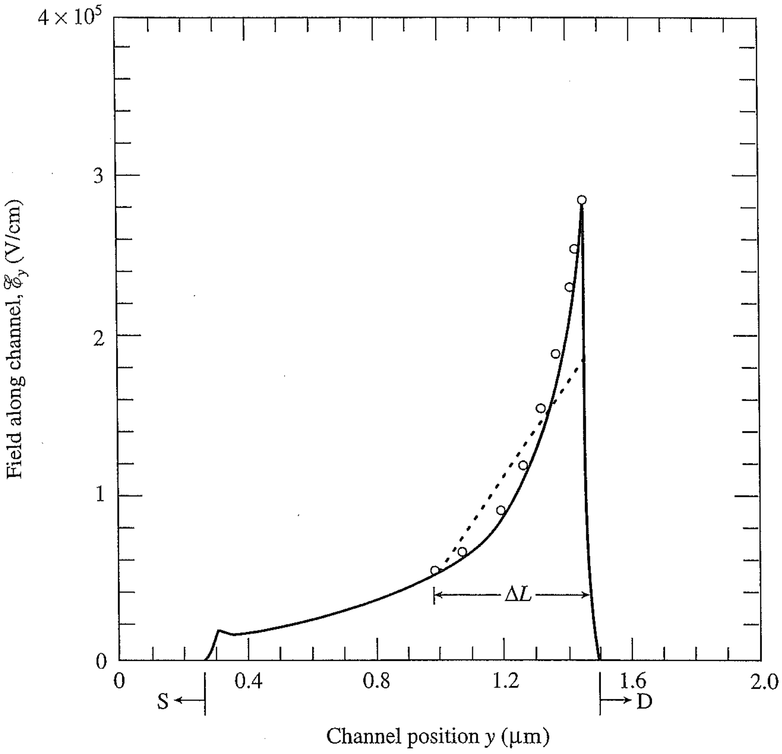


FIGURE 3.30. Calculated channel field versus distance between the source and drain. The velocity saturation region extends from a point where $\mathscr{E}_y \approx 5 \times 10^4$ V/cm to the drain. (After Ko *et al.*, 1981.)

channel to the terminal contact, there is a small voltage drop in the source and drain regions due to the finite silicon resistivity and metal contact resistance. In a long-channel device, the source–drain parasitic resistance is negligible compared with the channel resistance. In a short-channel device, however, the source–drain series resistance can be an appreciable fraction of the channel resistance and can therefore cause significant current degradation.

The most severe current degradation by series resistance occurs in the linear region (low V_{ds}) when the gate voltage is high. This is because the MOSFET channel resistance,

$$R_{ch} \equiv \frac{V_{ds}}{I_{ds}} = \frac{L}{\mu'_{eff} C_{ox} W (V_g - V_{on} - mV_{ds}/2)} \tag{3.100}$$

from Eq. (3.61), is the lowest under such bias conditions. It is instructive to estimate the sheet resistivity of a MOSFET channel,

$$\rho_{ch} \equiv R_{ch} \frac{W}{L} = \frac{1}{\mu'_{eff}(\varepsilon_{ox}/t_{ox})(V_g - V_{on} - mV_{ds}/2)} \approx \frac{1}{\mu_{eff}\varepsilon_{ox}\mathscr{E}_{ox}}. \tag{3.101}$$

Since the maximum oxide field $\mathscr{E}_{ox}$ is typically 2–5 MV/cm for most VLSI

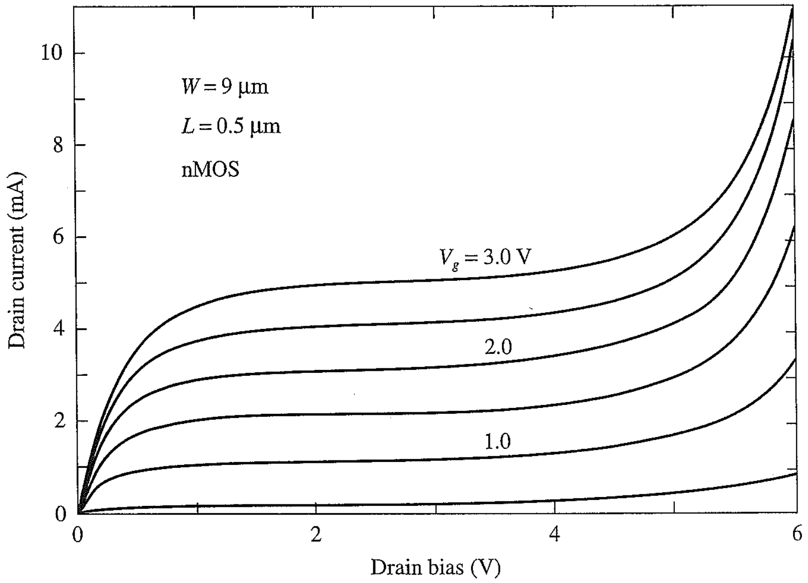


FIGURE 3.31. Example I_{ds}–V_{ds} curves of a short-channel nMOSFET showing breakdown at high drain bias voltages. (After Sun *et al.* 1987.)

technologies, the minimum channel sheet resistivity is about 2000 Ω/□ for nMOSFETs and 7000 Ω/□ for pMOSFETs.

The MOSFET current in the saturation region is least affected by the resistance degradation of source–drain voltage, since I_{ds} is essentially independent of V_{ds} in saturation. The saturation current is only affected through gate-voltage degradation by the voltage drop between the source contact and the source end of the channel (Fig. 4.20).

The effect of series resistance on linear I_{ds}–V_g curves used in channel-length extraction will be addressed in Section 4.3. Various contributions to the source–drain series resistance and their effect on circuit performance will be discussed in detail in Chapter 5.

3.2.5 MOSFET BREAKDOWN

Breakdown occurs in a short-channel MOSFET when the drain voltage exceeds a certain value, as shown in Fig. 3.31. It was discussed in Section 3.2.3 that the peak electric field given by Eq. (3.99) in the saturation region can attain large values at high drain voltages. When the field exceeds mid-10^5 V/cm, impact ionization (Section 2.4.1) takes place at the drain, leading to an abrupt increase of drain current. The breakdown voltage of nMOSFETs is usually lower than that of pMOSFETs because electrons have a higher rate of impact ionization (Fig. 2.37) and because n^+ source and drain junctions are more abrupt than p^+ junctions. There is also a weak dependence of the breakdown voltage on channel length; shorter devices have a lower breakdown voltage.

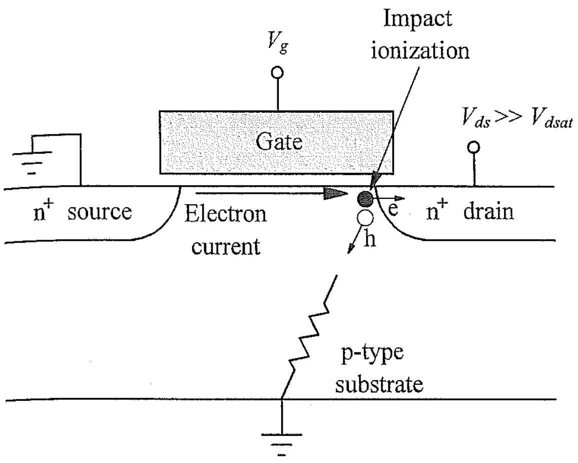


FIGURE 3.32. Schematic diagram showing impact ionization at the drain.

The breakdown process in an nMOSFET is shown schematically in Fig. 3.32. Electrons gain energy from the field as they move down the channel. Before they lose energy through collisions, they possess high kinetic energy and are capable of generating secondary electrons and holes by impact ionization. The generated electrons are attracted to the drain, adding to the drain current, while the holes are collected by the substrate contact, resulting in a substrate current. The substrate current in turn can produce a voltage (*IR*) drop from the spreading resistance in the bulk, which tends to forward-bias the source junction. This lowers the threshold voltage of the MOSFET and triggers a positive feedback effect, which further enhances the channel current. Substrate current is usually a good indicator of hot carriers generated by low-level impact ionization before runaway breakdown occurs.

Breakdown often results in permanent damage to the MOSFET as large amounts of hot carriers are injected into the oxide in the gate-to-drain overlap region. MOSFET breakdown is particularly a problem for VLSI technology during the elevated-voltage burn-in process. It can be relieved to some extent by using a lightly doped drain (LDD) structure (Ogura *et al.*, 1982), which introduces additional series resistance and reduces the peak field in a MOSFET. However, drain current and therefore device performance are traded off as a result. Ultimately, the devices should operate at a power-supply voltage far enough below the breakdown condition. This is one of the key CMOS design considerations in Chapter 4.

EXERCISES

3.1 It is commonly assumed that the surface potential ψ_s is pinned at $2\psi_B$ once the inversion layer is formed. In fact, ψ_s still rises slightly as the gate voltage and inversion charge density increase. Use Eq. (3.14) for $V = 0$ to show that a second-order correction term takes the form

$$\psi_s = 2\psi_B + \frac{2kT}{q}\ln\left(\frac{C_{ox}(V_g - V_{fb} - 2\psi_B)}{\sqrt{2\varepsilon_{si}kTN_a}}\right).$$

For $V_g = 5\ V$, $t_{ox} = 200$ Å, and $N_a = 10^{16}\ \text{cm}^{-3}$, show that ψ_s is about $8kT/q \approx 0.2$ V over $2\psi_B$.

3.2 Fill in the steps that lead to Eq. (3.29), the fraction of drift current component in the limit of $V \rightarrow 0$.

3.3 The effective field $\mathscr{E}_{eff}$ plays an important role in MOSFET channel mobility. Show that the definition

$$\mathscr{E}_{eff} \equiv \frac{\int_0^{x_i} n(x)\,\mathscr{E}(x)\,dx}{\int_0^{x_i} n(x)\,dx},$$

leads to Eq. (3.47), i.e., $\mathscr{E}_{eff} = (|Q_d| + |Q_i|/2)/\varepsilon_{si}$. Note that

$$|Q_i| = q \int_0^{x_i} n(x)\,dx$$

and

$$\mathscr{E}(x) = \frac{1}{\varepsilon_{si}}\left(|Q_d| + q\int_x^{x_i} n(x')\,dx'\right)$$

from Gauss's law. The inversion-layer depth x_i is assumed to be much smaller than the bulk depletion width.

3.4 An alternative threshold definition is based on the rate of change of inversion charge denisty with gate voltage. Equation (3.57) from Fig. 2.28(b) states that $d|Q_i|/dV_g$ is given by the serial combination of C_{ox} and $C_i \equiv d|Q_i|/d\psi_s$. Below threshold, $C_i \ll C_{ox}$, so that $d|Q_i|/dV_g \approx C_i$ and Q_i increases exponentially with V_g. Above threshold, $C_i \gg C_{ox}$, so that $d|Q_i|/dV_g \approx C_{ox}$ and Q_i increases linearly with V_g. The change of behavior occurs at an *inversion charge threshold voltage*, V_t^{inv}, where $C_i = C_{ox}$. Show that at $V_g = V_t^{inv}$ one has $d|Q_i|/dV_g = C_{ox}/2$ and $Q_i \approx (2kT/q)C_{ox}$. Note that such an inversion charge threshold is independent of depletion charge and is slightly higher than the conventional $2\psi_B$ threshold.

3.5 From Eq. (3.59) (neglecting the second term from inversion charge capacitance), show that the fractional loss of inversion charge due to the polysilicon depletion effect is $\Delta Q_i/Q_i \approx C_{ox}/2C_p$, where C_p is the small-signal polysilicon-depletion capacitance defined in Eq. (2.184). Explain why there is a factor-of-two difference between the loss of charge and the loss of capacitance.

3.6 *Charge-sharing model* (Yau, 1974): In Fig. 3.19, assume that both the source and drain depletion depths are equal to the gate depletion width W_{dm} (low-drain-bias condition) and that the junction curvatures under the gate edges are cylindrical. Show that

$$\frac{L+L'}{2L} = 1 - \frac{x_j}{L}\left(\sqrt{1+\frac{2W_{dm}}{x_j}} - 1\right).$$

In the linear region, the threshold voltage is largely determined by the total integrated depletion charge under the gate, instead of by the highest barrier in the channel as in the subthreshold regime. Show that the short-channel threshold roll-off is given by

$$\Delta V_t(\mathrm{SCE}) = \frac{qN_a W_{dm}}{C_{ox}} \left(\sqrt{1 + \frac{2W_{dm}}{x_j}} - 1 \right) \frac{x_j}{L}.$$

Note the $1/L$ dependence, in contrast with the exponential dependence in Eq. (3.66) for ΔV_t(SCE) under subthreshold conditions.

3.7 The small-signal transconductance in the saturation region is defined as $g_{msat} \equiv dI_{dsat}/dV_g$. Derive an expression for g_{msat} using Eq. (3.78) based on the $n = 1$ velocity saturation model. Show that g_{msat} approaches the saturation-velocity-limited value, Eq. (3.84), when $L \rightarrow 0$. What becomes of the expression for g_{msat} in the long-channel limit when $v_{sat} \rightarrow \infty$?

3.8 From Eq. (3.78) based on the $n = 1$ velocity saturation model, what is the carrier velocity at the source end of the channel? What are the limiting values when $L \rightarrow 0$ and when $v_{sat} \rightarrow \infty$?

3.9 Following a similar approach as in the text for the $n = 1$ velocity saturation model, derive an integral equation for the $n = 2$ velocity saturation model from which I_{ds} can be solved. It is very tedious to carry out the integration analytically (Taylor, 1984). Interested readers may attempt performing it numerically on a computer.

3.10 Assuming the $n = 1$ velocity saturation model, show that the total integrated inversion charge under the gate is

$$Q_i(\mathrm{total}) = -WLC_{ox}(V_g - V_t)\frac{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} + \frac{1}{3}}{\sqrt{1 + 2\mu_{eff}(V_g - V_t)/(mv_{sat}L)} + 1}$$

in the saturation region. Evaluate the intrinsic gate-to-channel capacitance, and show that it approaches Eq. (3.56) in the long-channel limit.

4 CMOS DEVICE DESIGN

This chapter examines the key device design issues in a modern CMOS VLSI technology. It begins with an extensive review of the concept of MOSFET scaling. Two important CMOS device design parameters, threshold voltage and channel length, are then discussed in detail.

4.1 MOSFET SCALING

CMOS technology evolution in the past twenty years has followed the path of device scaling for achieving density, speed, and power improvements. MOSFET scaling was propelled by the rapid advancement of lithographic techniques for delineating fine lines of 1 μm width and below. In Section 3.2.1, we discussed that reducing the source-to-drain spacing, i.e., the channel length of a MOSFET, led to short-channel effects. For digital applications, the most undesirable short-channel effect is a reduction in the gate threshold voltage at which the device turns on, especially at high drain voltages. Full realization of the benefits of the new high-resolution lithographic techniques therefore requires the development of new device designs, technologies, and structures which can be optimized to keep short-channel effects under control at very small dimensions. Another necessary technological advancement for device scaling is in ion implantation, which not only allows the formation of very shallow source and drain regions but also is capable of accurately introducing a sharply profiled, low concentration of doping atoms for optimum channel profile design.

4.1.1 CONSTANT-FIELD SCALING

In constant-field scaling (Dennard *et al.*, 1974), it was proposed that one can keep short-channel effects under control by scaling down the vertical dimensions (gate insulator thickness, junction depth, etc.) along with the horizontal dimensions, while also proportionally decreasing the applied voltages and increasing the substrate doping concentration (decreasing the depletion width). This is shown schematically in Fig. 4.1. ***The principle of constant-field scaling lies in scaling the device voltages***

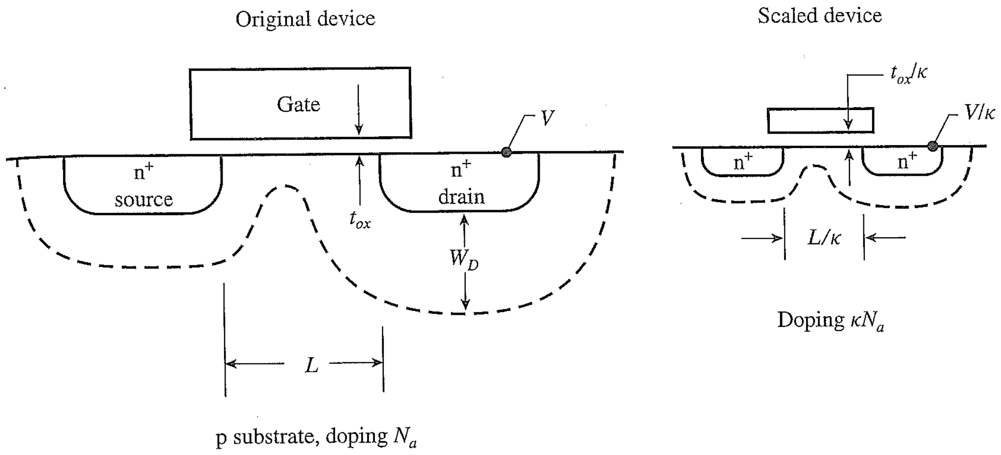


FIGURE 4.1. Principles of MOSFET constant-electric-field scaling. (After Dennard, 1986.)

and the device dimensions (both horizontal and vertical) by the same factor, κ (> 1), so that the electric field remains unchanged. This assures that the reliability of the scaled device is not worse than that of the original device.

4.1.1.1 RULES FOR CONSTANT-FIELD SCALING

Table 4.1 shows the scaling rules for various device parameters and circuit performance factors. The doping concentration must be increased by the scaling factor κ in order to keep Poisson's equation (3.64) invariant with respect to scaling. The

TABLE 4.1 Scaling of MOSFET Device and Circuit Parameters

	MOSFET Device and Circuit Parameters	**Multiplicative Factor ($\kappa > 1$)**
Scaling assumptions	Device dimensions (t_{ox}, L, W, x_j)	$1/\kappa$
	Doping concentration (N_a, N_d)	κ
	Voltage (V)	$1/\kappa$
Derived scaling behavior of device parameters	Electric field ($\mathscr{E}$)	1
	Carrier velocity (v)	1
	Depletion-layer width (W_d)	$1/\kappa$
	Capacitance ($C = \varepsilon A/t$)	$1/\kappa$
	Inversion-layer charge density (Q_i)	1
	Current, drift (I)	$1/\kappa$
	Channel resistance (R_{ch})	1
Derived scaling behavior of circuit parameters	Circuit delay time ($\tau \sim CV/I$)	$1/\kappa$
	Power dissipation per circuit ($P \sim VI$)	$1/\kappa^2$
	Power–delay product per circuit ($P\tau$)	$1/\kappa^3$
	Circuit density ($\propto 1/A$)	κ^2
	Power density (P/A)	1

maximum drain depletion width,

$$W_D = \sqrt{\frac{2\varepsilon_{si}(\psi_{bi} + V_{dd})}{qN_a}}, \tag{4.1}$$

from Eq. (2.70) scales down approximately by κ provided that the power-supply voltage V_{dd} is much greater than the built-in potential ψ_{bi}. All capacitances (including wiring load) scale down by κ, since they are proportional to area and inversely proportional to thickness. The charge per device ($\sim C \times V$) scales down by κ^2, while the inversion-layer charge density (per unit gate area), Q_i, remains unchanged after scaling. Since the electric field at any given point is unchanged, the carrier velocity ($v = \mu\mathscr{E}$) at any given point is also unchanged (the mobility is the same for the same vertical field). Therefore, any velocity saturation effects will be similar in the original and the scaled devices.

The drift current per MOSFET width, obtained by integrating the first term of the electron current density equation (2.43) over the inversion layer thickness, is

$$\frac{I_{drift}}{W} = Q_i v = Q_i \mu \mathscr{E}, \tag{4.2}$$

and is unchanged with respect to scaling. This means that the drift current scales down by κ, consistent with the behavior of both the linear and the saturation MOSFET currents in Eq. (3.19) and Eq. (3.23). A key implicit assumption is that the threshold voltage also scales down by κ. Note that the velocity saturated current, Eq. (3.78), also scales the same way, since both v_{sat} and μ_{eff} are constants, independent of scaling. However, the diffusion current per unit MOSFET width, obtained by integrating the second term of the current density equation (2.43) and given by

$$\frac{I_{diff}}{W} = D_n \frac{dQ_i}{dx} = \mu_n \frac{kT}{q}\frac{dQ_i}{dx}, \tag{4.3}$$

scales up by κ, since dQ_i/dx is inversely proportional to the channel length. Therefore, the diffusion current does not scale down the same way as the drift current. This has significant implications in the nonscaling of MOSFET subthreshold currents, as will be discussed in Section 4.1.3.

4.1.1.2 EFFECT OF SCALING ON CIRCUIT PARAMETERS

With both the voltage and the current scaled down by the same factor, it follows that the active channel resistance [e.g., Eq. (3.100)] of the scaled-down device remains unchanged. It is further assumed that parasitic resistance is either negligible or unchanged in scaling. The circuit delay, which is proportional to RC or CV/I, then scales down by κ. ***This is the most important conclusion of constant-field scaling: once the device dimensions and the power-supply voltage are scaled***

down, the circuit speeds up by the same factor. Moreover, power dissipation per circuit, which is proportional to VI, is reduced by κ^2. Since the circuit density has increased by κ^2, the power density, i.e., the active power per chip area, remains unchanged in the scaled-down device. This has important technological implications in that, in contrast to bipolar devices (Chapters 6, 7, and 8), packaging of the scaled CMOS devices does not require more elaborate heat-sinking. The power–delay product of the scaled CMOS circuit shows a dramatic improvement by a factor of κ^3 (Table 4.1).

4.1.1.3 THRESHOLD VOLTAGE

It was assumed earlier that the threshold voltage should be decreased by the scaling factor, κ, in proportion to the power-supply voltage. This is examined using the threshold equation (3.40) for a uniformly doped substrate:

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si} q N_a (2\psi_B + V_{bs})}}{C_{ox}}, \tag{4.4}$$

where V_{bs} is the substrate bias voltage. In silicon technology, the material-related parameters (energy gap, work function, etc.) do not change with scaling; hence, in general, V_t does not scale. However, in a conventional process, n$^+$-polysilicon gates are used for n-channel MOSFETs, and $V_{fb} = -E_g/2q - \psi_B$ from Eq. (2.181). It turns out that the first two terms on the RHS of Eq. (4.4) add up to approximately -0.15 V, which can be neglected. One can then argue (Dennard *et al.*, 1974) that by adjusting V_{bs} so that $2\psi_B + V_{bs}$ scales down by κ, the last term of Eq. (4.4), and therefore V_t, will also scale down by κ. However, at the present level of technology development, V_{bs} has been reduced to zero for most logic applications, though a reverse-biased body–source junction is still used for some dynamic memory array devices. Further reduction of the $2\psi_B + V_{bs}$ term with scaling would require a forward bias on the substrate. This is not commonly used in VLSI technologies, although it has been attempted in experimental devices (Sai-Halasz *et al.*, 1990). In practice, nonuniform doping profiles have been employed to tailor the threshold voltage of scaled devices, as will be discussed in Section 4.2.

p-channel MOSFETs with p$^+$-polysilicon gates scale similarly to their counterparts. However, in buried-channel devices, e.g., when an n$^+$-polysilicon gate is used for p-channel MOSFETs, the sum of the first two terms in Eq. (4.4) is nearly 1 V and therefore cannot be neglected. For this reason, it is difficult to scale buried-channel devices to low threshold voltages. More about threshold voltage design can be found in Section 4.2.

4.1.2 GENERALIZED SCALING

Even though constant-field scaling provides a basic guideline to the design of scaled MOSFETs, the requirement of reducing the voltage by the same factor as the

TABLE 4.2 CMOS VLSI Technology Generations

Feature Size (μm)	Power-Supply Voltage (V)	Gate Oxide Thickness (Å)	Oxide Field (MV/cm)
2	5	350	1.4
1.2	5	250	2.0
0.8	5	180	2.8
0.5	3.3	120	2.8
0.35	3.3	100	3.3
0.25	2.5	70	3.6

device physical dimension is too restrictive. Because of subthreshold nonscaling and reluctance to depart from the standardized voltage levels of the previous generation, the power-supply voltage was seldom scaled in proportion to channel length. Table 4.2 lists the supply voltage and device parameters of several generations of CMOS VLSI technology. It is clear that the oxide field has been increasing over the generations rather than staying constant. For device design purposes, therefore, it is necessary to develop a more general set of guidelines that allows the electric field to increase. In such a generalized scaling (Baccarani *et al.*, 1984), it is desired that both the vertical and the lateral electric fields change by the same multiplication factor so that the shape of the electric field pattern is preserved. This assures that 2-D effects, such as short-channel effects, do not become worse when scaling to a smaller dimension. Higher fields, however, do cause reliability concerns as mentioned in Section 2.4.

4.1.2.1 RULES FOR GENERALIZED SCALING

If we assume that the electric field intensity changes by a factor of α, i.e., $\mathscr{E} \rightarrow \alpha\mathscr{E}$, while the device physical dimensions (both lateral and vertical) scale down by κ (>1) in generalized scaling, the potential or voltage will change by a factor equal to the ratio α/κ. If $\alpha = 1$, it reduces back to constant-field scaling. To keep Poisson's equation invariant under the transformation, $(x, y) \rightarrow (x, y)/\kappa$ and $\psi \rightarrow \psi/(\kappa/\alpha)$ within the depletion region,

$$\frac{\partial^2(\alpha\psi/\kappa)}{\partial(x/\kappa)^2} + \frac{\partial^2(\alpha\psi/\kappa)}{\partial(y/\kappa)^2} = \frac{qN_a'}{\varepsilon_{si}}, \tag{4.5}$$

N_a' should be scaled to $(\alpha\kappa)N_a$. In other words, the doping concentration must be scaled up by an extra factor of α to control the depletion-region depth and thus avoid increased short-channel effects due to the higher electric field. Table 4.3 shows the generalized scaling rules of other device and circuit parameters.

Since the electric field intensity is usually increased in generalized scaling, the carrier velocity tends to increase as well. How much the velocity increases depends on how velocity-saturated the original device is. In the long-channel limit, carrier

TABLE 4.3 Generalized MOSFET Scaling

	MOSFET Device and Circuit Parameters	Multiplicative Factor ($\kappa > 1$)	
Scaling assumptions	Device dimensions (t_{ox}, L, W, x_j)	$1/\kappa$	
	Doping concentration (N_a, N_d)	$\alpha\kappa$	
	Voltage (V)	α/κ	
Derived scaling behavior of device parameters	Electric field ($\mathscr{E}$)	α	
	Depletion-layer width (W_d)	$1/\kappa$	
	Capacitance ($C = \varepsilon A/t$)	$1/\kappa$	
	Inversion-layer charge density (Q_i)	α	
		Long Ch.	Vel. Sat.
	Carrier velocity (v)	α	1
	Current, drift (I)	α^2/κ	α/κ
Derived scaling behavior of circuit parameters	Circuit delay time ($\tau \sim CV/I$)	$1/\alpha\kappa$	$1/\kappa$
	Power dissipation per circuit ($P \sim VI$)	α^3/κ^2	α^2/κ^2
	Power–delay product per circuit ($P\tau$)	α^2/κ^3	
	Circuit density ($\propto 1/A$)	κ^2	
	Power density (P/A)	α^3	α^2

velocities are far from saturation and will increase by the same factor, α, as the electric field. The drift current, which is proportional to WQ_iv, will then change by a factor of α^2/κ. This is consistent with the scaling behavior of long-channel currents, Eq. (3.19) and Eq. (3.23). On the other hand, if the original device is fully velocity-saturated, the carrier velocity cannot increase any more, in spite of the higher field in the scaled device. The current in this case will change only by a factor of α/κ, consistent with the velocity-saturated current, Eq. (3.80). The circuit delay scales down by a factor between κ and $\alpha\kappa$, depending on the degree of velocity saturation. The most serious issue with generalized scaling is the increase of the power density by a factor of α^2 to α^3. This puts a great burden on VLSI packaging technology to dissipate the extra heat generated on the chip. The power–delay product is also a factor of α^2 higher than for constant-field scaling.

4.1.2.2 CONSTANT-VOLTAGE SCALING

Even though Poisson's equation within the depletion region is invariant under generalized scaling, the same is not true in the inversion layer when mobile charges are present. This is because mobile charge densities are exponential functions of potential which do not scale linearly with either physical dimensions or voltage. Furthermore, even in the depletion region, not all the boundary conditions scale consistently under generalized scaling. ***This is due to the fact that the band bending at the source junction is given by the built-in potential (Appendix 6), which***

does not scale with voltage. Strictly speaking, the shape of the field pattern is preserved only if $\alpha = \kappa$, i.e., constant-voltage scaling. Under constant-voltage scaling, the electric field scales up by κ and the doping concentration N_a scales up by κ^2. The maximum gate depletion width (long-channel) [Eq. (3.69)],

$$W_{dm}^0 = \sqrt{\frac{4\varepsilon_{si}kT\ln(N_a/n_i)}{q^2N_a}}, \tag{4.6}$$

then scales down by κ. Here $\ln(N_a/n_i)$ is a weak function of N_a and can be treated as a constant. This allows the short-channel V_t roll-off [Eq. (3.66)],

$$\Delta V_t = \frac{24t_{ox}}{W_{dm}}\sqrt{\psi_{bi}(\psi_{bi}+V_{ds})}e^{-\pi L/2(W_{dm}+3t_{ox})}, \tag{4.7}$$

to remain unchanged, as both t_{ox} and W_{dm} are scaled down by the same factor as the channel length L. Both the power-supply voltage and the threshold voltage [Eq. (4.4)],

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{2\varepsilon_{si}qN_a(2\psi_B+V_{bs})}}{C_{ox}}, \tag{4.8}$$

also remain unchanged. From Eq. (2.166), the inversion-layer charge per unit area is related to the electron concentration at the silicon surface, $n(0)$, by

$$Q_i = \sqrt{2\varepsilon_{si}kTn(0)}. \tag{4.9}$$

Since Q_i scales up by κ in constant-voltage scaling, $n(0)$ scales up by κ^2. Therefore, the mobile charge density scales the same way as the fixed charge density N_a. The inversion-layer thickness, being proportional to $Q_i/qn(0)$, scales down by κ just like other linear dimensions. The Debye length, $L_D = (\varepsilon_{si}kT/q^2N_a)^{1/2}$, also scales down by κ under constant-voltage scaling.

Although constant-voltage scaling leaves the solution of Poisson's equation for the electrostatic potential unchanged except for a constant multiplicative factor in the electric field, it cannot be practiced without limit, since the power density increases by a factor of κ^2 to κ^3. Higher fields also cause hot-electron and oxide reliability problems. In reality, CMOS technology evolution has followed mixed steps of constant-voltage and constant-field scaling, as is evident in Table 4.2.

4.1.3 NONSCALING EFFECTS

4.1.3.1 PRIMARY NONSCALING FACTORS

From the above discussions, it is clear that although constant-field scaling provides a basic framework for shrinking CMOS devices to gain higher density and speed without degrading reliability and power, there are several factors that scale neither with the physical dimensions nor with the operating voltage. ***The primary reason for the nonscaling effects is that neither the thermal voltage kT/q nor the silicon bandgap E_g changes with scaling.*** The former leads to subthreshold nonscaling; i.e., the threshold voltage cannot be scaled down like other parameters. The latter leads to nonscalability of the built-in potential, depletion-layer width, and short-channel effect.

From Eq. (3.36), the *off current* of a MOSFET is given by

$$I_{ds}(V_g = 0, V_{ds} = V_{dd}) = \mu_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{-qV_t/mkT}. \tag{4.10}$$

Because of the exponential dependence, the threshold voltage cannot be scaled down significantly without causing a substantial increase in the off current. In fact, even if the threshold voltage is held unchanged, the off current per device still increases by a factor of κ (from the C_{ox} factor) when the physical dimensions are scaled down by κ. This imposes a serious limitation on how low the threshold voltage can be, especially in dynamic circuits and random-access memories. The threshold voltage limitation in turn sets a lower limit on the power-supply voltage V_{dd}, since the circuit delay increases rapidly with the ratio V_t/V_{dd} when the latter exceeds about 0.3, as will be discussed in Chapter 5.

Another nonscaling factor related to kT/q is the inversion-layer thickness, which is unchanged in constant-field scaling. Since the inversion-layer capacitance arising from the finite thickness is in series with the oxide capacitance, the total gate capacitance per unit area of the scaled device increases by a factor less than κ (Baccarani and Wordeman, 1983). This degrades the inversion charge density and therefore the current, especially at low gate voltages, as can be seen from Eq. (3.58).

Because both the junction built-in potential [Eq. (2.69)] and the maximum surface potential [Eq. (2.155)] are in the range of 0.6–1.0 V and do not change significantly with device scaling, the depletion-region widths, Eq. (4.1) and Eq. (4.6), do not scale quite as much as other linear dimensions. This results in worse short-channel effects in the scaled MOSFET, as is evident from Eq. (4.7). To compensate for these effects, the doping concentration must increase more than that suggested by constant-field scaling or generalized scaling.

4.1.3.2 SECONDARY NONSCALING FACTORS

Because of subthreshold nonscaling, the voltage level cannot be scaled down as much as the linear dimensions, and the electric field has increased as a result. This triggers several secondary nonscaling effects. First, in our discussions so far, it was implicitly assumed that carrier mobilities are constant, independent of scaling. However, as discussed in Section 3.1.5, the mobility decreases with increasing electric field:

$$\mu_{eff} \approx 32500 \mathscr{E}_{eff}^{-1/3}, \tag{4.11}$$

in units of cm^2/V-s for $\mathscr{E}_{eff} \leq 5 \times 10^5$ V/cm. Beyond $\mathscr{E}_{eff} = 5 \times 10^5$ V/cm, the mobility decreases even faster due to surface roughness scattering (Fig. 3.13). Since it is inevitable that the electric field increases with scaling, carrier mobilities are degraded in scaled MOSFETs. As a result, both the current and the delay improve less than the factors listed in Table 4.3 for generalized scaling. Furthermore, higher fields tend to push device operation more into the velocity-saturated regime. This means that the current gain and the delay improvement are closer to the velocity-saturated column of Table 4.3, and there is little to gain by operating at an even higher field or voltage.

The most serious problems associated with the higher field intensity are reliability and power. The power density increases by a factor of α^2 to α^3 as discussed before. Reliability problems arise from higher oxide fields, higher channel fields, and higher current densities. Even under the fully velocity-saturated condition, the current density increases by $\alpha\kappa$. This aggravates the problem of electromigration in aluminum lines, which is already becoming worse under constant-field scaling (Dennard *et al.*, 1974). Higher fields also drive gate oxides closer to the breakdown condition, making it difficult to maintain oxide integrity. In fact, in order to curb the growing oxide field, the gate oxide thickness has been reduced less than the lateral device dimensions, e.g., the channel length, as is evident in Table 4.2. This means that the channel doping concentration must be increased more than called for in Table 4.3 to keep short-channel effects [Eq. (4.7)] under control. In other words, the maximum gate depletion width W_{dm} must be reduced more than the oxide thickness t_{ox}. This triggers another set of nonscaling effects, including the subthreshold slope $\propto m = 1 + (3t_{ox}/W_{dm})$, and the substrate sensitivity $dV_t/dV_{bs} = m - 1$ [Eq. (3.41)]. These will be discussed in detail in Section 4.2.3.

4.1.3.3 OTHER NONSCALING FACTORS

In practice, there is yet another set of nonscaling factors encountered in CMOS technology evolution. One kind of nonscaling effects is related to the gate and source–drain doping levels. If not properly scaled up, they may lead to gate depletion and source–drain series resistance problems. From Eq. (2.185), polysilicon gate depletion contributes a capacitance $C_p = \varepsilon_{si} q N_p / Q_p$ in series with the oxide capacitance C_{ox}. As C_{ox} increases by a factor of κ while Q_p remains unchanged in constant-field scaling, N_p must scale up by κ also to keep C_p in step with C_{ox}. In generalized scaling, N_p must scale up even more (by $\alpha\kappa$). In reality, this is seldom done, for process reasons. The total gate capacitance then scales up by less than C_{ox}, leading to degradation of the inversion charge density and transconductance. Similarly, it is difficult to scale up the source–drain doping level and make the profile more abrupt while scaling down the junction depth. In practice, the source–drain series resistance has not been reduced in proportion to channel resistance, Eq. (3.100). This causes loss of current drive as the parasitic component becomes a more significant fraction of the total resistance in the scaled device.

Another class of nonscaling factors arise from process tolerances. The full benefit of scaling cannot be realized unless all process tolerances are reduced by the same factor as the device parameters. These include channel length tolerance, oxide thickness tolerance, threshold voltage tolerance, etc. ***It is a key requirement and challenge in VLSI technology development to keep the tolerance to a constant percentage of the device parameter as the dimension is scaled down.*** This could be a major factor in manufacturing costs as one tries to control a few hundred angstroms of channel length or a few atomic layers of gate oxide.

4.2 THRESHOLD VOLTAGE

This section focuses on a key design parameter in CMOS technology: threshold voltage. Although the threshold voltage was introduced in Chapter 3, the discussions there were restricted to the case of uniform doping. In this section, threshold voltages under nonuniform doping conditions are discussed, leading to the design of MOSFET channel profile.

4.2.1 THRESHOLD-VOLTAGE REQUIREMENT

4.2.1.1 VARIOUS DEFINITIONS OF THRESHOLD VOLTAGE

First, we examine the various definitions of threshold voltage and the threshold-voltage requirement from a technology point of view. There are quite a number of different ways to define the threshold voltage of a MOSFET device. In Chapter 3 we followed the most commonly used definition [$\psi_s(\text{inv}) = 2\psi_B$] of V_t. The advantage of this definition lies in its popularity and ease of incorporation into analytical solutions. However, it is not directly measurable from experimental I–V characteristics (it can be determined from a split C–V measurement; see Exercise 2.6). In Section 3.1.6, we introduced the linearly extrapolated threshold voltage, V_{on}, determined by the intercept of a tangent through the maximum-slope (linear transconductance) point of the low-drain I_{ds}–V_g curve. This is easily measured experimentally, but is about $3kT/q$ higher than the $2\psi_B$ threshold voltage, due to inversion-layer capacitance effects illustrated in Fig. 3.16.

Another commonly employed definition of threshold voltage is based on the subthreshold current, Eq. (3.36). For $V_{ds} > 2kT/q$ and $V_g < V_t$,

$$I_{ds}(V_g) = \mu_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT}{q}\right)^2 e^{q(V_g-V_t)/mkT}. \tag{4.12}$$

For a given constant current level I_0 (say, 50 nA/□), one can define a threshold voltage V_t^{sub} such that $I_{ds}(V_g = V_t^{sub}) = I_0(W/L)$. The advantages of such a threshold-voltage definition are twofold. First, it is easy to extract from hardware data and is therefore suitable for automated measurement of a large number of devices. Second, the device off current, $I_{off} = I_{ds}(V_g = 0)$, can be directly calculated from I_0, V_t^{sub}, and the subthreshold slope. However, there is a serious problem when the definition $I_{ds}(V_g = V_t^{sub}) = I_0(W/L)$ is used on a short-channel device whose exact channel length is not known. Even if the channel length is extracted from the currents in the linear region (Section 4.3.2), it is not necessarily the same channel length needed in Eq. (4.12) for the subthreshold region (Nguyen, 1984). This is because of the different current conduction mechanisms involved (diffusion versus drift). In subsequent discussions, we will adhere to the $2\psi_B$ definition of V_t. In general, V_t depends on temperature (temperature coefficient), substrate bias (body-effect coefficient), channel length, and drain voltage (short-channel effect, or SCE).

4.2.1.2 OFF-CURRENT REQUIREMENT AND MINIMUM THRESHOLD VOLTAGE

It is evident from Chapter 3 that the lower the threshold voltage, the higher the current drive, hence the faster the switching speed. From a CMOS performance point of view, it is desirable to have a threshold voltage as low as possible. This, of course, is counterbalanced by the off-current requirement that the MOSFET be turned off properly at $V_g = 0$. To meet the maximum-off-current requirement, one needs to consider the worst case when the threshold voltage is the lowest. Many tolerance factors must be taken into consideration in the worst-case design of a VLSI technology, for example, process tolerances (film thickness, implant dose, etc.), dimension tolerances (lithography, etching, etc.), operating temperature range, and bias conditions. Since the threshold voltage increases with substrate bias and decreases with temperature as discussed in Section 3.1.4, the worst-case condition is at zero substrate bias and maximum operating temperature, $T_{\max}$. Depending on the application, $T_{\max}$ is typically 100°C or so, driven by both environmental factors and heat dissipation of the VLSI chip in operation. From Eq. (4.12), the minimum threshold voltage for a maximum off current I_{off} at $T = T_{\max}$ is

$$V_{t\,\min} = \frac{mkT_{\max}}{q} \ln\left(\frac{I_{ds}(V_g = V_t)}{I_{off}}\right), \tag{4.13}$$

where

$$I_{ds}(V_g = V_t) = \mu_{eff} C_{ox} \frac{W}{L}(m-1)\left(\frac{kT_{\max}}{q}\right)^2 \tag{4.14}$$

is the drain current at threshold voltage. Note that $I_{ds}(V_g = V_t)$ is rather insensitive to temperature, since $\mu_{eff} \propto T_{\max}^{-3/2}$. However, it does depend on technology. For example, $I_{ds}(V_g = V_t)/W$ varies from 10^{-8} A/µm for 1-µm CMOS technology to 10^{-6} A/µm for 0.1-µm CMOS technology. (Note that these numbers are for nMOSFETs; pMOSFET currents are about 3 times lower. Also note that the extrapolated subthreshold currents at the linearly extrapolated threshold voltage V_{on} are about 10 times higher than these numbers, as discussed at the end of Section 3.1.6.). Taking the median value of 10^{-7} A/µm for $I_{ds}(V_g = V_t)/W$ and assuming a worst-case off-current requirement of $I_{off}/W = 10^{-8}$ A/µm, one obtains $V_{t\,\mathrm{min}} \approx 0.1$ V for $T_{\mathrm{max}} = 100$°C and $m = 1.3$. For a given I_{off}/W, $V_{t\,\mathrm{min}}$ increases as the channel length is scaled down. However, this is opposite to the downward scaling trend of the power-supply voltage. As a result, the device designer often encounters a tradeoff between the performance and the off-current requirement in scaled CMOS technologies (Mii *et al.*, 1994).

The above figures are acceptable for CMOS logic technologies. In a dynamic memory technology, however, the off-current requirement is much more stringent for the access transistor in the cell: on the order of $I_{off}/W = 10^{-12}$ A/µm or so (Dennard, 1984). This means $V_{t\,\mathrm{min}} \approx 0.5$ V for the DRAM access device. It should be understood that Eq. (4.12) is an analytical expression based on some simplifying approximations. It is used here for the purpose of illustration. More exact values of off current for a particular design should be obtained from numerical simulations.

4.2.1.3 THRESHOLD-VOLTAGE TOLERANCES

Equation (4.13) gives the minimum threshold voltage at the highest operating temperature for the worst-case process conditions. To obtain the nominal design threshold voltage at room temperature, one needs to add ΔV_t due to the temperature difference as well as the sum of all V_t tolerances to $V_{t\,\mathrm{min}}$. In other words,

$$V_t(\text{nominal, }23^\circ\text{C}) = V_{t\,\mathrm{min}} + \Delta V_t(\text{temp.}) + \sqrt{[\Delta V_t(\text{SCE})]^2 + [\Delta V_t(\text{process})]^2}, \tag{4.15}$$

where ΔV_t(SCE) and ΔV_t(process) are the 3σ tolerances of V_t reduction from short-channel effects and from process variations, respectively. They are added in an root-mean-square (RMS) fashion, since there are no correlations between them. ΔV_t(temp.) is about 55–75 mV, since $dV_t/dT \approx -0.7$ to -1 mV/°C, insensitive to technology (Section 3.1.4).

An approximate expression for ΔV_t(SCE) is given by Eq. (3.68) for the worst-case high drain bias:

$$\Delta V_t(\text{SCE}) = 8(m-1)\sqrt{\psi_{bi}(\psi_{bi} + V_{ds})}\,e^{-\pi L_{\mathrm{min}}/2mW_{dm}}, \tag{4.16}$$

where $m \approx 1 + (3t_{ox}/W_{dm})$, and L_{min} is the 3σ worst-case channel length of the manufacturing process. The preexponential factor in Eq. (4.16) does not scale much with technology, since ψ_{bi} is largely determined by the silicon bandgap and does not vary significantly with either device dimension or supply voltage. From 1-µm CMOS technology with $\psi_{bi} \approx 0.7$ V and $V_{ds} = 5$ V to 0.1-µm CMOS technology

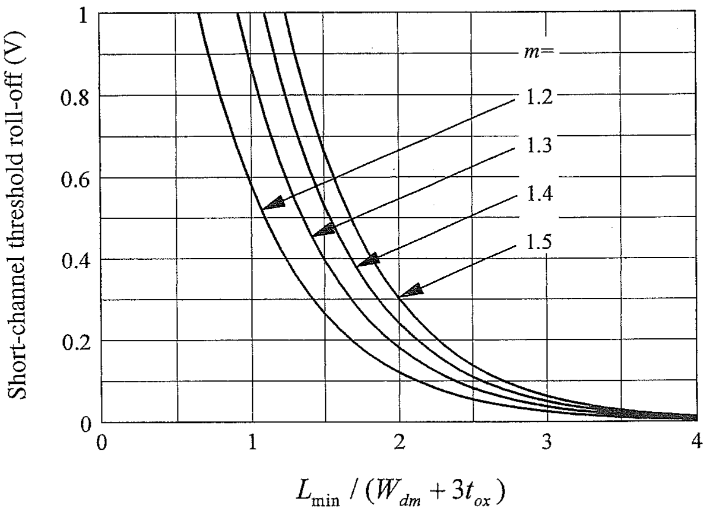


FIGURE 4.2. Short-channel threshold voltage roll-off, $\Delta V_t(\text{SCE})$, versus $L_{\min}/mW_{dm}$ or $L_{\min}/(W_{dm} + 3t_{ox})$ for several typical values of m, based on an approximate analytical expression.

with $\psi_{bi} \approx 0.94$ V and $V_{ds} = 1.5$ V, the square-root factor in Eq. (4.16) changes only slightly, from 2.0 to 1.5 V. Taking an average value of 1.75 V, one can plot $\Delta V_t(\text{SCE})$ as a function of $L_{\min}/mW_{dm}$, or equivalently, as a function of $L_{\min}/(W_{dm} + 3t_{ox})$, as in Fig. 4.2, for several possible values of m. Because of the exponential factor in Eq. (4.16), $\Delta V_t(\text{SCE})$ is very sensitive to $L_{\min}/mW_{dm}$. ***A good choice of $L_{\min}/mW_{dm}$ is 2***, which gives $\Delta V_t(\text{SCE}) \approx 0.2$ V for a median value of $m = 1.3$. Lower values of $L_{\min}/mW_{dm}$ result in too severe a short-channel effect or large $\Delta V_t(\text{SCE})$. Higher values of $L_{\min}/mW_{dm}$ improve the short-channel effect but also raise the junction capacitance (smaller W_{dm} for a given $L_{\min}$) or increase the oxide field (smaller t_{ox} for a given V_{dd}). The last quantity in Eq. (4.15), $\Delta V_t(\text{process})$ due to t_{ox} and N_a variations, can be estimated from the threshold equation (3.20). In most cases, 3σ variations of t_{ox} and N_a can be controlled to within 5–10% of their nominal values. Therefore, $\Delta V_t(\text{process})$ should be less than 10% of V_t, or less than about 50 mV, and $[\Delta V_t(\text{process})]^2$ can be neglected in comparison with $[\Delta V_t(\text{SCE})]^2$ in Eq. (4.15). The threshold voltage requirements in above examples are then $V_t(\text{nominal, } 23°\text{C}) \approx 0.4$ V for logic technologies and ≈ 0.8 V for DRAM technologies.

Another consideration that may further limit how low the threshold voltage can be is the *burn-in* procedure. Burn-in is required in most VLSI technologies to remove early failures and ensure product reliability. It is usually carried out at elevated temperatures and overvoltages to accelerate the degradation process. Both of these conditions further lower the threshold voltage and aggravate the leakage currents. Ideally, the burn-in procedure should be designed so that it does not require a compromise on the device performance.

4.2.2 NONUNIFORM DOPING

When the MOSFET channel is uniformly doped, the maximum gate depletion width (long-channel),

$$W_{dm}^0 = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a}}, \tag{4.17}$$

and the threshold voltage,

$$V_t = V_{fb} + 2\psi_B + \frac{\sqrt{4\varepsilon_{si}qN_a\psi_B}}{C_{ox}}, \tag{4.18}$$

are coupled through the parameter N_a, and therefore cannot be varied independently (for a given V_{fb} and t_{ox}). It was discussed in the previous subsection that in order to control short-channel effects, W_{dm} or W_{dm}^0 should be on the order of $L_{\min}/(2m)$. The doping concentration that satisfies this requirement may not give the desired threshold voltage that satisfies the off-current requirement. Nonuniform channel doping gives the device designer an additional degree of freedom to tailor the profile for meeting both requirements. Such an optimization is made possible by the ion implantation technology.

4.2.2.1 INTEGRAL SOLUTION TO POISSON'S EQUATION

In this subsection, the surface potential, electric field, and threshold voltage for the case of nonuniform channel doping are solved for under the depletion approximation. Mathematically, a general expression can be derived as follows. For a nonuniform p-type doping profile $N(x)$, the electric field is obtained by integrating Poisson's equation once (neglecting mobile carriers in the depletion region):

$$\mathscr{E}(x) = \frac{q}{\varepsilon_{si}}\int_x^{W_d} N(x)\,dx, \tag{4.19}$$

where W_d is the depletion-layer width. Integrating again gives the surface potential,

$$\psi_s = \frac{q}{\varepsilon_{si}}\int_0^{W_d}\int_x^{W_d} N(x')\,dx'\,dx. \tag{4.20}$$

Using integration by parts, one can show that Eq. (4.20) is equivalent to (Brews, 1979)

$$\psi_s = \frac{q}{\varepsilon_{si}}\int_0^{W_d} xN(x)\,dx. \tag{4.21}$$

The maximum depletion-layer width (long-channel) W_{dm}^0 is determined by the condition $\psi_s = 2\psi_B$ when $W_d = W_{dm}^0$. ***The threshold voltage of a nonuniformly doped MOSFET is then determined by both the integral (depletion charge density) and the first moment of N(x) within (0, W_{dm}^0).***

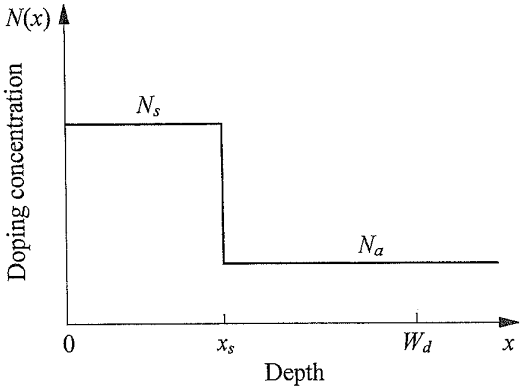


FIGURE 4.3. A schematic diagram showing the high–low step doping profile. $x = 0$ denotes the silicon–oxide interface.

4.2.2.2 A HIGH–LOW STEP PROFILE

Consider the idealized step doping profile shown in Fig. 4.3 (Rideout *et al.*, 1975). It can be formed by making one or more low-dose, shallow implants into a uniformly doped substrate of concentration N_a. After drive-in, the implanted profile is approximated by a region of constant doping N_s that extends from the surface to a depth x_s. If the entire depletion region at the threshold condition is contained within x_s, the MOSFET can be considered as uniformly doped with a concentration N_s. The case of particular interest analyzed here is when the depletion width W_d exceeds x_s, so that part of the depletion region has a charge density N_s and part of it N_a. The integration in Eq. (4.21) can be easily carried out for this profile to yield the surface potential, or the band bending at the surface,

$$\psi_s = \frac{qN_s}{2\varepsilon_{si}}x_s^2 + \frac{qN_a}{2\varepsilon_{si}}\left(W_d^2 - x_s^2\right). \tag{4.22}$$

Inversely, the depletion layer width W_d can be found as a function of the surface potential ψ_s:

$$W_d = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(\psi_s - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}. \tag{4.23}$$

This is less than the depletion width in the uniformly doped (N_a) case for the same surface potential. The electric field at the surface is obtained by evaluating the integral in Eq. (4.19) with $x = 0$:

$$\mathscr{E}_s = \frac{qN_s x_s}{\varepsilon_{si}} + \frac{qN_a(W_d - x_s)}{\varepsilon_{si}}. \tag{4.24}$$

From Gauss's law, the total depleted charge per unit area in silicon is given by

$$Q_s = -\varepsilon_{si}\mathscr{E}_s = -qN_s x_s - qN_a(W_d - x_s), \tag{4.25}$$

as would be expected from Fig. 4.3. ***The effect of the nonuniform surface doping***

is then to increase the depletion charge within $0 \le x \le x_s$ by $(N_s - N_a)x_s$ and, at the same time, reduce the depletion layer width as indicated by Eq. (4.23).

For an applied gate voltage V_g, the quantities ψ_s and Q_s are related by Eq. (3.14):

$$V_g = V_{fb} + \psi_s - \frac{Q_s}{C_{ox}} = V_{fb} + \psi_s + \frac{qN_s x_s + qN_a(W_d - x_s)}{C_{ox}}. \tag{4.26}$$

Substituting Eq. (4.23) for W_d yields

$$V_g = V_{fb} + \psi_s + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si} q N_a \left(\psi_s - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} \tag{4.27}$$

$$+ \frac{q(N_s - N_a)x_s}{C_{ox}}.$$

By definition, the threshold voltage is the gate voltage at which $\psi_s = 2\psi_B$, i.e.,

$$V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si} q N_a \left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)} \tag{4.28}$$

$$+ \frac{q(N_s - N_a)x_s}{C_{ox}}.$$

The maximum depletion width (long-channel) at threshold is given by Eq. (4.23) with $\psi_s = 2\psi_B$:

$$W_{dm}^0 = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)}. \tag{4.29}$$

There is some ambiguity as to whether $2\psi_B$ is defined in terms of N_s or N_a. We adopt the convention that $2\psi_B$ is defined in terms of the p-type concentration at the depletion-layer edge, i.e., $2\psi_B = (2kT/q)\ln(N_a/n_i)$. In fact, it makes very little difference which concentration we use, since $2\psi_B$ is a rather weak function of the doping concentration anyway. Further refinement of the threshold condition would require a numerical simulation of the specific profile.

In Section 3.1.3, we showed that the subthreshold slope is given by $2.3mkT/q$, where $m = dV_g/d\psi_s$ at $\psi_s = 2\psi_B$. Here m is also referred to as the body-effect coefficient, $1 + (C_{dm}/C_{ox})$, defined in Eq. (3.22). In the nonuniformly doped case, m can be evaluated from Eq. (4.27):

$$m \equiv \frac{dV_g}{d\psi_s}(\psi_s = 2\psi_B) \tag{4.30}$$

$$= 1 + \frac{\sqrt{2\varepsilon_{si} q N_a}}{2C_{ox}}\left(2\psi_B - \frac{q(N_s - N_a)x_s^2}{2\varepsilon_{si}}\right)^{-1/2}.$$

It can be expressed in terms of W_{dm}^0 using Eq. (4.29):

$$m = 1 + \frac{\varepsilon_{si}/W_{dm}^0}{C_{ox}} = 1 + \frac{C_{dm}}{C_{ox}} = 1 + \frac{3t_{ox}}{W_{dm}^0}. \tag{4.31}$$

These expressions are consistent with Eq. (3.67) for a uniformly doped channel. Similarly, the threshold voltage in the presence of a substrate bias $-V_{bs}$ is given by Eq. (4.28) with the $2\psi_B$ term in the square root replaced by $2\psi_B + V_{bs}$. Using Eq. (4.29), one can show that the substrate sensitivity is

$$\frac{dV_t}{dV_{bs}} = \frac{\varepsilon_{si}/W_{dm}^0}{C_{ox}} = \frac{C_{dm}}{C_{ox}} = m - 1. \tag{4.32}$$

Therefore, ***all the previous expressions for the depletion capacitance, subthreshold slope, and body-effect coefficient in terms of W_{dm}^0 for the uniformly doped case remain valid for the nonuniformly doped case***. The only difference is that the maximum depletion layer width W_{dm}^0 in the nonuniformly doped case is given by Eq. (4.29) instead of Eq. (4.17).

4.2.2.3 GRAPHICAL INTERPRETATION

The above solutions of potential, field, and threshold voltage for a nonuniformly doped channel are best illustrated graphically by plotting the electric field versus depth as shown in Figs. 4.4 and 4.5. We start with the uniformly doped case in Fig. 4.4(a), where $\mathscr{E}(x)$ is a straight line with a negative slope whose magnitude is proportional to the substrate doping concentration N_a. The x-intercept gives the depletion-layer width where $\mathscr{E} = 0$. The y-intercept gives the surface electric field $\mathscr{E}_s$, which from Gauss's law is proportional to the total depletion charge per unit area. Since $\mathscr{E} = -d\psi/dx$, the triangular area under $\mathscr{E}(x)$ equals the surface potential, or the band bending ψ_s. As the gate voltage increases, both W_d and $\mathscr{E}_s$ increase, and so does ψ_s until it reaches $2\psi_B$. At this point, surface inversion occurs and the depletion-layer width has reached its maximum value. The threshold voltage is largely determined by the y-intercept or the surface field $\mathscr{E}_s$ when $\psi_s = 2\psi_B$, since

$$V_t = V_{fb} + 2\psi_B + \frac{\varepsilon_{si}\mathscr{E}_s}{C_{ox}}, \tag{4.33}$$

and the first two terms on the RHS nearly cancel each other for n$^+$-polysilicon gates on nMOSFETs and vice versa, as discussed in Section 4.1.1. For a lower N_a, the magnitude of the $\mathscr{E}(x)$ slope decreases, and therefore V_t decreases while W_{dm}^0 increases as depicted in Fig. 4.4(b). The two triangular areas under the different $\mathscr{E}(x)$ lines are approximately the same at the threshold condition, since $2\psi_B$ is a rather weak function of N_a and can be considered as a constant for practical purposes.

Figure 4.5(a) shows an $\mathscr{E}(x)$ plot for the nonuniformly doped case of a high–low step profile discussed above. With a higher doping N_s in the surface

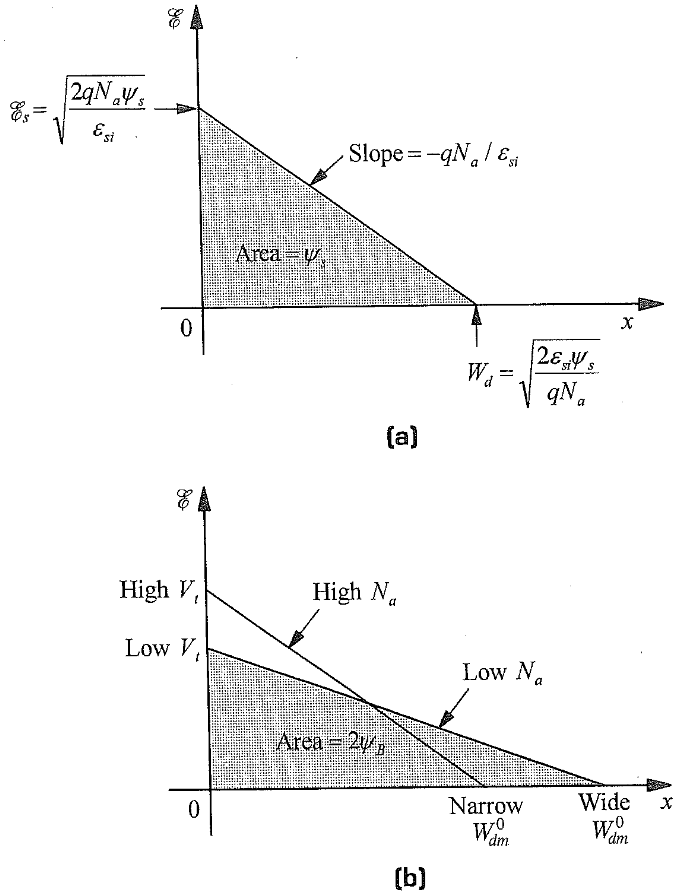

$\mathscr{E}$
$\mathscr{E}_s = \sqrt{\frac{2qN_a\psi_s}{\varepsilon_{si}}}$
Slope $= -qN_a / \varepsilon_{si}$
Area $= \psi_s$
0
x
$W_d = \sqrt{\frac{2\varepsilon_{si}\psi_s}{qN_a}}$
(a)
$\mathscr{E}$
High V_t
High N_a
Low V_t
Low N_a
Area $= 2\psi_B$
0
Narrow W_{dm}^0
Wide W_{dm}^0
x
(b)

FIGURE 4.4. Graphical interpretation of relationships among doping concentration, depletion width, surface potential, and threshold voltage. (a) Uniformly doped case. The shaded area equals the surface potential ψ_s. (b) Two uniformly doped cases at $\psi_s = 2\psi_B$, one with a lower doping than the other. Both triangular areas equal $2\psi_B$ at threshold. V_t is directly related to the y-intercept.

region $0 \le x \le x_s$, the function $\mathscr{E}(x)$, instead of being a straight line, changes slope at $x = x_s$. At threshold, the shaded area under the two-sloped line $\mathscr{E}(x)$ equals $2\psi_B$ just as in the uniformly doped case. In other words, the shaded area in Fig. 4.5(a) equals the triangular area under the straight line $\mathscr{E}(x)$, but the x- and y-intercepts change quite differently from the uniformly doped cases depicted in Fig. 4.4(b). How much V_t shifts in response to a given change of W_{dm}^0 depends on where x_s is. If x_s is very close to the surface and N_s is high, it is possible to increase V_t with very little or almost no change in W_{dm}^0. ***This additional degree of freedom allows V_t and W_{dm}^0 to decouple from the uniformly doped relations (4.17) and (4.18).*** Mathematically, the threshold shift and the change of depletion width for the nonuniformly doped case are given by Eq. (4.28) and Eq. (4.29).

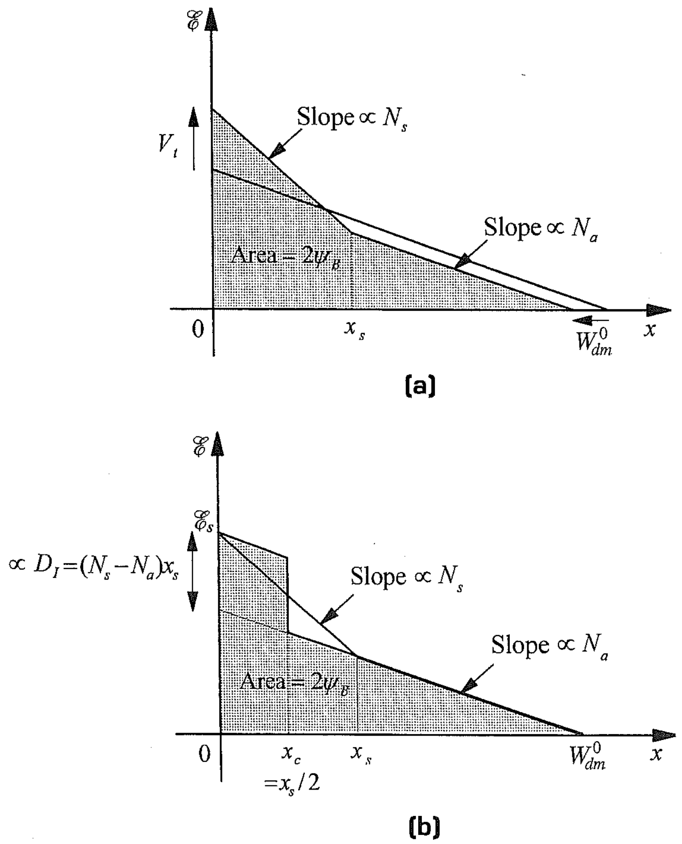


FIGURE 4.5. Graphical interpretation for nonuniformly doped cases. (a) High–low step profile compared with a uniformly doped profile (N_a). (b) A delta-function profile equivalent to the high–low profile in (a).

4.2.2.4 GENERALIZATION TO A GAUSSIAN PROFILE

Using the graphical representation in Fig. 4.5(b), one can show that the nonuniform step doping profile discussed above is equivalent to the delta-function profile shown in Fig. 4.6 with an equivalent dose of

$$D_I = (N_s - N_a)x_s \tag{4.34}$$

centered at $x_c = x_s/2$. This is because both the area under $\mathscr{E}(x)$ and the y-intercept (i.e., $\mathscr{E}_s$ or V_t) are identical between the two cases. The same result follows from Eq. (4.19) and Eq. (4.21).

Similar arguments apply to a general Gaussian (or other symmetric) profile with a dopant distribution,

$$N(x) = \frac{D_I}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-x_c)^2}{2\sigma^2}\right), \tag{4.35}$$

where σ is the implant straggle. The effect of such an implanted profile on threshold voltage and depletion-layer width is equivalent to that of the step doping profile

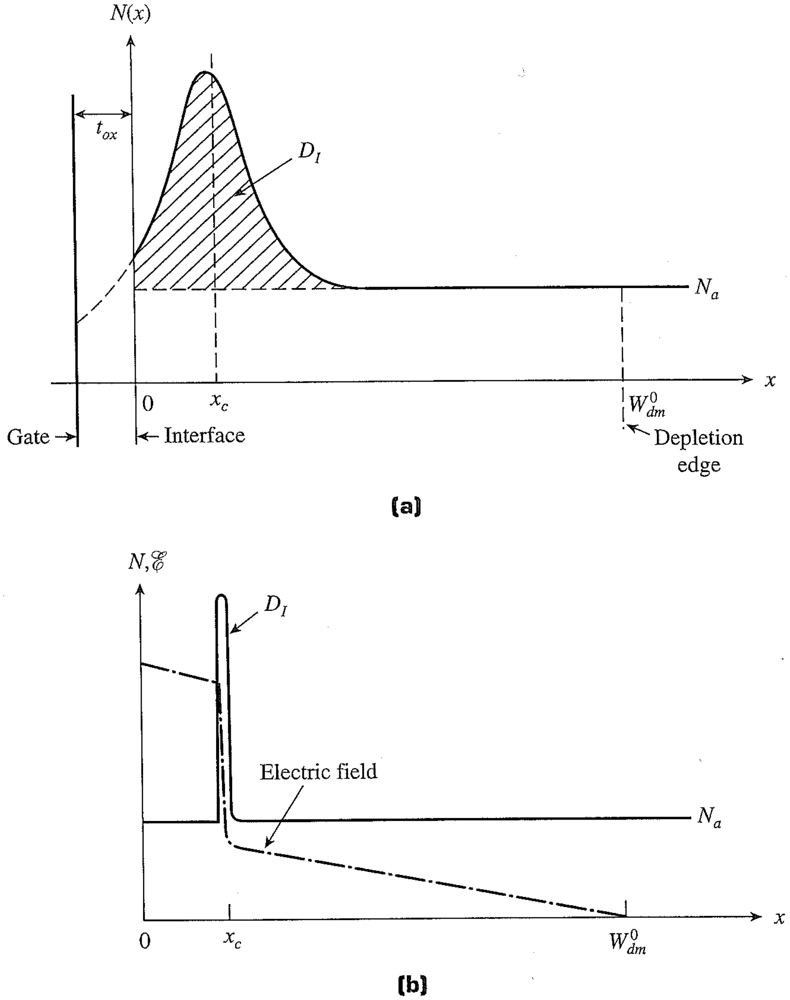


FIGURE 4.6. Schematic diagrams showing (a) an implanted Gaussian profile and (b) a delta-function profile equivalent to (a). (After Brews, 1979.)

discussed above, independent of σ. Substituting Eq. (4.34) and $x_c = x_s/2$ into the threshold voltage equation (4.28) yields

$$V_t = V_{fb} + 2\psi_B + \frac{1}{C_{ox}}\sqrt{2\varepsilon_{si} q N_a\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_{si}}\right)} + \frac{qD_I}{C_{ox}}. \tag{4.36}$$

Similarly, the maximum depletion width, Eq. (4.29), becomes

$$W^0_{dm} = \sqrt{\frac{2\varepsilon_{si}}{qN_a}\left(2\psi_B - \frac{qD_I x_c}{\varepsilon_{si}}\right)}. \tag{4.37}$$

For a given implanted dose D_I, the resulting threshold voltage shift depends on the location of the implant, x_c. ***For shallow surface implants, $x_c = 0$, there is no change in the depletion width. The V_t shift is simply given by qD_I/C_{ox}, as with a sheet of charge at the silicon–oxide interface.*** All other device parameters, e.g., substrate sensitivity and subthreshold slope, remain unchanged. As x_c increases for a given dose, both the maximum depletion width and the V_t shift decrease. However, if x_c is not too large, one can always readjust the background doping N_a to a lower value N_a' to restore W_{dm}^0 to its original value. The threshold voltage, in the meantime, is shifted by an amount somewhat less than the shallow implant case.

Although the above analysis on nonuniform doping assumes $N_s > N_a$, the results remain equally valid if $N_s < N_a$. Such a profile is referred to as the *retrograde channel doping* and will be discussed in detail in the next subsection.

4.2.3 CHANNEL PROFILE DESIGN

4.2.3.1 CMOS DESIGN CONSIDERATIONS

CMOS device design involves choosing a set of parameters that are coupled to a variety of circuit characteristics to be optimized. The choice of these device parameters is further subject to technology constraints and system compatibility requirements. Figure 4.7 shows a schematic diagram of the design process and the parameters involved. Because various circuit characteristics are interrelated through the device parameters, tradeoffs among them are often necessary. For example, reduction of W_{dm} improves the short-channel effects, but degrades the substrate

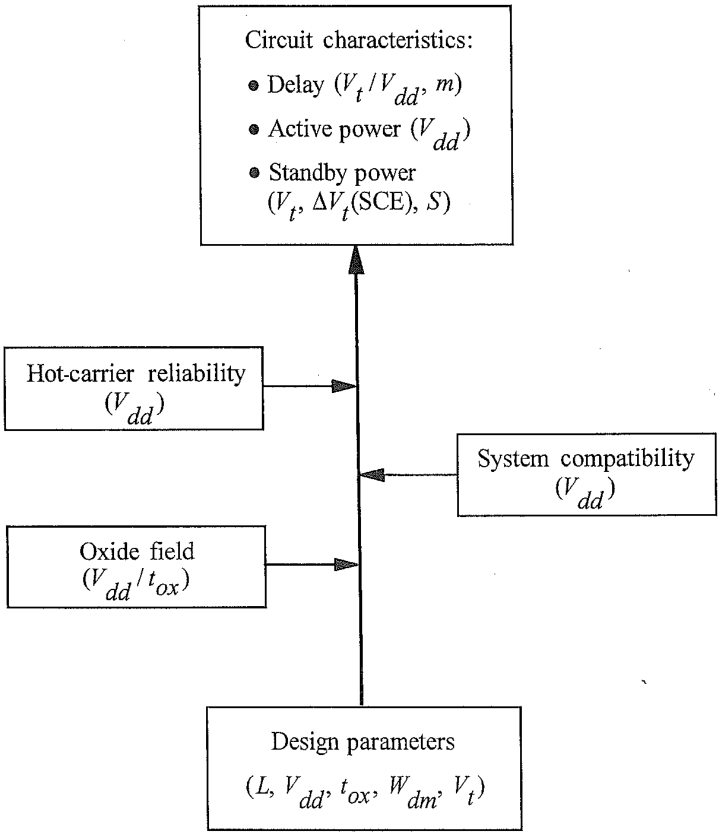

Circuit characteristics:
• Delay (V_t/V_{dd}, m)
• Active power (V_{dd})
• Standby power (V_t, ΔV_t(SCE), S)
Hot-carrier reliability (V_{dd})
System compatibility (V_{dd})
Oxide field (V_{dd}/t_{ox})
Design parameters (L, V_{dd}, t_{ox}, W_{dm}, V_t)

FIGURE 4.7. A CMOS design flowchart showing device parameters, technology constraints, and circuit objectives.

sensitivity; thinner t_{ox} increases the current drive, but causes reliability concerns, etc. There is no unique way of designing CMOS devices for a given technology generation. Nevertheless, we attempt here to give a general guideline of how these device parameters should be chosen.

It was discussed in Section 4.2.1 that in order to keep short-channel effects under control, a good choice of the maximum gate depletion width W_{dm} is such that $L_{\min}/mW_{dm} \approx 2$, where $L_{\min}$ is the minimum channel length. Since $L_{\min}$ is some fraction shorter than the nominal channel length L of the technology, this requirement can be stated as $L/mW_{dm} > 2$, or equivalently, as $W_{dm} + 3t_{ox} < L/2$. Here the body-effect coefficient is $m = 1 + (3t_{ox}/W_{dm}^0) \approx 1 + (3t_{ox}/W_{dm})$ from Eq. (4.31), for either a uniformly doped or a nonuniformly doped channel. Since both the subthreshold slope, $2.3mkT/q$, and the substrate sensitivity, $dV_t/dV_{bs} = m - 1$, degrade with higher m, m should be kept close to 1. A larger m also results in a lower saturation current in the long-channel limit [Eq. (3.23)]. Typically, one requires $m < 1.5$, or $3t_{ox}/W_{dm} < \frac{1}{2}$. These design considerations are illustrated in Fig. 4.8. A lower limit on t_{ox} imposed by technology constraints is $V_{dd}/\mathscr{E}_{ox}^{\max}$, where $\mathscr{E}_{ox}^{\max}$ is the maximum oxide field. ***For a given L and V_{dd}, the allowable parameter space in the t_{ox}–W_{dm} design plane is a triangular area bounded by requirements on the SCE, oxide field, and subthreshold slope (also substrate sensitivity).***

4.2.3.2 TRENDS OF POWER-SUPPLY VOLTAGE AND THRESHOLD VOLTAGE

Channel profile design is largely dictated by threshold-voltage requirements. The lower limit of threshold voltage is given by off-current specifications outlined in Section 4.2.1: $V_t \geq 0.4$ V. The upper limit of threshold voltage is imposed by circuit delay or performance considerations. It will be shown in Chapter 5 that CMOS delay degrades rapidly once V_t exceeds 25% of V_{dd}. Therefore, one should keep $V_t \leq V_{dd}/4$ if possible (Mii *et al.*, 1994). Figure 4.9 shows the trends in

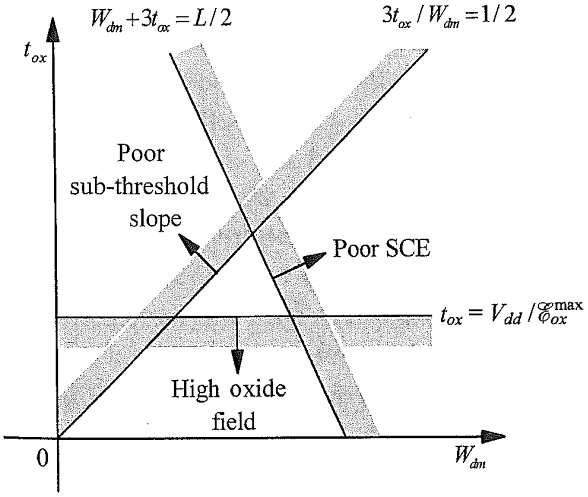


FIGURE 4.8. The t_{ox}–W_{dm} design plane. Some tradeoff among the various factors can be made within the parameter space bounded by SCE, body-effect, and oxide-field considerations.

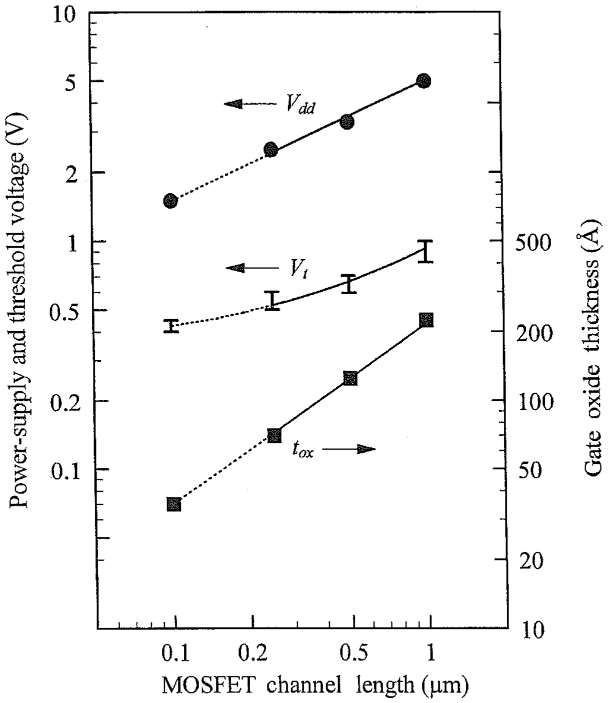


FIGURE 4.9. Trends of power-supply voltage, threshold voltage, and gate oxide thickness versus channel length for CMOS technologies from 1 to 0.1 μm. (After Taur *et al.*, 1995a.)

power-supply voltage, threshold voltage, and oxide thickness for CMOS logic technologies from 1.0- to 0.1-μm channel length (Taur *et al.*, 1995a). When V_{dd} is high, there is plenty of design room to choose a threshold voltage that satisfies both requirements: $0.4\text{ V} \le V_t \le V_{dd}/4$. For example, $V_{dd} = 5$ V and $V_t = 0.8$–1.0 V for 1-μm CMOS technology; and $V_{dd} = 3.3$ V and $V_t = 0.6$–0.7 V for 0.5-μm CMOS technology. When V_{dd} is reduced toward shorter channel lengths, it becomes increasingly difficult to satisfy both the performance and the off-current requirements. One often faces a tradeoff of leakage current versus circuit speed. This stems from subthreshold nonscalability. For this reason and for compatibility with the standardized power-supply voltage of earlier-generation systems, ***the general trend is that V_{dd} has not been scaled down in proportion to L, and V_t has not been scaled down in proportion to V_{dd}***, as is evident in Fig. 4.9.

A higher value of V_{dd}/L leads to a precipitous shrinkage of the design space in Fig. 4.8. If one assumes $W_{dm} + 3t_{ox} \approx L/2.5$ for short-channel effects and applies $3t_{ox}/W_{dm} \approx m - 1$, the oxide thickness can be expressed as

$$t_{ox} \approx \frac{m-1}{m}\frac{L}{8}, \tag{4.38}$$

and the oxide field as

$$\mathscr{E}_{ox} \equiv \frac{V_{dd}}{t_{ox}} \approx 8\frac{m}{m-1}\frac{V_{dd}}{L}. \tag{4.39}$$

Equation (4.38) implies that an oxide thickness of $t_{ox} \approx L/50$ to $L/30$ is desired for control of short-channel effects. From Eq. (4.39), the increase of V_{dd}/L inevitably leads to higher oxide fields. This trend is clearly seen in Table 4.2. Some drain engineering, such as a lightly doped drain (LDD) structure (Ogura *et al.*, 1982) is also necessary to relieve hot-electron reliability problems at higher voltages. However, yield and reliability considerations constrain the maximum oxide field to about 5 MV/cm, since oxide breakdown occurs at slightly beyond 10 MV/cm as mentioned in Section 2.4. When that limit is reached, the power-supply voltage must be reduced for thinner oxides and shorter channel lengths. (Another reason for voltage reduction comes from the active-power consideration to be addressed in Chapter 5.) When V_{dd} becomes less than about 2 V, a tradeoff between off current and device delay is necessary. For example, $V_{dd} = 1.5$ V and $V_t = 0.4$ V for 0.1-μm CMOS devices (Taur *et al.*, 1993c). More details on this tradeoff will be given in Chapter 5.

4.2.3.3 EFFECT OF GATE WORK FUNCTION

The gate work function has a major effect on channel profile design, since, through the V_{fb} term, it has a strong influence on the MOSFET threshold voltage:

$$V_t = V_{fb} + 2\psi_B + \frac{-Q_d}{C_{ox}}. \tag{4.40}$$

When n$^+$-polysilicon gates are used for n-channel MOSFETs, $V_{fb} = -E_g/2q - \psi_B$. This results in a low threshold voltage. For example, consider a 1-μm nMOSFET with $t_{ox} = 250$ Å. A p-type doping of $N_a = 10^{16}$ cm^{-3}, which gives $\psi_B = 0.35$ V and a maximum depletion width [Eq. (4.17)] of $W_{dm}^0 = 0.3$ μm, is sufficient to control the short-channel effect. For a uniformly doped channel, $-Q_d = qN_aW_{dm}^0$. Since $V_{fb} = -0.91$ V for the n$^+$-polysilicon gate, the threshold voltage calculated from Eq. (4.40) is $V_t = 0.14$ V. This is not high enough to satisfy the off-current requirement discussed in Section 4.2.1. ***A nonuniform high–low channel doping as described in the last subsection can be used to increase $|Q_d|$ and therefore the threshold voltage without significantly altering the gate depletion width.*** This is usually carried out with a shallow, p-type (^{11}B or ^{11}B ^{19}F$_2$) implant for nMOSFETs. Since W_{dm}^0 remains essentially unchanged, neither the substrate sensitivity, $dV_t/dV_{bs} = m - 1 = 3t_{ox}/W_{dm}^0$, nor the subthreshold slope, $2.3mkT/q$, is degraded.

Although high–low channel doping allows a higher threshold voltage without degrading the substrate sensitivity, the surface field at threshold, $\mathscr{E}_s = |Q_d|/\varepsilon_{si}$, becomes higher due to the increased depletion charge. This results in degradation of channel mobility, as discussed in Section 3.1.5. Ideally, the threshold voltage can be adjusted by choosing a proper gate work function without increased fields. For example, if a midgap-work-function gate is used in the 1-μm case above, then $V_{fb} = -\psi_B$ and $V_t = 0.7$ V without additional doping in the channel. This will result in the same electric field and channel mobility as the uniform,

10^{16}-cm^{-3}-doped case. A midgap-work-function gate is also symmetrical for nMOSFETs and pMOSFETs. In reality, however, no midgap-work-function gate material has been used in VLSI production, although that has been attempted in research laboratories (Davari *et al.*, 1987). Technology issues such as compatibility of gate material with thin gate oxides are the main obstacles.

4.2.3.4 BURIED-CHANNEL MOSFETs

Further positive shifts in V_t are possible if a p$^+$-polysilicon gate is used for nMOSFETs, which gives $V_{fb} = +E_g/2q - \psi_B$. For the above example, $V_{fb} = +0.21$ V and $V_t = 1.26$ V for a uniform, 10^{16}-cm^{-3}-doped channel. Now the threshold voltage is too high! To reduce V_t, the channel must be counterdoped to lower $|Q_d|$. This means a shallow n-type implant for nMOSFETs, and an n–p junction is formed near the surface. At zero gate voltage, the n-type region is depleted of electrons by the gate field so there is no conduction between the source and drain. The surface field at threshold is lower than in the uniform, 10^{16}-cm^{-3}-doped case, since $|Q_d|$ is lower. This improves channel mobility. The n-type counterdoping can be increased to the point that Q_d becomes positive and the third term on the RHS of Eq. (4.40) becomes negative. When this happens, the surface field at threshold is negative and the MOSFET is called a *buried-channel device*, as inversion first takes place at a point of maximum potential below the surface.

In reality, buried-channel nMOSFETs have not been utilized in VLSI manufacturing, since p$^+$-polysilicon gates have a higher resistance and are more difficult to process because of possible boron penetration problems (Sun *et al.*, 1989). However, their counterparts, n$^+$-polysilicon gated pMOSFETs, have been employed in VLSI manufacturing for CMOS technologies of 0.5-μm channel length and above. In those technologies, n$^+$-polysilicon gates are used for both n- and pMOSFETs, and boron or BF_2 channel implants are made for both types of devices (Taur *et al.*, 1985). Figure 4.10 shows the band diagram of a buried-channel pMOSFET. It can be seen that at the threshold, the surface field is negative (pushing holes away from the surface), and the channel for holes is formed at a potential minimum slightly below the surface. As the gate voltage increases beyond threshold, the field changes sign and the channel moves to the surface, but the effective field is still lower than that of a conventional surface-channel device. ***Although a buried-channel device offers higher mobilities, its short-channel effect is inherently worse than that of a surface-channel device*** (Nguyen and Plummer, 1981). This is because the counterdoping (especially boron) at the surface tends to diffuse deeper into the silicon during subsequent thermal cycles in the process. As the channel length and power-supply voltage are scaled down, a lower threshold voltage is required. It becomes increasingly more difficult to build a buried-channel device, since higher counterdoping in the channel invariably leads to wider gate depletion widths and poorer short-channel effects. ***For CMOS logic technologies of 0.25-μm channel length***

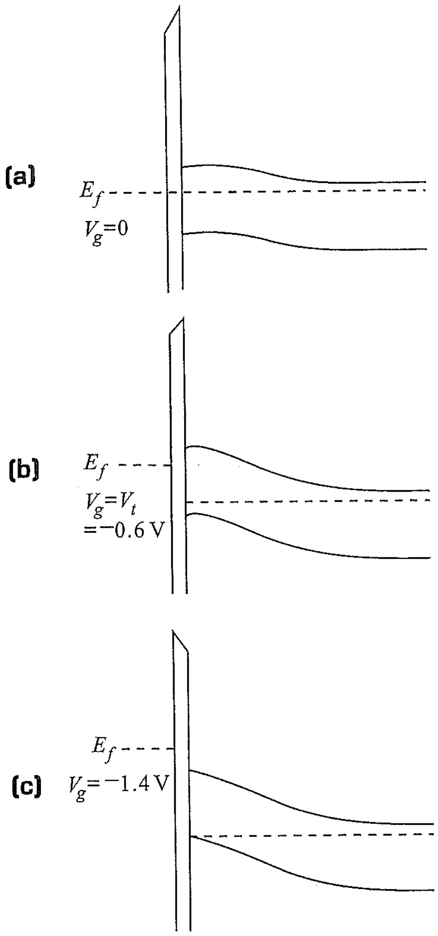


FIGURE 4.10. Band diagram of a buried-channel pMOSFET with n$^+$-polysilicon gate. A shallow p-type layer is implanted at the surface to lower the magnitude of threshold voltage. The gate voltage is (a) below threshold, (b) at threshold, and (c) above threshold. (After Taur *et al.*, 1985.)

and below, dual polysilicon gates (n$^+$ polysilicon for nMOSFET and p$^+$ polysilicon for pMOSFET) are used, so that both types of devices are surface-channel devices (Wong *et al.*, 1988).

4.2.3.5 RETROGRADE (LOW–HIGH) CHANNEL PROFILE

When the channel length is scaled to 0.15 μm and below, a much higher doping concentration is needed in the channel to reduce W_{dm} and control short-channel effects. If a uniform profile were used, the depletion-charge term would increase disproportionately and the threshold voltage would become too high even with dual polysilicon gates. This can be seen by writing the threshold voltage equation (4.18) for the uniformly doped case as

$$V_t = V_{fb} + 2\psi_B + 2(m-1)2\psi_B, \tag{4.41}$$

using Eq. (4.17) and Eq. (4.31). For an n$^+$-polysilicon-gated nMOSFET, $V_{fb} = -E_g/2q - \psi_B$, and therefore

$$V_t = -0.56\text{ V} + (4m-3)\psi_B. \tag{4.42}$$

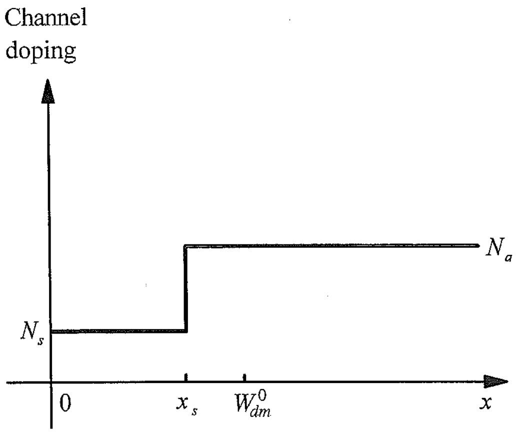


FIGURE 4.11. A schematic diagram showing the low–high (retrograde) step doping profile. $x = 0$ denotes the silicon–oxide interface.

As the channel length scales down, both W_{dm}^0 and t_{ox} are reduced. However, t_{ox} tends not to scale as much as W_{dm}^0 because of the desire to limit the increase of oxide fields and to delay entry into the direct tunneling regime below 30 Å. As a result, m becomes higher. ψ_B has increased as well because of the higher N_a. Both of these factors tend to raise V_t, which is opposite to the V_{dd}-reduction trend required for shorter devices. For example, for 0.1-μm-channel CMOS devices, a channel doping of $N_a = 10^{18}$ cm^{-3}, which gives $\psi_B = 0.47$ V and $W_{dm}^0 = 350$ Å, is needed to control the short-channel effect. If $t_{ox} = 35$ Å, then $m = 1.3$ and $V_t = 0.47$ V, which is too high with respect to the 1.5-V supply voltage for 0.1-μm devices (Fig. 4.9). The problem is further aggravated by quantum effects, which, as will be discussed in Section 4.2.4, can raise the threshold voltage by another 0.1–0.2 V at such high fields (van Dort *et al.*, 1994).

To reduce the threshold voltage without significantly increasing the gate depletion width, a retrograde channel profile, i.e., a low–high doping profile as shown schematically in Fig. 4.11, is required (Sun *et al.*, 1987; Shahidi *et al.*, 1989). Such a profile is formed using higher-energy implants that peak below the surface. It is assumed that the maximum gate depletion width extends into the higher-doped region. All the equations in Section 4.2.2 remain valid for $N_s < N_a$. For simplicity, we assume an ideal retrograde channel profile for which $N_s = 0$. Equation (4.28) then becomes

$$V_t = V_{fb} + 2\psi_B + \frac{qN_a}{C_{ox}}\sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2} - \frac{qN_a x_s}{C_{ox}}. \tag{4.43}$$

Similarly, Eq. (4.29) gives the maximum depletion width (long-channel),

$$W_{dm}^0 = \sqrt{\frac{4\varepsilon_{si}\psi_B}{qN_a} + x_s^2}. \tag{4.44}$$

The net effect of low–high doping is that the threshold voltage is reduced, but

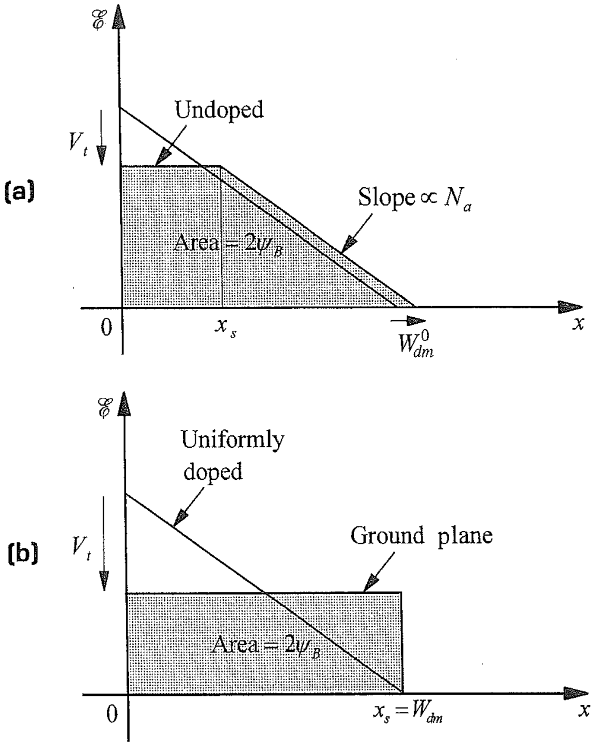


FIGURE 4.12. Graphical interpretation of retrograde doping profiles. (a) A low–high step profile compared with a uniformly doped profile (N_a). (b) An extreme retrograde profile that degenerates into a ground-plane MOSFET; the band bending in this case is given by the rectangular area, which equals $2\psi_B$ at threshold.

the depletion width has increased, just opposite to that of high-low doping. Note that Eq. (4.44) has the same form as Eq. (2.76) for a p–i–n diode discussed in Section 2.2.2. All other expressions, such as those for the subthreshold slope and the substrate sensitivity, in Section 4.2.2 apply with W_{dm}^0 replaced by Eq. (4.44).

A graphical representation of the retrograde channel profile is shown in Fig. 4.12(a). As described in Section 4.2.2.3, when the electric field is plotted against depth at threshold condition, the x-intercept is the maximum depletion width while the y-intercept is proportional to the depletion charge (third) term of V_t in Eq. (4.40). The area under the $\mathscr{E}(x)$ curve equals $2\psi_B$. With a retrograde doping profile, it is possible to reduce the y-intercept, and hence V_t, with only a slight increase in the depletion width while keeping the area under the curve unchanged. Note that $\mathscr{E}(x)$ is flat within the undoped region, $0 < x < x_s$, where there is no depletion charge.

4.2.3.6 EXTREME RETROGRADE PROFILE AND GROUND-PLANE MOSFET

Two limiting cases are worth discussing. If $x_s \ll (4\varepsilon_{si}\psi_B/qN_a)^{1/2}$, then W_{dm}^0 remains essentially unchanged from the uniformly doped value [Eq. (4.44)], while V_t is lowered by a net amount equal to $qN_a x_s/C_{ox}$ [Eq. (4.43)]. To reduce V_t even further, x_s must increase, assuming there is no counterdoping of the channel. If N_a

stays the same, it can be seen from Fig. 4.12(a) that W_{dm}^0 will widen significantly, which degrades the short-channel effect. To keep W_{dm}^0 unchanged, the concentration (i.e., slope) beyond x_s must be raised from N_a to N_a' while x_s is increased. In the limiting case shown in Fig. 4.12(b), $x_s = W_{dm}^0$, and the entire depletion region is undoped. All the depletion charge is concentrated at the edge of the depletion region. In order for this to occur, N_a' must be high enough that $x_s \gg (4\varepsilon_{si}\psi_B/qN_a')^{1/2}$. With N_a replaced by N_a', Eq. (4.43) can be expanded under this limit to yield

$$V_t = V_{fb} + 2\psi_B + \frac{\varepsilon_{si}/x_s}{C_{ox}} 2\psi_B. \tag{4.45}$$

This result is expected from Fig. 4.12(b), since the y-intercept equals the area divided by the x-intercept, or $\mathscr{E}_s = 2\psi_B/x_s$. It is interesting to note that in this case, the maximum depletion width becomes independent of channel length. In other words, there is no need to distinguish between W_{dm} and W_{dm}^0. Using $m = 1 + 3t_{ox}/W_{dm} = 1 + 3t_{ox}/x_s$, one can write Eq. (4.45) as

$$V_t = V_{fb} + 2\psi_B + (m - 1)2\psi_B. \tag{4.46}$$

Comparison with Eq. (4.41) shows that, with the extreme retrograde profile, the depletion-charge term of V_t is reduced to half of the uniformly doped value. This is also clear from Fig. 4.12(b). Substituting $V_{fb} = -E_g/2q - \psi_B$ in Eq. (4.46) yields

$$V_t = -0.56\ \text{V} + (2m - 1)\psi_B. \tag{4.47}$$

For the above 0.1-μm MOSFET example, $\psi_B = 0.47$ V and $m = 1.3$, which gives $V_t = 0.19$ V. This value is low enough for $V_{dd} = 1.5$ V, even with the quantum correction of V_t (Section 4.2.4) taken into account. If there is a substrate bias $-V_{bs}$ present, the factor $2\psi_B$ in the last term of Eq. (4.46) is replaced by $2\psi_B + V_{bs}$, i.e.,

$$V_t = V_{fb} + 2m\psi_B + (m - 1)V_{bs}. \tag{4.48}$$

Further reduction of V_t can be accomplished by either counterdoping the channel or forward-biasing the substrate. A forward substrate bias also helps improve short-channel effects, as it effectively reduces the built-in potential, ψ_{bi} in Eq. (3.66), between the source–drain and the p-type substrate. However, forward substrate bias also causes source junction leakage, increases the drain-to-substrate capacitance, and degrades the subthreshold slope and body effect.

Since ψ_B is a weak function of N_a', the above results are independent of the exact value of N_a' as long as it is high enough to satisfy $x_s \gg (4\varepsilon_{si}\psi_B/qN_a')^{1/2}$. All the essential device characteristics, such as SCE (W_{dm}), subthreshold slope (m), and threshold voltage, are determined by the depth of the undoped layer, x_s. ***The limiting case of retrograde channel profile therefore degenerates into a ground-plane MOSFET*** (Yan *et al.*, 1991). The band diagram and charge distribution of

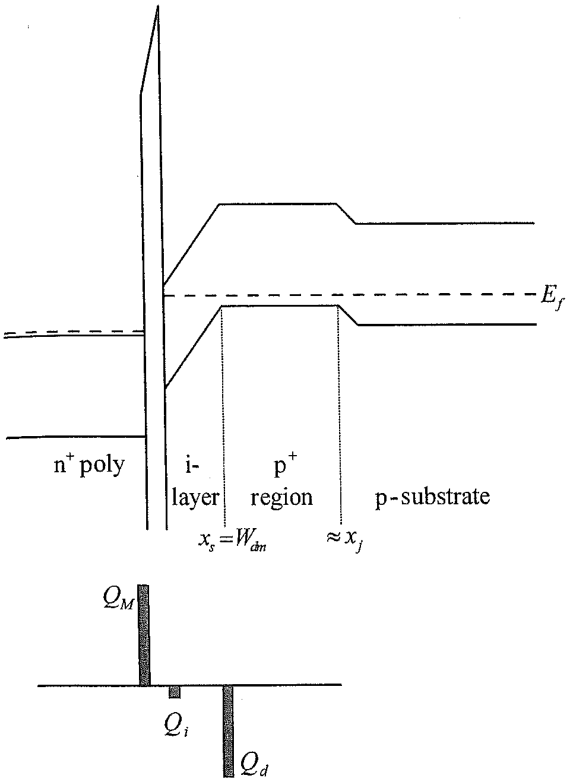


FIGURE 4.13. Band diagram and charge distribution of an extreme retrograde-doped or ground-plane nMOSFET at threshold condition.

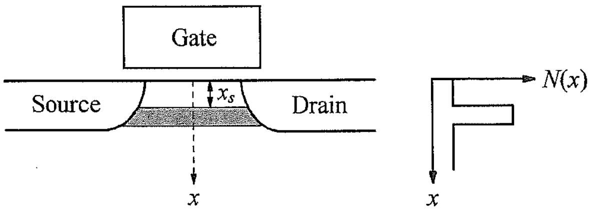


FIGURE 4.14. Schematic cross section of a low–high–low, or pulse-shaped, or delta-doped MOSFET. The doping concentration along the dashed line is depicted in the profile to the right. The highly doped region corresponds to the shaded area in the cross section.

such a device at threshold condition are shown schematically in Fig. 4.13. Note that the field is constant (no potential curvature) in the undoped region between the surface and x_s. There is an abrupt change of field at $x = x_s$, where a delta function of depletion charge (area $= 2\varepsilon_{si}\psi_B/x_s$) is located. Beyond x_s, the bands are essentially flat. It is desirable not to extend the p$^+$ region under the source and drain junctions, since that increases the parasitic capacitance. The ideal channel doping profile is then that of a low–high–low type shown in Fig. 4.14, in which the narrow p$^+$region is used only to confine the gate depletion width. Such a profile is also referred to as *pulse-shaped doping* or *delta doping* in the literature. The integrated dose of the p$^+$ region must be at least $2\varepsilon_{si}\psi_B/qx_s$ to provide the gate depletion charge needed. It is advisable to use somewhat higher than the minimum dose to supply additional depletion charge for the source–drain fields in short-channel devices. However, too high a p$^+$ dose or concentration may result in band-to-band tunneling leakage between the source or drain and the substrate, as mentioned in Section 2.4.2.

4.2.3.7 LATERALLY NONUNIFORM CHANNEL DOPING

So far we have discussed nonuniform channel doping in the vertical direction. Another type of nonuniform doping used in very short-channel devices is in the lateral direction. ***For nMOSFETs, more highly p-type-doped regions near the two ends of the channel are beneficial to the suppression of short-channel effect, since they help compensate charge-sharing effects from the source–drain fields*** described in Section 3.2.1 (Ogura *et al.*, 1982). This can be implemented by a moderate-dose p-type implant carried out together with the n$^+$ source–drain implant after gate patterning. Such a self-aligned, laterally nonuniform channel doping is often referred to as *halo* or *pocket* implants (Taur *et al.*, 1993c). With an optimally designed 2-D nonuniform doping profile called the *superhalo*, it is possible to counteract the short-channel effect and achieve nearly identical I_{on} and I_{off} for devices of different channel lengths within the process tolerances (Taur and Nowak, 1997).

4.2.4 QUANTUM EFFECT ON THRESHOLD VOLTAGE

It was discussed in Section 2.3.2 that in the inversion layer of a MOSFET, carriers are confined in a potential well very close to the silicon surface. The well is formed by the oxide barrier (essentially infinite except for tunneling calculations) and the silicon conduction band, which bends down severely toward the surface due to the applied gate field. Because of the confinement of motion in the direction normal to the surface, inversion-layer electrons must be treated quantum-mechanically as a 2-D gas (Stern and Howard, 1967), especially at high normal fields. Thus the energy levels of the electrons are grouped in discrete *subbands*, each of which corresponds to a quantized level for motion in the normal direction, with a continuum for motion in the plane parallel to the surface. An example of the quantum-mechanical energy levels and band bending is shown in Fig. 4.15. The electron concentration peaks below the silicon–oxide interface and goes to nearly zero at the interface, as dictated by the boundary condition of the electron wave function. This is in contrast to the classical model in which the electron concentration peaks at the surface, as shown in Fig. 4.16. Quantum-mechanical behavior of inversion-layer electrons affects MOSFET operation in two ways. ***First, at high fields, threshold voltage becomes higher, since more band bending is required to populate the lowest subband, which is some energy above the bottom of the conduction band. Second, once the inversion layer forms below the surface, it takes a higher gate-voltage overdrive to produce a given level of inversion charge density***. In other words, the effective gate oxide thickness is slightly larger than the physical thickness. This reduces the transconductance and the current drive of a MOSFET.

4.2.4.1 TRIANGULAR POTENTIAL APPROXIMATION FOR THE SUBTHRESHOLD REGION

A full solution of the silicon inversion layer involves numerically solving coupled Poisson's and Schrödinger's equations self-consistently (Stern and Howard, 1967).

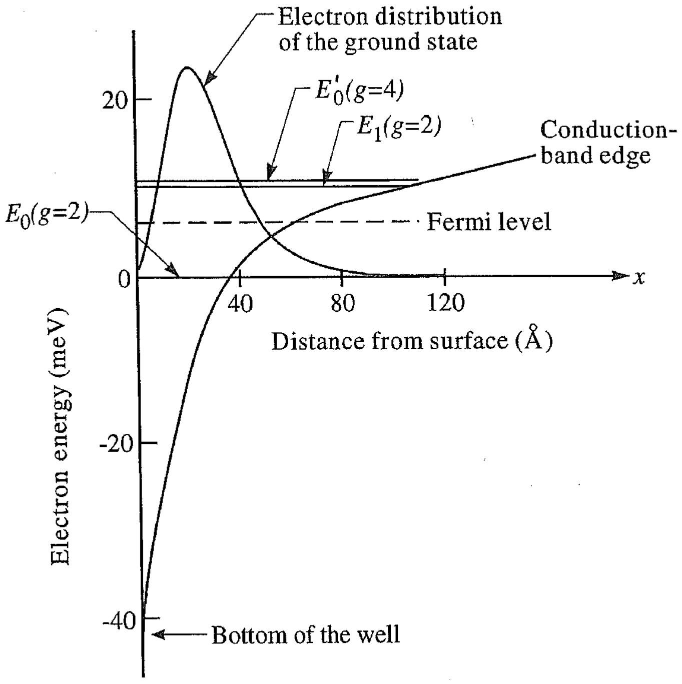


FIGURE 4.15. An example of quantum-mechanically calculated band bending and energy levels of inversion-layer electrons near the surface of an MOS device. The ground state is about 40 meV above the bottom of the conduction band at the surface. The dashed line indicates the Fermi level for 10^{12} electrons/cm^2 in the inversion layer. (After Stern and Howard, 1967.)

Under subthreshold conditions when the inversion charge density is low, band bending is solely determined by the depletion charge. It is then possible to decouple the two equations and obtain some insight into the quantum-mechanical (QM) effect on the threshold voltage. Since the inversion electrons are located in a narrow region close to the surface where the electric field is nearly constant ($\mathscr{E}_s$), it is a good approximation to consider the potential well as composed of an infinite oxide barrier for $x < 0$, and a triangular potential $V(x) = q\mathscr{E}_s x$ due to the depletion charge for $x > 0$. The Schrödinger equation is solved with the boundary conditions that the electron wave function goes to zero at $x = 0$ and at infinity. The solutions are Airy functions with eigenvalues E_j given by (Stern, 1972)

$$E_j = \left[\frac{3hq\mathscr{E}_s}{4\sqrt{2m_x}}\left(j + \frac{3}{4}\right)\right]^{2/3}, \qquad j = 0, 1, 2, \ldots, \tag{4.49}$$

where $h = 6.63 \times 10^{-34}$ J-s is Planck's constant, and m_x is the effective mass of electrons perpendicular to the surface. Note that MKS units are used throughout this subsection (i.e., length must be in meters, not centimeters). The average distance from the surface for electrons in the jth subband is given by

$$x_j = \frac{2E_j}{3q\mathscr{E}_s}. \tag{4.50}$$

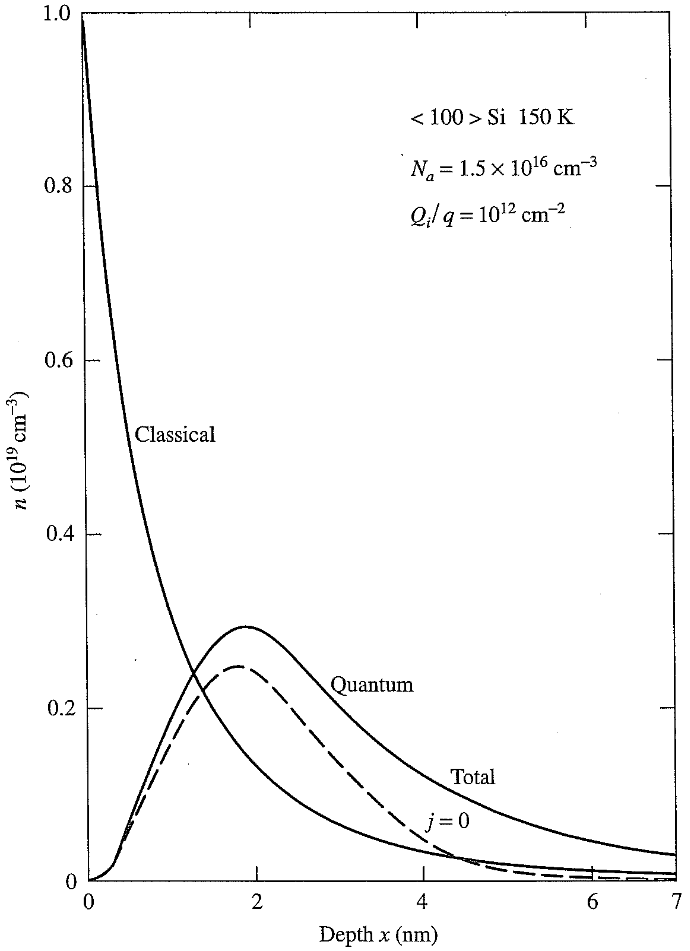


FIGURE 4.16. Classical and quantum-mechanical electron density versus depth for a ⟨100⟩ silicon inversion layer. The dashed curve shows the electron density distribution for the lowest subband. (After Stern, 1974.)

For silicon in the ⟨100⟩ direction, there are two groups of subbands, or *valleys*. The lower valley has a twofold degeneracy ($g = 2$) with $m_x = 0.92m_0$, where $m_0 = 9.1 \times 10^{-31}$ kg is the free-electron mass. These energy levels are designated as $E_0, E_1, E_2 \ldots$. The higher valley has a fourfold degeneracy ($g' = 4$) with $m'_x = 0.19m_0$. The energy levels are designated as $E'_0, E'_1, E'_2, \ldots$. Note that

$$E'_j = \left[\frac{3hq\mathscr{E}_s}{4\sqrt{2m'_x}}\left(j + \frac{3}{4}\right)\right]^{2/3}, \qquad j = 0, 1, 2, \ldots \tag{4.51}$$

At room temperature, several subbands in both valleys are occupied near threshold, with a majority of the electrons in the lowest subband of energy E_0 above the bottom of the conduction band. From Appendix 7, the total inversion charge per unit area is expressed as (Stern and Howard, 1967)

$$Q_i^{QM} = \frac{4\pi qkT}{h^2}\left(gm_d\sum_j \ln\left(1+e^{(E_f-E_c'-E_j)/kT}\right)\right.$$
$$\left. + g'm_d'\sum_j \ln\left(1+e^{(E_f-E_c'-E_j')/kT}\right)\right), \tag{4.52}$$

where $m_d = 0.19m_0$ and $m_d' = 0.42m_0$ are the density-of-states effective masses of the two valleys, and $E_f - E_c'$ is the difference between the Fermi level and the bottom of the conduction band at the surface. It is shown in Appendix 7 that in the subthreshold region, Eq. (4.52) can be simplified to

$$Q_i^{QM} = \frac{4\pi qkTn_i^2}{h^2N_cN_a}\left(2m_d\sum_j e^{-E_j/kT}\right.$$
$$\left. + 4m_d'\sum_j e^{-E_j'/kT}\right)e^{q\psi_s/kT}, \tag{4.53}$$

where N_c is the effective density of states in the conduction band.

4.2.4.2 THRESHOLD-VOLTAGE SHIFT DUE TO QUANTUM EFFECT

When $\mathscr{E}_s < 10^4$–10^5 V/cm at room temperature, both the lowest energy level E_0 and the spacings between the subbands are comparable to or less than kT. A large number of subbands are occupied, and Q_i^{QM} is essentially the same as the classical inversion charge density per unit area given by Eq. (3.31) for the subthreshold region,

$$Q_i = \frac{kTn_i^2}{\mathscr{E}_sN_a}e^{q\psi_s/kT}. \tag{4.54}$$

(The expression has been generalized to cover nonuniformly doped cases where $\mathscr{E}_s$ is the electric field at the surface and N_a is the doping concentration at the edge of the depletion layer.) When $\mathscr{E}_s > 10^5$ V/cm, however, the subband spacings become greater than kT and Q_i^{QM} is significantly less than Q_i. ***The Q_i^{QM}–ψ_s curve [Eq. (4.53)] exhibits a positive parallel shift with respect to the classical Q_i–ψ_s curve [Eq. (4.54)] on a semilogarithmic scale, which means that additional band bending is required to achieve the same inversion charge per unit area as the classical value***. The classical threshold condition, $\psi_s = 2\psi_B$, should therefore be modified to $\psi_s = 2\psi_B + \Delta\psi_s^{QM}$, where $Q_i^{QM}(\psi_s = 2\psi_B + \Delta\psi_s^{QM}) = Q_i(\psi_s = 2\psi_B)$. Based on the last expression,

$$\Delta\psi_s^{QM} = \frac{kT}{q}\ln\left(\frac{Q_i(\psi_s = 0)}{Q_i^{QM}(\psi_s = 0)}\right) \tag{4.55}$$

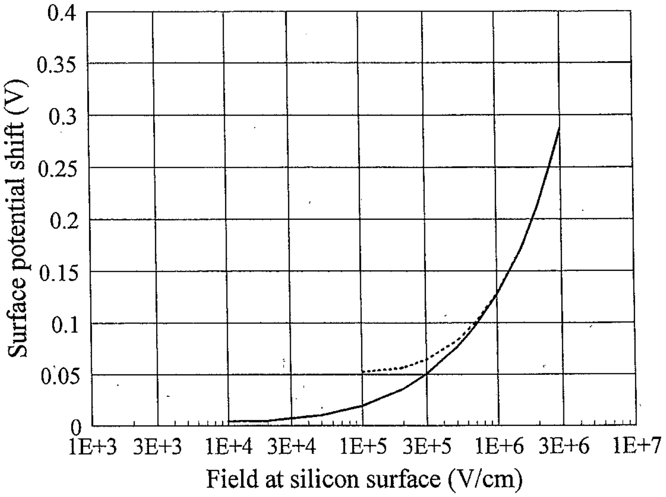


FIGURE 4.17. Additional band bending $\Delta\psi_s^{QM}$ (over the classical $2\psi_B$ value) required for reaching the threshold condition as a function of surface electric field. The dotted curve is calculated by keeping only the lowest term (twofold degeneracy) in Eq. (4.53.)

can be evaluated from the preexponential factors in Eqs. (4.54) and (4.53). Figure 4.17 shows the calculated $\Delta\psi_s^{QM}$ as a function of $\mathscr{E}_s$. Beyond 10^6 V/cm, only the lowest subband is occupied by electrons, and

$$\Delta\psi_s^{QM} \approx \frac{E_0}{q} - \frac{kT}{q}\ln\left(\frac{8\pi q m_d \mathscr{E}_s}{h^2 N_c}\right), \tag{4.56}$$

as indicated by the dotted curve in Fig. 4.17. Knowing $\Delta\psi_s^{QM}$, one can easily calculate the threshold voltage shift due to the quantum effect:

$$\Delta V_t^{QM} = \frac{dV_g}{d\psi_s}\,\Delta\psi_s^{QM} = m\;\Delta\psi_s^{QM}, \tag{4.57}$$

where $m = 1 + (3t_{ox}/W_{dm}^0)$ as before. For the 0.1-μm MOSFET example discussed in Section 4.2.3, $N_a = 10^{18}$ cm^{-3} and $m = 1.3$. If the channel is uniformly doped, $\mathscr{E}_s = 5.4 \times 10^5$ V/cm, $\Delta\psi_s^{QM} = 0.08$ V, and $\Delta V_t^{QM} = 0.10$ V, which makes the threshold voltage (0.57 V) unacceptably high. With an extreme retrograde doping profile, the surface electric field is reduced by a factor of two to 2.7×10^5 V/cm, for which $\Delta\psi_s^{QM} = 0.046$ V. This brings $\Delta V_t^{QM} = 0.06$ V and a minimum V_t of 0.25 V. It is not very difficult to adjust the retrograde profile to obtain a V_t higher than the extreme value, e.g., 0.4 V, suitable for the 1.5-V power-supply voltage for such devices (Fig. 4.9).

4.2.4.3 QUANTUM EFFECT ON INVERSION-LAYER DEPTH

After strong inversion, the inversion charge density builds up rapidly and the triangular potential-well model is no longer valid. If the field is high enough that only the lowest subband is populated, a variational approach leads to an approximate

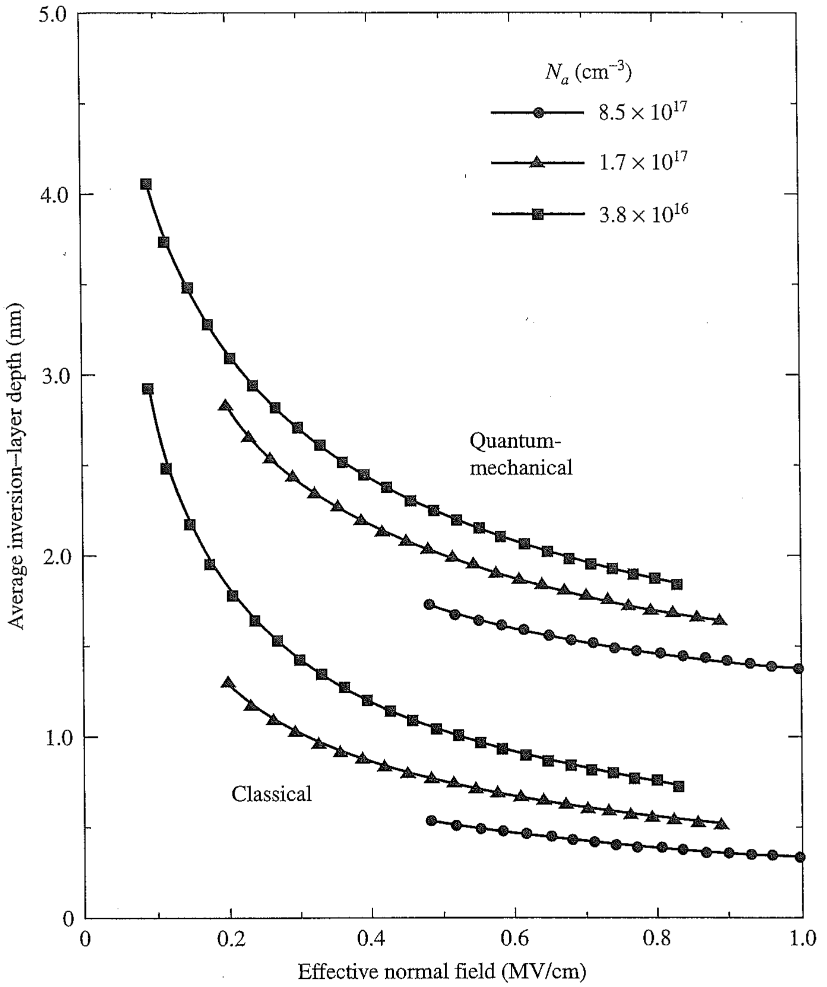


FIGURE 4.18. Calculated QM and classical inversion-layer depth versus effective normal field for several uniform doping concentrations. (After Ohkura, 1990.)

expression for the average distance of electrons from the surface (Stern, 1972):

$$x_{av}^{QM} = \left(\frac{9\varepsilon_{si}h^2}{16\pi^2 m_x q Q^*} \right)^{1/3}, \tag{4.58}$$

where $Q^* = Q_d + \frac{11}{32}Q_i$ is a combination of the depletion and inversion charge per unit area in the channel. For intermediate fields, the solution must be obtained numerically. Figure 4.18 shows a comparison of the classical and QM inversion-layer depths versus the effective normal field defined in Eq. (3.47) (Ohkura, 1990). The QM value is consistently larger than the classical value by about 10–12 Å for a wide range of channel doping (uniform) and effective fields. Since

$$V_g = V_{fb} + \psi_s + \frac{Q_d}{C_{ox}} + \frac{Q_i}{C_{ox}}, \tag{4.59}$$

where $C_{ox} = \varepsilon_{ox}/t_{ox}$ and the band bending at the surface is [Eq. (4.21)]

$$\psi_s = \frac{Q_i}{\varepsilon_{si}} x_{av} + \frac{q}{\varepsilon_{si}} \int_0^{W_d} N_a(x) x \, dx, \tag{4.60}$$

the quantum-mechanical effect adds $\Delta t_{ox} = (\varepsilon_{ox}/\varepsilon_{si})\, \Delta x_{av} = (x_{av}^{QM} - x_{av}^{CL})/3$ ***or about 3–4 Å to the gate oxide thickness for calculation of the inversion charge density***. This effectively reduces the current drive and the transconductance of thin-oxide MOSFETs.

4.2.5 DISCRETE DOPANT EFFECTS ON THRESHOLD VOLTAGE

As CMOS devices are scaled down, the number of dopant atoms in the depletion region of a minimum geometry device decreases. Due to the discreteness of atoms, there is a statistical random fluctuation of the number of dopants within a given volume around its average value. For example, in a uniformly doped $W = L = 0.1$-μm nMOSFET, if $N_a = 10^{18}$ cm^{-3} and $W_{dm}^0 = 350$ Å, the average number of acceptor atoms in the depletion region is $N = N_a L W W_{dm}^0 = 350$. The actual number fluctuates from device to device with a standard deviation $\sigma_N = \langle(\Delta N)^2\rangle^{1/2} = N^{1/2} = 18.7$, which is a significant fraction of the average number N. Since the threshold voltage of a MOSFET depends on the charge of ionized dopants in the depletion region, this translates into a threshold-voltage fluctuation which could affect the operation of VLSI circuits.

4.2.5.1 A SIMPLE FIRST-ORDER MODEL

To estimate the effect of depletion charge fluctuation on threshold voltage, we consider a small volume $dx\,dy\,dz$ at a point (x, y, z) in the depletion region of a uniformly doped (N_a) MOSFET. The x-axis is in the depth direction, the y-axis in the length direction, and the z-axis in the width direction. The average number of dopant atoms in this small volume is $N_a\,dx\,dy\,dz$. The actual number fluctuates around this value with a standard deviation of $\sigma_{dN} = (N_a\,dx\,dy\,dz)^{1/2}$. This fluctuation can be thought of as a small delta function of nonuniform doping (either positive or negative) at (x, y, z) superimposed on a uniformly doped background N_a. Here we focus on the linearly extrapolated threshold voltage V_{on}, as defined in Fig. 3.16. When there is a slight local nonuniformity of doping in either the channel-width or the channel-length direction, the first-order influence on the linear threshold voltage is through its effect on the depletion charge density averaged over the entire channel area (Nguyen, 1984). This is similar to the assumption made in the charge-sharing model for short-channel effects in Section 3.2.1. The effect of the above doping fluctuation on the linear threshold voltage is then equivalent to that of a uniform delta-function implant of dose (number of ions per unit area) ΔD and depth x, where $\langle(\Delta D)^2\rangle^{1/2} = \sigma_{dN}/WL = (N_a\,dx\,dy\,dz)^{1/2}/WL$. The threshold-voltage shift is obtained by substituting $D_I = \Delta D$ and $x_c = x$ in Eq. (4.36) and retaining only the first-order terms in ΔD:

$$\Delta V_{on} = \frac{q\,\Delta D}{C_{ox}}\left(1 - x\sqrt{\frac{qN_a}{2\varepsilon_{si}(2\psi_B)}}\right) = \frac{q\,\Delta D}{C_{ox}}\left(1 - \frac{x}{W_{dm}^0}\right). \tag{4.61}$$

The last expression is quite general and is applicable to a nonuniformly doped background as well. It follows directly from Eq. (4.21) or can be seen from the graphical representation in Fig. 4.5(b). The mean square deviation (variance) of threshold voltage due to the depletion charge fluctuation in $dx\,dy\,dz$ is then

$$\left\langle \Delta V_{on}^2 \right\rangle_{x,y,z} = \frac{q^2 N_a}{C_{ox}^2 L^2 W^2}\left(1 - \frac{x}{W_{dm}^0}\right)^2 dx\,dy\,dz. \tag{4.62}$$

Since dopant number fluctuations at various points are completely random and uncorrelated, the total mean square fluctuation of the threshold voltage is obtained by integrating Eq. (4.62) over the entire depletion region:

$$\sigma_{V_{on}}^2 = \frac{q^2 N_a}{C_{ox}^2 L^2 W^2}\int_0^W \int_0^L \int_0^{W_{dm}^0} \left(1 - \frac{x}{W_{dm}^0}\right)^2 dx\,dy\,dz. \tag{4.63}$$

It is straightforward to carry out the integration and obtain

$$\sigma_{V_{on}} = \frac{q}{C_{ox}}\sqrt{\frac{N_a W_{dm}^0}{3LW}}. \tag{4.64}$$

In the above 0.1-µm example, $\sigma_{V_{on}} = 17.5$ mV if $t_{ox} = 35$ Å. This is small compared with the worst-case short-channel threshold roll-off in Section 4.2.1, but can be significant in minimum-geometry devices, for example, in an SRAM cell.

In the above analysis, it was assumed that the surface potential is uniform in both the length and the width directions of the device. In other words, all the lumpiness due to local fluctuations of the depletion charge is smoothed out and the surface potential depends only on the average (or total) depletion charge of the device. This assumption is not valid in the subthreshold region, where current injection is dominated by the highest potential barrier in the channel rather than by the average value (Nguyen, 1984). In general, the problem needs to be solved by 3-D numerical simulations (Wong and Taur, 1993). The results indicate that in addition to the threshold fluctuations of a similar magnitude to that expected from Eq. (4.64), there is also a negative shift of the average threshold voltage, especially in the subthreshold region. This is believed to be due to the inhomogeneity of surface potential resulting from the microscopic random distribution of discrete dopant atoms in the channel. For the same reason, the source–drain current may exhibit some statistical asymmetry under high-drain-bias conditions (Wong and Taur, 1993).

4.2.5.2 DISCRETE DOPANT EFFECTS IN A RETROGRADE-DOPED CHANNEL

Threshold voltage fluctuations due to discrete dopants are greatly reduced in a retrograde-doped channel. Consider the profile in Fig. 4.11 with $N_s = 0$, i.e., the channel is undoped within $0 < x < x_s$. The average threshold voltage and the maximum depletion width W_{dm}^0 are given by Eq. (4.43) and Eq. (4.44), respectively. For a small volume of dopants at (x, y, z) where $x_s < x < W_{dm}^0$, Eq. (4.62) still holds. The x-integral in Eq. (4.63), however, is carried out from x_s to W_{dm}^0, which results in

$$\sigma_{V_{on}} = \frac{q}{C_{ox}}\sqrt{\frac{N_a W_{dm}^0}{3LW}}\left(1 - \frac{x_s}{W_{dm}^0}\right)^{3/2} \tag{4.65}$$

for a retrograde-doped channel. ***In the extreme retrograde or ground-plane limit shown in Fig. 4.12(b), $x_s = W_{dm}^0$, and the threshold voltage fluctuation goes to zero.*** This is also clear from Eq. (4.45), where the threshold voltage is essentially independent of N_a (or N_a'). Of course, the technological challenge is then to control the tolerance of the undoped-layer thickness x_s so that it does not introduce a different kind of threshold voltage variations.

4.3 MOSFET CHANNEL LENGTH

Channel length is a key parameter in CMOS technology used for performance projection (circuit models), short-channel design, and model–hardware correlation. This section focuses on MOSFET channel length: its definition, extraction, and physical interpretation.

4.3.1 VARIOUS DEFINITIONS OF CHANNEL LENGTH

A number of quantities, e.g., *mask length* (L_{mask}), *gate length* (L_{gate}), *metallurgical channel length* (L_{met}), and *effective channel length* (L_{eff}), have been used to describe the length of a MOSFET. Even though they are all related to each other, their relationships are strongly process-dependent.

Figure 4.19 shows schematically how various channel lengths are defined. L_{mask} is the design length on the polysilicon etch mask. It is reproduced on the wafer as L_{gate} through lithography and etching processes. Depending on the lithography and etching biases, L_{gate} can be either longer or shorter than L_{mask}. There are also process tolerances associated with L_{gate}. For the same L_{mask} design, L_{gate} may vary

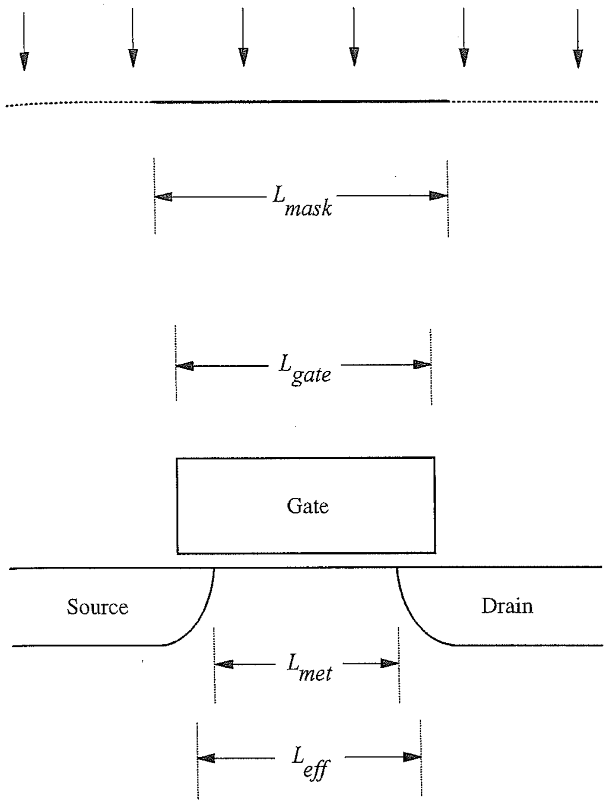


FIGURE 4.19. Schematic diagram showing the definitions of and relationship among the various notions of channel length. The physical interpretation of L_{eff} is examined in Section 4.3.3.

from chip to chip, wafer to wafer, and run to run. Although L_{gate} is an important parameter for process control and monitoring, there is no simple way of making a large number of measurements of it. Usually, L_{gate} is measured with a scanning electron microscope (SEM) and only sporadically across the wafer. There is also an uncertainty in the precise definition of L_{gate} when the polysilicon etch profile is not vertical, as to whether L_{gate} refers to the top or to the bottom dimension of the gate.

L_{met} is defined as the distance between the metallurgical junctions of the source and drain diffusions at the silicon surface. In a modern CMOS process, the source and drain regions are self-aligned to the polysilicon gate by performing the source–drain implant after gate patterning (Kerwin *et al.*, 1969). As a result, there is a close correlation between L_{met} and L_{gate}. Usually, L_{met} is shorter than L_{gate} by a certain amount due to the lateral implant straggle and the lateral source–drain diffusion in the process. Accurate physical measurement of L_{met} in actual hardware is very difficult. Normally, L_{met} is used only in 2-D models for short-channel device design. Even for that purpose, difficulties arise in defining L_{met} when dealing with a buried-channel device or a retrograde channel profile with zero surface doping, where there are no metallurgical junctions at the silicon surface.

The parameter L_{eff} is different from all other channel lengths discussed above in that it is defined through some electrical characteristics of the MOSFET device and is not a physical parameter. ***Basically, L_{eff} is a measure of how much gate-controlled current a MOSFET delivers and is therefore most suitable for circuit models.*** L_{eff} also allows for a large number of automated measurements, since it can be extracted from electrically measured terminal currents. The basis of the L_{eff} definition lies in the fact that the channel resistance of a MOSFET in the linear or low-drain bias region is proportional to the channel length, as indicated by Eq. (3.100) (Dennard *et al.*, 1974). Further details of the definition and the extraction of L_{eff} are given in the next subsection.

For submicron CMOS technologies, it is important to distinguish among the various notions of channel length. The errors can be significant, since lithography and etching bias, junction depletion width, and lateral source–drain diffusions are all becoming an appreciable fraction of the channel length.

4.3.2 EXTRACTION OF THE EFFECTIVE CHANNEL LENGTH

As discussed in the last subsection, the effective channel length L_{eff} is defined by its proportionality to the linear or low-drain channel resistance. That is,

$$R_{ch} \equiv \frac{V_{ds}}{I_{ds}} = \frac{L_{eff}}{\mu'_{eff} C_{ox} W (V_g - V_{on} - m V_{ds}/2)} \tag{4.66}$$

from Eq. (3.61), where V_{on} is the linearly extrapolated threshold voltage and μ'_{eff} is the modified effective mobility, which contains the inversion-layer capacitance effect. μ'_{eff} is a weak function of V_g. For different L_{mask}, L_{eff} differs but is assumed to be related to L_{mask} by a constant *channel length bias* ΔL:

$$L_{eff} = L_{mask} - \Delta L. \tag{4.67}$$

All the lithography and etch biases as well as the lateral source–drain implant straggle and diffusion are lumped into ΔL. The assumption that the channel length bias is constant is a reasonable one when the channel length is not too short. However, ΔL can be linewidth-dependent when L_{mask} approaches the resolution limit of the lithography tool used in the process. This issue will be addressed later.

In the simplest scheme of channel-length extraction (Dennard *et al.*, 1974), R_{ch} is measured for a set of devices with different L_{mask}. Based on Eq. (4.66) and Eq. (4.67), a plot of R_{ch} for a given V_g versus L_{mask} should yield a straight line whose intercept with the x-axis gives ΔL and therefore L_{eff}. In practice, however, two issues must be addressed for short-channel devices. The first one is the source–drain series resistance. The second one is the short-channel effect (SCE), which causes V_{on} in Eq. (4.66) to depend on L_{mask}.

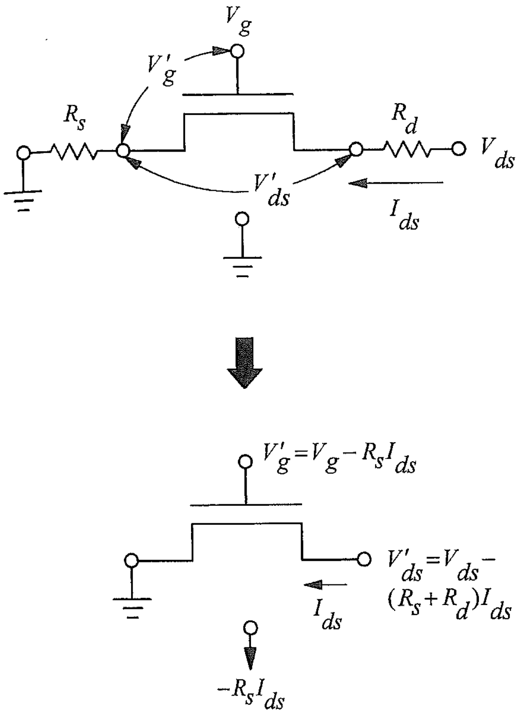


FIGURE 4.20. Equivalent circuit of MOSFET with source and drain series resistance. The intrinsic part of the top circuit is equivalent to the bottom circuit with redefined terminal voltages.

4.3.2.1 CHANNEL-RESISTANCE METHOD

The effect of source–drain resistance is examined using the equivalent circuit in Fig. 4.20. A source resistance R_s and a drain resistance R_d are assumed to connect an *intrinsic MOSFET* to the external terminals where voltages V_{ds} and V_g are applied. The internal voltages are V'_{ds} and V'_g for the intrinsic MOSFET. One can write the following relations:

$$V'_{ds} = V_{ds} - (R_s + R_d) I_{ds} \tag{4.68}$$

and

$$V'_g = V_g - R_s I_{ds}. \tag{4.69}$$

As shown in Fig. 4.20, the intrinsic part of an actual device with parasitic resistance is equivalent to an intrinsic MOSFET with a grounded source, with V'_g and V'_{ds} at the gate and the drain terminals, and with a reverse bias $-R_s I_{ds}$ on the substrate. Based on Eq. (4.66), but with redefined voltage symbols on the intrinsic nodes, the channel resistance of the intrinsic device is given by

$$R_{ch} \equiv \frac{V'_{ds}}{I_{ds}} = \frac{L_{eff}}{\mu'_{eff} C_{ox} W \left(V'_g - V'_{on} - m V'_{ds}/2\right)}, \tag{4.70}$$

where V'_{on} is the linear threshold voltage with the reverse bias on the substrate. It is related to the zero-substrate-bias threshold voltage V_{on} by

$$V'_{on} = V_{on} + (m - 1)R_s I_{ds}, \tag{4.71}$$

where $m - 1$ is the substrate sensitivity [Eq. (4.32)]. In a normal CMOS process, the source and drain regions are symmetrical, and therefore $R_s = R_d = R_{sd}/2$, where R_{sd} is the total source–drain parasitic resistance. Using Eqs. (4.67)–(4.71), one can write the externally measured total device resistance as

$$R_{tot} \equiv \frac{V_{ds}}{I_{ds}} = R_{sd} + R_{ch} = R_{sd} + \frac{L_{mask} - \Delta L}{\mu'_{eff} C_{ox} W (V_g - V_{on} - m V_{ds}/2)}. \tag{4.72}$$

Here all the internal voltages have been replaced by the voltages at the external terminals, since $V'_g - V'_{on} - m V'_{ds}/2 = V_g - V_{on} - m V_{ds}/2$ from Eqs. (4.68), (4.69), and (4.71). Note that V_{on} is defined in terms of the intrinsic device, i.e., the one that would be obtained from linear extrapolation if there were no parasitic resistances.

For a set of devices with different L_{mask} but the same W, the parameters R_{sd}, ΔL, and C_{ox} are the same within process tolerances. It is also assumed that μ'_{eff} does not change with channel length, an assumption which will be examined later. The linear threshold voltage V_{on}, however, does depend on channel length because of short-channel effects. When comparing R_{tot} of devices with different L_{mask}, therefore, it is important to measure V_{on} for each device and adjust V_g so that the gate overdrive $V_g - V_{on}$ is the same from device to device. ***A plot of R_{tot} (at small V_{ds}) versus L_{mask} for a given $V_g - V_{on}$ will then yield a straight line that passes through the point (ΔL, R_{sd}).*** An example is shown in Fig. 4.21. The slope of the line depends on the specific value of the gate overdrive. ΔL and R_{sd} are determined by the common intercept of several lines, each for a different $V_g - V_{on}$ (Chern *et al.*, 1980).

4.3.2.2 SHIFT-AND-RATIO METHOD

Despite the simplicity of the channel-resistance method described above, two main issues remain. First, it is not always straightforward to find the intrinsic V_{on} of short-channel devices. The presence of R_{sd} adds considerable difficulty in the usual linear extrapolation of V_{on} from the measured I_{ds}–V_g curve (Sun *et al.*, 1986). Typically, one tends to underestimate V_{on} in short-channel devices, as the degradation of I_{ds} by R_{sd} is more severe at higher currents. This introduces errors in channel-length extraction. The problem is further aggravated by a strong dependence of mobility on gate voltage, for example, in low-temperature and/or 0.1-μm MOSFETs. The second problem with the resistance method is that the R_{tot}-versus-L_{mask} lines for different gate overdrives may not intersect at a common point. Significant errors may result if only a limited number of $V_g - V_{on}$ are investigated.

An improved channel-length extraction algorithm, called the *shift-and-ratio* (S&R) method, is able to circumvent the above problems (Taur *et al.*, 1992). This method is based on the same channel-resistance concept described above. It starts

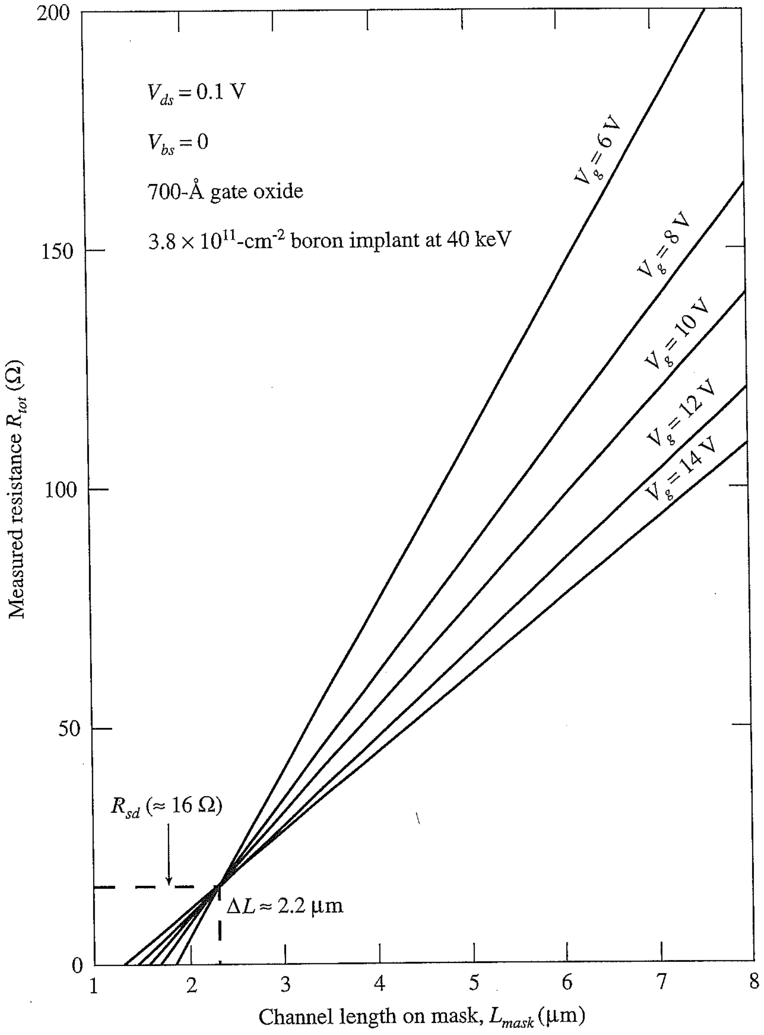


FIGURE 4.21. Measured R_{tot} at a low drain voltage versus L_{mask} for several different values of $V_g - V_{on}$. The common intercept determines both ΔL and R_{sd}. (After Chern *et al.*, 1980.)

with a generalization of Eq. (4.72) to the form

$$R^i_{tot}(V_g) = R_{sd} + L^i_{eff} f\left(V_g - V^i_{on}\right), \tag{4.73}$$

where f is a general function of gate overdrive common to all the measured devices. The superscript i denotes the ith device, with an unknown effective channel length $L^i_{eff} = L^i_{mask} - \Delta L$ and linear threshold voltage V^i_{on}. The key assumption behind Eq. (4.73) is that the modified effective mobility μ'_{eff} is a common function of $V_g - V_{on}$ for all the measured devices. This is a reasonable assumption in view of Eq. (3.50), Eq. (3.51), Eq. (3.58), and Eq. (3.60). Note that for this argument, V_t and V_{on} are interchangeable as long as their difference is a constant [≈(2–3)kT/q], independent of channel length. Strictly speaking, short-channel devices have a slightly higher mobility because of the lower threshold voltage, and therefore a lower vertical field ($\mathscr{E}_{eff}$) for the same $V_g - V_{on}$. Assuming a long-channel V_t (or V_{on}) of 0.5 V and a power-supply voltage V_{dd} of 2.0 V, one can estimate the effect of SCE on $\mathscr{E}_{eff}$ using Eq. (3.50). For a V_t (low-drain) roll-off of 100 mV, the vertical field is 14% lower at $V_g = V_t$ and 7% lower at $V_g = V_{dd}$. Since $\mu_{eff} \propto \mathscr{E}_{eff}^{-1/3}$ from Eq. (3.51), an average of $(14\% + 7\%)/2 = 10.5\%$ lower $\mathscr{E}_{eff}$ translates into a mobility increase of only 3.5%. Such a small error is deemed acceptable for most practical purposes of channel-length extraction.

The task is to calculate R_{sd}, L_{eff}^i, and V_{on}^i in Eq. (4.73) from the measured data on $R_{tot}^i(V_g)$. The S&R algorithm simplifies the procedure by differentiating Eq. (4.73) with respect to V_g. Since the parasitic resistance R_{sd} is either independent or a weak function of V_g, its derivative can be neglected:

$$S^i(V_g) \equiv \frac{dR_{tot}^i}{dV_g} = L_{eff}^i \frac{df\left(V_g - V_{on}^i\right)}{dV_g}. \tag{4.74}$$

Here df/dV_g is also a general function of gate overdrive common to all the devices measured. An important benefit of working with the derivatives is that R_{sd} drops completely out of the picture, so it does not matter if R_{sd} varies from device to device as long as it is constant. S&R extraction is usually carried out with two devices: one long-channel and one short-channel. Equation (4.74) with superscript i represents the short-channel device, while superscript 0 refers to the long-channel device:

$$S^0(V_g) \equiv \frac{dR_{tot}^0}{dV_g} = L_{eff}^0 \frac{df\left(V_g - V_{on}^0\right)}{dV_g}. \tag{4.75}$$

An example is shown in Fig. 4.22 for two devices: $S^0(V_g)$ for $L_{mask}^0 = 10$ µm, and $S^i(V_g)$ for $L_{mask}^i = 0.25$ µm.

It would have been easy if $V_{on}^i = V_{on}^0$, in which case S^i and S^0 would be similar functions of V_g and L_{eff}^i would be simply obtained from the ratio $S^i/S^0 = L_{eff}^i/L_{eff}^0 \approx L_{eff}^i/L_{mask}^0$. In general, however, $V_{on}^i \neq V_{on}^0$, and the two S-functions must be shifted with respect to each other before the ratio is taken. For example, in Fig. 4.22, we can shift one curve (S^i) horizontally to the right by a varying amount δ and compute the ratio between the two curves,

$$r(\delta, V_g) \equiv \frac{S^0(V_g)}{S^i(V_g - \delta)}, \tag{4.76}$$

as a function of V_g. The purpose here is to find the δ-value for which the ratio r is a constant, independent of V_g. If δ is zero or too small, r is a monotonically decreasing function of V_g. On the other hand, if δ is too large, r becomes a monotonically increasing function of V_g. These are shown in Fig. 4.23. ***Only when S^i is shifted by an amount δ equal to the threshold voltage difference between the two devices,***

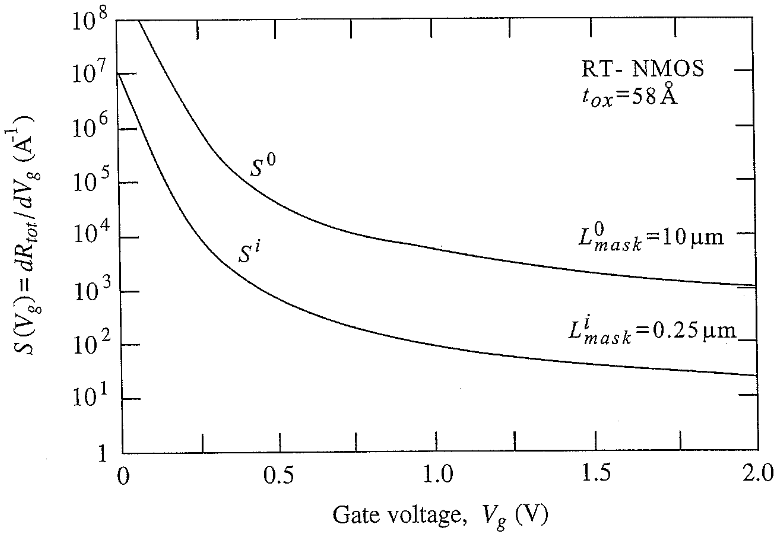


FIGURE 4.22. Examples of $S(V_g) = dR_{tot}(V_g)/dV_g$ curves measured from long-channel ($L_{mask} = 10$ μm) and short-channel ($L_{mask} = 0.25$ μm) devices. (After Taur *et al.*, 1992.)

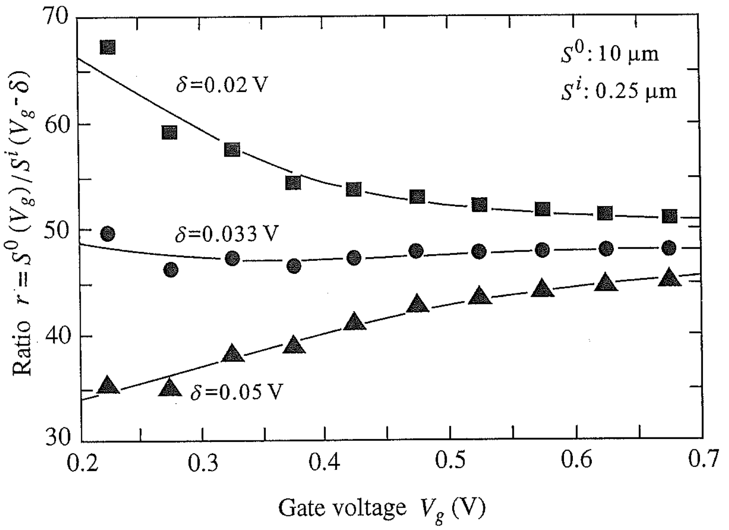


FIGURE 4.23. Curves of $r(\delta, V_g)$ (ratio of S-functions after shift) versus V_g for three different amounts of shift δ. The data are taken from the device examples in Fig. 4.22.

$V^0_{on} - V^i_{on}$*, does *r* become nearly independent of *V_g. Once the correct shift is found, it is a simple matter to find L^i_{eff} from the ratio r evaluated at that shift.

The above procedure can be automated by computing the average r and the mean square deviation of r from its average value, i.e.,

$$\langle r \rangle = \frac{\int_{\Delta V_g} r(\delta, V_g)\, dV_g}{\int_{\Delta V_g} dV_g} \tag{4.77}$$

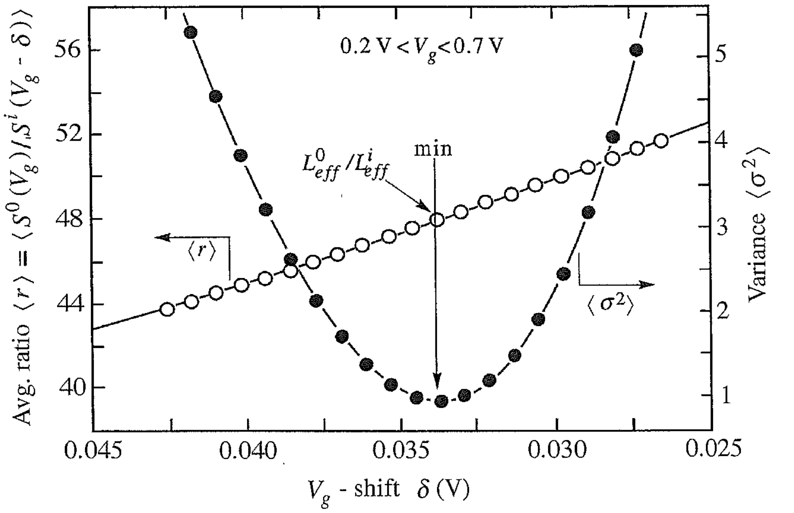


FIGURE 4.24. Average ratio $\langle r \rangle$ (open circles) and variance $\langle \sigma^2 \rangle$ (solid dots) versus shift δ. The data are taken from the device examples in Fig. 4.22 and Fig. 4.23. (After Taur *et al.*, 1992.)

and

$$\langle \sigma^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2, \tag{4.78}$$

as functions of δ for a selected range of gate voltage, ΔV_g. Figure 4.24 plots the example case where the $\langle \sigma^2 \rangle$-versus-δ curve exhibits a sharp minimum at the point of best match (constant ratio between S^0 and S^i). This occurs at

$$\delta_{\min} = V^0_{on} - V^i_{on}, \tag{4.79}$$

and the channel length L^i_{eff} (or ΔL) can be obtained from the average ratio $\langle r \rangle$ at this point:

$$\langle r \rangle_{\delta_{\min}} = \frac{L^0_{eff}}{L^i_{eff}} = \frac{L^0_{mask} - \Delta L}{L^i_{mask} - \Delta L}. \tag{4.80}$$

Note that even if ΔL is different between the long-channel and the short-channel devices, very little error is introduced, as $\Delta L \ll L^0_{mask}$ and hence $L^0_{eff} \approx L^0_{mask}$, insensitive to ΔL.

Once $\delta_{\min}$ and $\langle r \rangle_{\delta_{\min}}$ are found, R_{sd} can be calculated from the measured resistances by using Eq. (4.73) and the corresponding equation for the long-channel device (superscript 0):

$$R_{sd} = \frac{\langle r \rangle_{\delta_{\min}} R^i_{tot}(V_g - \delta_{\min}) - R^0_{tot}(V_g)}{\langle r \rangle_{\delta_{\min}} - 1}. \tag{4.81}$$

For a given long-channel reference, the above extraction procedure can be repeated for a number of short-channel devices with different L^i_{mask}. A by-product of the process is the low-drain short-channel threshold-voltage roll-off given by $\delta_{\min}$. The S&R algorithm has been shown to yield consistent results for effective channel lengths down to the 0.1-μm range (Taur *et al.*, 1992). Although it is not necessary to assume a constant ΔL and R_{sd} for the S&R algorithm to work, comparable ΔL and R_{sd} results among different devices are a good indication that the channel-length extraction has been carried out properly.

It is important in the S&R algorithm to choose a proper gate voltage range ΔV_g for $\langle r \rangle$ and $\langle \sigma^2 \rangle$ calculations.The subthreshold region should be avoided, where the current conduction mechanism is different from that in the linear region (Nguyen, 1984). ***A good choice for ΔV_g is from about V_{on} + 0.2 V all the way up to V_{dd}, which covers most regions of interest in the I_{ds}–V_g curve***. The sensitivity of the extracted L_{eff} to ΔV_g depends on how abrupt the lateral source–drain profile is. This is discussed in more detail in the next subsection.

4.3.3 PHYSICAL MEANING OF EFFECTIVE CHANNEL LENGTH

This subsection examines the physical meaning of L_{eff} extracted from electrically measured terminal currents. The effective channel length is defined through the linear channel resistance by Eq. (4.66). This equation is derived for long-channel devices and is not strictly valid for short-channel devices. By its definition, L_{eff} represents a measure of the effective current-carrying capability of the device and is not associated with any fixed physical quantity. When the channel profile is reasonably uniform and the source–drain doping is not too graded, L_{eff} is approximately equal to L_{met} (Laux, 1984). In general, however, one cannot take $L_{eff} = L_{met}$ for granted, especially in very short-channel MOSFETs.

An example is illustrated in Fig. 4.25, where L_{eff}, extracted (by the S&R algorithm) from currents calculated using a 2-D device simulator, is plotted versus L_{met} for a variety of source–drain and channel doping conditions (Taur *et al.*, 1995b). A 2-D Gaussian profile is used to simulate the falloff of the source and drain doping concentration near the gate edge:

$$N_d(x, y) = N_0 e^{-(x-x_0)^2/2\sigma_V^2} e^{-(y-y_0)^2/2\sigma_L^2}, \tag{4.82}$$

where x is in the vertical direction and y is in the lateral direction. The junction depth x_j is mainly determined by the vertical straggle σ_V. A key parameter for L_{eff} is the lateral source–drain doping gradient characterized by the lateral straggle σ_L. For each of the doping cases in Fig. 4.25, L_{eff} varies with L_{met} linearly with a slope of one. In other words, $L_{eff} - L_{met}$ is essentially independent of L_{met}, indicating that the linear relationship between R_{tot} and L_{mask} assumed in Fig. 4.21 and Eq. (4.73) can be extended to short-channel devices. However, $L_{eff} - L_{met}$ varies considerably with the doping profile. For an infinitely abrupt source–drain

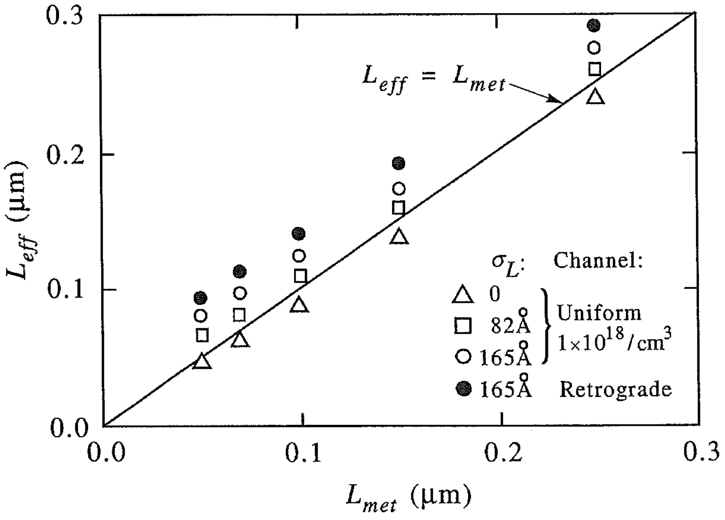


FIGURE 4.25. L_{eff} extracted from simulated currents versus L_{met} for four doping cases. Open symbols are for a uniformly doped channel with different lateral source–drain gradients. Solid dots are for a retrograde-doped channel. (After Taur *et al.*, 1995b.)

profile ($\sigma_L = 0$), $L_{eff} - L_{met}$ is slightly negative. As the lateral straggle σ_L increases, $L_{eff} - L_{met}$ becomes increasingly more positive. The difference grows even larger in the retrograde channel case where L_{eff} is significantly longer than L_{met}. Such a deviation can be understood in terms of the spatial dependence of channel sheet resistivity as discussed below.

4.3.3.1 SHEET RESISTIVITY IN SHORT-CHANNEL DEVICES

Equation (4.66) implicitly assumes that the sheet resistivity, ρ_{ch} given by Eq. (3.101), is uniform in both the MOSFET width and length directions. If the device is wide enough, ρ_{ch} can be considered uniform in that direction. However, the variation of ρ_{ch} in the length direction cannot be ignored in a short-channel device. From Eq. (3.8),

$$I_{ds} = -\mu_{eff} W Q_i(y) \frac{dV}{dy}, \tag{4.83}$$

where $V(y)$ is the quasi-Fermi level at a point y along the channel length direction. I_{ds} is a constant independent of y as required by current continuity. One can define a laterally varying sheet resistivity as

$$\rho_{ch}(y) = \frac{dV/dy}{I_{ds}/W} = \frac{1}{-\mu_{eff} Q_i(y)}. \tag{4.84}$$

Note that $Q_i < 0$ for nMOSFETs. This expression is valid as long as the current flow is largely parallel to the y-direction and the equipotential contours are perpendicular to the silicon surface. The total resistance is given by

$$\frac{V_{ds}}{I_{ds}} = \frac{1}{W}\int_{S-D} \rho_{ch}(y)\,dy, \qquad (4.85)$$

where the integration is carried out from the heavily doped source region to the heavily doped drain region.

Figure 4.26 plots $\rho_{ch}(y)$ calculated from the 2-D device simulator versus distance from the source to the drain for $L_{met} = 0.10$ µm at three different gate voltages (Taur *et al.*, 1995b). The area under each curve gives the total source-to-drain resistance

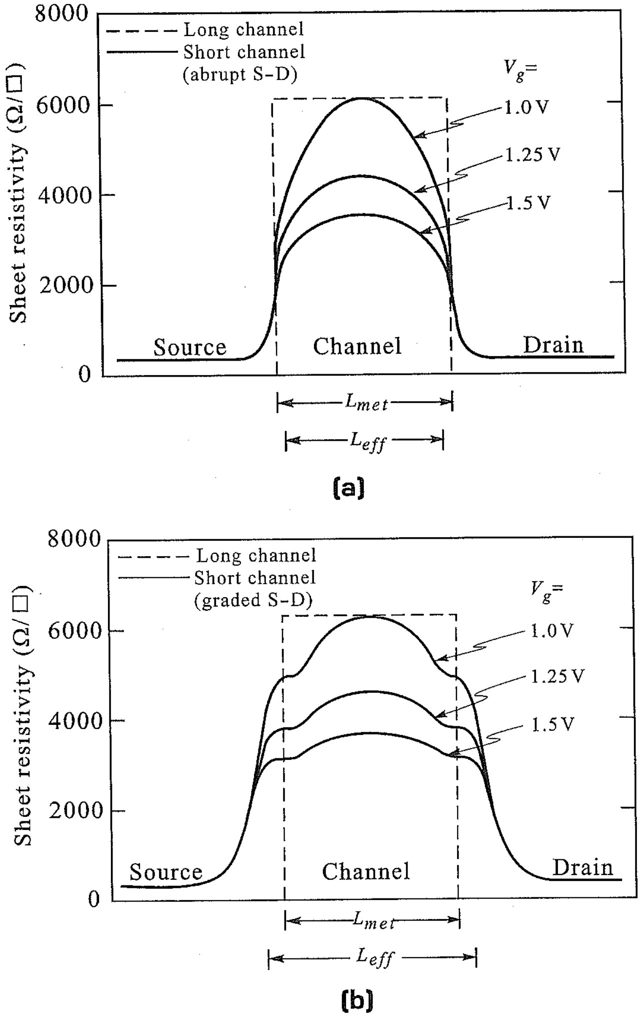


FIGURE 4.26. Simulated channel sheet resistivity at three different gate voltages versus distance from source to drain of an $L_{met} = 0.10$-µm MOSFET. The curves in (a) are for an infinitely abrupt (laterally) source–drain in which $L_{eff} = 0.091$ µm. The curves in (b) are for a graded ($\sigma_L = 165$ Å) source–drain in which $L_{eff} = 0.124$ µm. In both cases, the dashed lines represent the ideal, uniform-sheet resistivity of a scaled long-channel device. (After Taur *et al.*, 1995b.)

ases,
ırger
ıch a
ꞩheet

ı by
f the
ever,
ınnel

ł.83)

tion.
ɜfine

ł.84)

flow
ɔular

as indicated by Eq. (4.85). In Fig. 4.26(a) for an infinitely abrupt (laterally) source–drain junction, the sheet resistivity is modulated by gate voltage inside the (metallurgical) channel and independent of gate voltage outside the (metallurgical) channel. However, in contrast to a long-channel device, $\rho_{ch}(y)$ is highly nonuniform, with a peak near the middle of the channel and decreasing toward the edges. This is due to SCEs from the source–drain fields, which help lower the potential barrier near the junctions and raise the local inversion charge density (Wordeman *et al.*, 1985). This effect is more pronounced at low gate voltages near threshold. The resulting L_{eff} extracted by the S&R method is slightly shorter than L_{met}.

Figure 4.26(b) shows similar plots for the same $L_{met} = 0.10\ \mu\text{m}$, but with a finite lateral source–drain gradient. $\rho_{ch}(y)$ again is nonuniform inside the channel, being modulated by the gate voltage. In this case, however, a nonnegligible portion of the sheet resistivity outside the metallurgical channel is also gate-voltage-dependent. This is because of accumulation (Section 2.3.1) or gate modulation of the series resistance associated with the finite source–drain doping gradient. Since, according to the L_{eff} definition in Eq. (4.73), any part of the sheet resistivity that is gate-voltage-dependent contributes to the effective channel length, the extracted L_{eff} is substantially longer than L_{met}. At the same time, the extracted R_{sd}, which represents the constant part of the resistance in Eq. (4.73), only accounts for a portion of the series resistance outside the metallurgical channel.

4.3.3.2 GATE-MODULATED ACCUMULATION-LAYER RESISTANCE

Because of the finite lateral gradient of source–drain doping in practical devices, current injection from the surface inversion layer into the bulk source–drain region does not occur immediately at the metallurgical junction. When the gate voltage is high enough to turn on the MOSFET channel, an n$^+$ surface accumulation layer is also formed in the gate-to-source or -drain overlap region, as shown schematically in Fig. 4.27 (Ng and Lynch, 1986). Near the metallurgical junction and away from

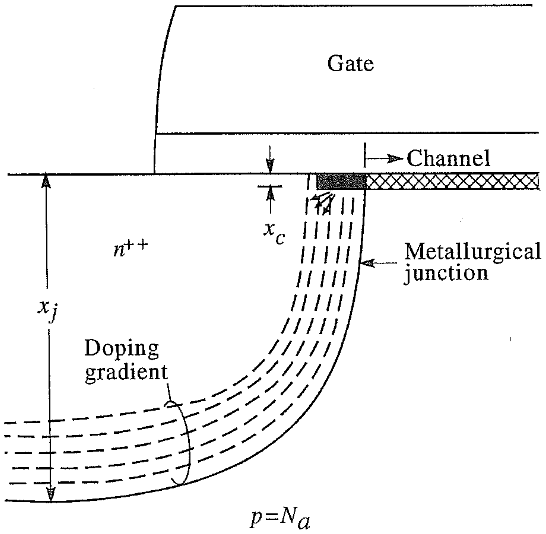


FIGURE 4.27. Schematic diagram showing doping distribution and current flow pattern near the end of the channel and the beginning of the source or drain. The dashed lines are contours of constant donor concentration, i.e., constant resistivity. The dark region represents the accumulation layer. (After Ng and Lynch, 1986.)

the surface, the donor concentration (also compensated by the p-type background) is low and the conductivity of the accumulation layer is higher than that of the bulk source–drain. As a result, current flow stays in the accumulation layer near the surface. This continues until the source–drain doping becomes high enough that the bulk conductance exceeds that of the accumulation layer. The point or region of current injection into the bulk depends on the lateral source–drain doping gradient. The more graded the profile is, the farther away the injection point is from the metallurgical junction.

The sheet resistivity of the accumulation layer can be estimated by applying Eq. (2.180) to the gate-to-source–drain overlap region:

$$V_g = V_{fb} + \psi_s - \frac{Q_{ac}}{C_{ox}}, \qquad (4.86)$$

where $Q_{ac} < 0$ is the accumulation charge (electrons) per unit area induced by the gate field, ψ_s is the band bending at the surface with respect to the bulk n-type region, and V_{fb} is the flat-band voltage largely determined by the work-function difference between the gate electrode and the n-type silicon. For an n$^+$-polysilicon-gated nMOSFET, $V_{fb} = -E_g/2q + \psi_B$, where ψ_B is given by Eq. (2.37) in terms of the local n-type doping concentration. The band bending in accumulation is approximately given by the distance between the n-type Fermi level and the conduction-band edge, i.e., $\psi_s \approx E_g/2q - \psi_B$. Therefore, V_{fb} and ψ_s in Eq. (4.86) nearly cancel each other and one obtains $V_g \approx -Q_{ac}/C_{ox}$. The sheet resistivity of the accumulation layer is then

$$\rho_{ac} = \frac{1}{\mu_{ac}|Q_{ac}|} = \frac{1}{\mu_{ac}C_{ox}V_g}, \qquad (4.87)$$

where μ_{ac} is the average electron mobility in the accumulation layer. If, for process reasons (gate reoxidation), the oxide thickness in the gate to source–drain overlap region is different from t_{ox} in the channel region, a different C_{ox} should be used in Eq. (4.87). The electron mobility in the accumulation layer has a similar field dependence to that in the inversion layer (Sun and Plummer, 1980). From Gauss's law, the average electric field in the accumulation layer is $\mathscr{E}_{eff} = |Q_{ac}|/2\varepsilon_{si} = C_{ox}V_g/2\varepsilon_{si}$. Knowing $\mathscr{E}_{eff}$, one can look up μ_{ac} from Fig. 3.13. Like the channel mobility, μ_{ac} is not limited by the impurity scattering which usually dominates the bulk mobility (Fig. 2.7) for the moderately high doping levels ($N_d \approx 10^{18}$ cm^{-3}) at the surface. This is because of the screening of Coulomb scattering when the carrier concentration in the accumulation layer greatly exceeds the donor concentration at that point.

4.3.3.3 INTERPRETATION OF L_{eff} IN TERMS OF CURRENT INJECTION POINTS

The dependence of ρ_{ac} on V_g in Eq. (4.87) is too similar to that of ρ_{ch} in Eq. (3.101) to allow separation of the accumulation-layer resistance from the channel resistance.

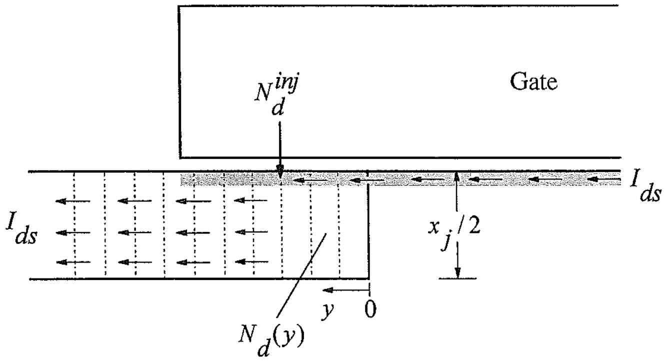


FIGURE 4.28. Schematic diagram of a simple 1-D model for estimating the source–drain doping concentration at the point of current injection from the surface into the bulk. As far as the resistance is concerned, the source or drain region is modeled as uniform stripes of doping concentration $N_d(y)$ and width $x_j/2$.

The region where the current flows predominantly in the accumulation layer is therefore considered as a part of L_{eff}. To estimate the source–drain doping concentration at which current injection into the bulk takes place, we consider a simple 1-D current model shown in Fig. 4.28. The donor concentration at the surface is assumed to be $N_d(y)$, which varies along the direction of current flow as dictated by the lateral doping gradient. If the bulk resistivity corresponding to N_d is $\rho(N_d)$ and the source–drain junction depth is x_j, the average sheet resistivity of a thin stripe perpendicular to the current flow is approximately $\rho(N_d)/(x_j/2)$. Here the effective depth of uniform current flow is taken as $x_j/2$, since for any given stripe the n-type conductivity is highest at the surface and drops to zero at x_j. As N_d increases toward the heavily doped source–drain region, $\rho(N_d)/(x_j/2)$ decreases accordingly. Current injection into the bulk takes place at $N_d = N_d^{inj}$, where the sheet resistivity $\rho(N_d^{inj})/(x_j/2)$ equals ρ_{ac} of Eq. (4.87). Since ρ_{ac} depends on the gate voltage, so does the point of injection (Hu *et al.*, 1987). At low gate overdrives, the injection point is closer to the metallurgical junction edge. As the gate voltage increases, the injection point moves out toward the more heavily doped source–drain region. For reasonably abrupt source–drain doping profiles, $N_d(y)$ is a strong (exponential) function of y and the injection points do not spread too far apart. If a V_g-range from slightly above V_t to V_{dd} is used in the S&R algorithm, the resulting L_{eff} is, to the first-order approximation, given by the injection point corresponding to $V_g \approx V_{dd}/2$ in the middle of the V_g-range. Setting $\rho(N_d^{inj})/(x_j/2)$ equal to ρ_{ac} of Eq. (4.87) with $V_g = V_{dd}/2$, one can then find N_d^{inj} from

$$\rho(N_d^{inj}) = \frac{x_j}{\mu_{ac} C_{ox} V_{dd}}, \tag{4.88}$$

where μ_{ac} is evaluated at an average normal field of $\mathscr{E}_{eff} = C_{ox} V_{dd}/4\varepsilon_{si}$.

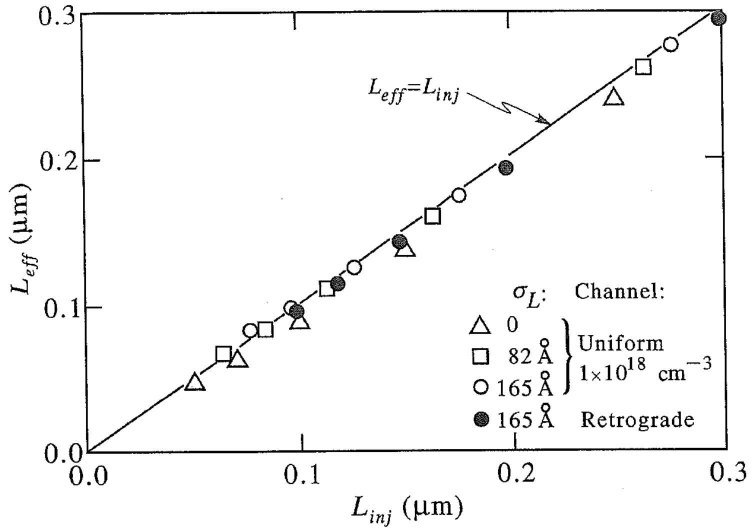


FIGURE 4.29. Same L_{eff} data as in Fig. 4.25, but plotted versus L_{inj} in terms of current injection points. Open symbols are for a uniformly doped channel with different lateral source–drain gradients. Solid dots are for a retrograde-doped channel.

In general, N_d^{inj} increases as the device dimensions are scaled down. For 1-μm CMOS technology, $N_d^{inj} \approx 10^{17}$ cm^{-3}. For the 0.1-μm devices in Fig. 4.25, with $V_{dd} = 1.5$ V, $t_{ox} = 30$ Å, $x_j = 500$ Å, one obtains $\mathscr{E}_{eff} = 4.2 \times 10^5$ V/cm, $\mu_{ac} =$ 420 cm^2/V-s (Fig. 3.13), and $\rho(N_d^{inj}) = 0.007\ \Omega$-cm. The last number corresponds to $N_d^{inj} \approx 8 \times 10^{18}$ cm^{-3} from the n-type resistivity curve in Fig. 2.8. Figure 4.29 replots the simulated L_{eff} data in Fig. 4.25 against L_{inj}, defined as the distance between the points where the source–drain doping equals N_d^{inj}. All the points lie within 100 Å of the $L_{eff} = L_{inj}$ line, independent of the lateral doping gradient and channel profile. This supports the physical interpretation of L_{eff} in terms of the current injection points from the accumulation layer.

4.3.3.4 IMPLICATIONS FOR SHORT-CHANNEL EFFECTS

The fact that L_{eff} can be much longer than L_{met} has significant implications for the short-channel V_t roll-off curves. Figure 4.30 shows the low-drain threshold voltage roll-off versus L_{eff} for several different source–drain doping gradients. The abrupt doping profile has the best short-channel effect. As the lateral straggle increases, the short-channel effect becomes progressively worse. This can be understood from Fig. 4.31, where the net doping concentration $N_d - N_a$ at the surface is plotted along the channel length direction for the various source–drain profiles studied. For a given L_{eff} or L_{inj} (=0.1 μm), the distance between the points where the doping concentration falls to N_d^{inj} (=8×10^{18} cm^{-3} in this case) is fixed. It is clear that the more graded the source–drain profile is, the deeper the n-type doping tail penetrates into the channel and compensates or reverses the p-type doping inside

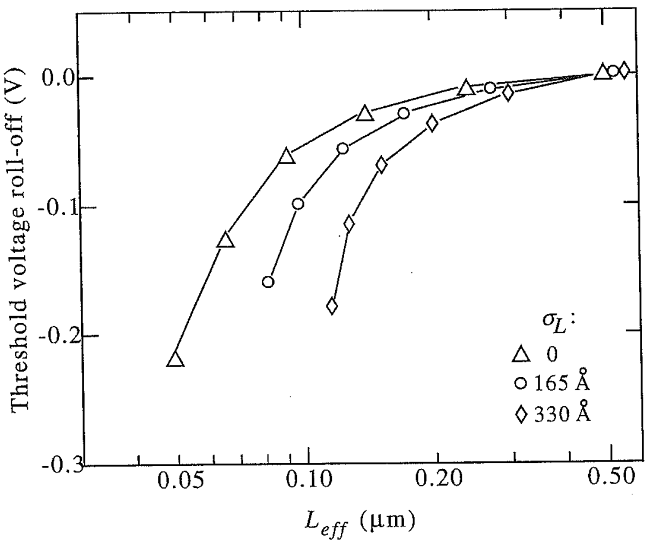


FIGURE 4.30. Simulated short-channel threshold roll-off versus L_{eff} for three different lateral source–drain doping gradients. On each curve, the points are for $L_{met} =$ 0.05, 0.07, 0.10, 0.15, 0.25, and 0.50 μm. (After Taur *et al.*, 1995b.)

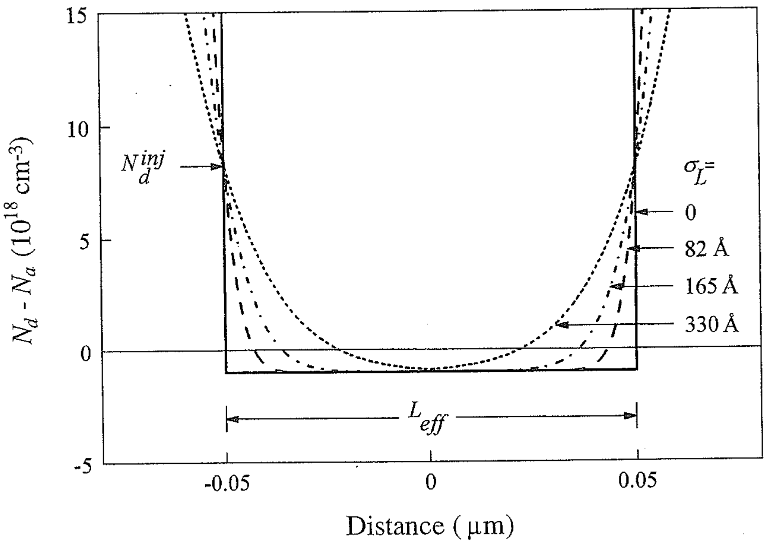


FIGURE 4.31. Net n-type concentration versus distance from source to drain at the surface of a 0.1-μm (L_{inj} or L_{eff}) nMOSFET. The injection points (N_d^{inj}) are kept the same for different lateral doping gradients.

the channel. This is detrimental to the short-channel effect, as the edge regions become more easily depleted and inverted by the source–drain fields (opposite to the halo effect). ***It is therefore very important to reduce the width of the (laterally) graded source–drain region as the channel length is scaled down.***

4.3
Ligl
exte
higl
dop
Cur
plac
mat
L_{me}
I
a pa
thro
dep
brea
0.25
inte
acc
R_{sd}
up
bec
T
I_{ds}–
and
thes
figu
leav

4.3
BY
In a
fror
and
as t
inte
for
T
MC
the

Hei

4.3.3.5 EFFECTIVE CHANNEL LENGTH OF LDD DEVICES

Lightly-doped-drain (LDD) MOSFETs (Ogura *et al.,* 1982) are designed with an extended moderately doped (10^{17}–10^{18}-cm^{-3} range) source–drain region to relieve high electric fields and related hot-electron effects. The presence of such a lightly doped region poses great difficulties in the interpretation of effective channel length. Current injection from the surface accumulation layer into the bulk region takes place over an extended distance with a location dependent on gate voltage. No matter what kind of L_{eff} extraction method is used, L_{eff} is significantly longer than L_{met}, especially at high gate overdrives (Sun *et al.*, 1986).

In one approach, channel resistance extraction (Fig. 4.21) is carried out for a pair of closely spaced V_g-values (Hu *et al.*, 1987). The procedure is repeated throughout the entire voltage range, allowing both L_{eff} and R_{sd} to be gate-voltage-dependent. Similar methods can be implemented in the S&R algorithm as well by breaking up the full voltage range used in Eq. (4.77) into small intervals of 0.5 or 0.25 V each. A set of (L_{eff}, R_{sd}) can be extracted from the current data in each interval by repeating the S&R algorithm. However, most circuit models do not accommodate gate-voltage-dependent channel lengths. Furthermore, once L_{eff} and R_{sd} are allowed to be gate-voltage-dependent, there is no unique way of breaking up R_{ch} and R_{sd} from the measured R_{tot}, and the solutions to Eq. (4.72) or Eq. (4.73) become ambiguous and method-dependent.

There is no consensus on how to define the L_{eff} of an LDD MOSFET. If the entire I_{ds}–V_g characteristics can be fitted within acceptable tolerances by a constant L_{eff} and a constant R_{sd} based on Eq. (4.73), a circuit model can be formulated with these parameters. Otherwise, one needs to define a constant L_{eff}, perhaps using the figure extracted from the current data at low gate overdrives (Sun *et al.*, 1986), and leave the rest of the resistance as a gate-voltage-dependent series resistance.

4.3.3.6 EXTRACTION OF CHANNEL LENGTH BY *C–V* MEASUREMENTS

In an entirely different approach, another type of channel length has been extracted from the measured *C–V* data of a series of MOSFETs with different L_{mask} (Sheu and Ko, 1984). Capacitance measurements in general are more difficult to perform, as they require specially designed test sites. It is by no means straightforward to interpret the capacitively measured channel length and apply it to circuit models for current calculations.

The capacitive extraction of channel length is based on the fact that when a MOSFET is turned on, the intrinsic gate-to-channel capacitance is proportional to the channel length:

$$C_{gc} = C_{ox} W L_{cap}. \tag{4.89}$$

Here L_{cap} is the capacitively defined channel length, which may or may not be

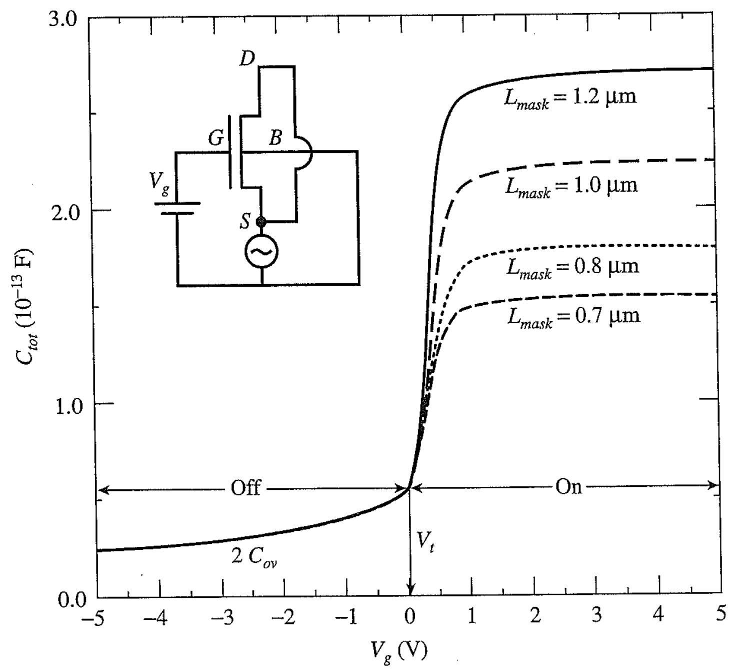


FIGURE 4.32. Example of measured capacitance from gate to source–drain versus gate voltage for MOSFETs of different mask lengths. The inset shows the split *C–V* measurement setup. (After Guo *et al.*, 1994.)

the same as L_{eff} or L_{met}. The gate-to-channel capacitance is usually measured in a split *C–V* setup that separates the majority-carrier response from the minority-carrier response, as shown in the inset of Fig. 4.32. The total measured capacitance consists of both the intrinsic gate-to-channel capacitance and a parasitic overlap capacitance from the gate to source–drain which is independent of channel length:

$$C_{tot} = C_{gc} + 2C_{ov} = C_{ox}WL_{cap} + 2C_{ov}. \tag{4.90}$$

Here C_{ov} is the overlap capacitance per gate edge (see Fig. 5.16). Typical examples of C_{tot}–V_g curves are shown in Fig. 4.32 (Guo *et al.*, 1994). Using a large-area MOS capacitor, one can easily calibrate C_{ox}, taking all the polysilicon depletion and inversion–layer quantum effects into account. To find out L_{cap}, it is critical to determine what $2C_{ov}$ to subtract from the measured C_{tot}. In principle, $2C_{ov}$ in Eq. (4.90) is the parasitic capacitance at a gate voltage when the MOSFET is on and C_{gc} is given by Eq. (4.89). In practice, $2C_{ov}$ cannot be separated from C_{gc}, since, unlike channel resistance, channel capacitance does not vary significantly with gate voltage once the device is turned on. What is usually done is to take $2C_{ov}$ as the measured capacitance when the MOSFET is off. However, from Fig. 4.32 it is clear that $2C_{ov}$ varies with the gate voltage (Oh *et al.*, 1990). There is no guarantee that $2C_{ov}$ in the off state is the same as $2C_{ov}$ in the on state. If $2C_{ov}$ is taken as the capacitance right below the threshold voltage, it will contain an unwanted inner-fringe term that is absent when the conducting channel is formed. If $2C_{ov}$ is taken at a negative gate voltage where the substrate is accumulated to eliminate the inner-fringe component, the lightly doped source–drain in the direct overlap region will be depleted (Sheu and Ko, 1984). Any such error in $2C_{ov}$ translates into a large error in L_{cap} when dealing with short-channel devices having small intrinsic capacitances.

A better interpretation of the capacitively extracted channel length is in terms of the gate length, L_{gate} (Fig. 4.19), since as the gate voltage varies, the same amount of charge per unit area is induced at the silicon surface whether it is in the inversion channel or in the source–drain overlap region under the gate. In other words, as far as the capacitance is concerned, the direct overlap length should be lumped into the channel length. This also circumvents the problem with the inner-fringe component mentioned above. One still needs to estimate the outer fringe capacitance and subtract it from the measured capacitance. But this can be done using a simple formula (Section 5.2.2) and therefore should have less error associated with it.

EXERCISES

4.1 Apply constant-field scaling rules to the long-channel currents [Eq. (3.19) for the linear region and Eq. (3.23) for the saturation region], and show that they behave as indicated in Table 4.1.

4.2 Apply constant-field scaling rules to the subthreshold current, Eq. (3.36), and show that instead of decreasing with scaling ($1/\kappa$), it actually increases with scaling (note that $V_g < V_t$ in subthreshold). What if the temperature is also scaled down by the same factor ($T \rightarrow T/\kappa$)?

4.3 Apply constant-field scaling rules to the saturation current from the $n = 1$ velocity saturation model [Eq. (3.78)] and the fully saturation-velocity limited current [Eq. (3.80)], and show that they behave as indicated in Table 4.1.

4.4 Apply generalized scaling rules to the saturation current from the $n = 1$ velocity saturation model [Eq. (3.78)], and show that it behaves as indicated in Table 4.3 (between the two limits).

4.5 In Eqs. (4.59) and (4.60), how should x_{av} be defined in terms of $n(x)$, the electron volume concentration as a function of depth? The same definition applies to x_j in Eq. (4.50) and x_{av}^{QM} in Eq. (4.58).

4.6 *Nonuniform V_t in the width direction.* A MOSFET is nonuniformly doped in the width direction. Part of the width (W_1) has a linear threshold voltage V_{on1}. The other part of the width (W_2) has a linear threshold voltage V_{on2}. Show that as far as the linear region characteristics [Eq. (3.61)] are concerned, this device is equivalent to a uniform MOSFET of width $W_1 + W_2$ with a linear threshold voltage $V_{on} = (W_1 V_{on1} + W_2 V_{on2})/(W_1 + W_2)$. Ignore any fringing fields that may exist near the boundary between the two regions.

4.7 *Nonuniform V_t in the length direction.* A MOSFET is nonuniformly doped in the length direction. Part of the length (L_1) has a linear threshold voltage V_{on1}. The other part of the length (L_2) has a linear threshold voltage V_{on2}. Assume $V_{on1} \approx V_{on2}$, and consider only the first-order terms of $V_{on1} - V_{on2}$. Show that as far as the linear region characteristics [Eq. (3.61)] are concerned, this device is equivalent to a uniform MOSFET of length $L_1 + L_2$ with a linear threshold voltage $V_{on} = (L_1 V_{on1} + L_2 V_{on2})/(L_1 + L_2)$. Ignore any fringing fields that may exist near the boundary between the two regions.

4.8 In the top equivalent circuit of Fig. 4.20, the source–drain current can be considered either as a function of the internal voltages: $I_{ds}(V_g', V_{ds}')$, or as a function of the external voltages: $I_{ds}(V_g, V_{ds})$. The internal voltages are related to the external voltages by Eqs. (4.68) and (4.69). Show that the transconductance of the intrinsic MOSFET can be expressed as

$$g_m' \equiv \left(\frac{\partial I_{ds}}{\partial V_g'}\right)_{V_{ds}'} = \frac{g_m}{1 - g_m R_s - g_{ds}(R_s + R_d)},$$

where

$$g_m \equiv \left(\frac{\partial I_{ds}}{\partial V_g}\right)_{V_{ds}}$$

is the extrinsic transconductance, and

$$g_{ds} \equiv \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)_{V_g}$$

is the extrinsic output conductance.

4.9 Show that in the subthreshold region and when the drain bias is low, Eq. (3.12) leads to Eq. (4.54):

$$Q_i = \frac{kTn_i^2}{\mathscr{E}_s N_a} e^{q\psi_s/kT},$$

where ψ_s is the surface potential and $\mathscr{E}_s$ is the surface electric field. This equation is more general than Eq. (3.31) since it is valid for nonuniform (vertically) dopings with N_a being the p-type concentration at the edge of the depletion layer. (Note that the factor N_a merely reflects the fact in Fig. 2.24 that the band bending ψ_s is defined with respect to the bands of the neutral bulk region of doping N_a.)

4.10 In a short-channel device or in a nonuniformly doped (laterally) MOSFET, ψ_s may vary along the channel length direction from the source to drain. Generalize the expression in Exercise 4.9 and show that

$$\frac{V_{ds}}{I_{ds}} = \frac{1}{\mu_{eff} W} \int_0^L \frac{dy}{Q_i(y)} = \frac{N_a}{\mu_{eff} WkTn_i^2} \int_0^L \mathscr{E}_s(y) e^{-q\psi_s(y)/kT}\, dy$$

ed in
V_{on1}.
sume
v that
evice
shold
s that

con-
ction
o the
ce of

(3.12)

equa-
cally)
letion
band
ion of

SFET,
drain.

for the subthreshold region at low drain biases. Since $\mathscr{E}_s(y) \approx [V_g - V_{fb} - \psi_s(y)]/3t_{ox}$ is not a strong function of ψ_s, the exponential factor dominates. This implies that the subthreshold current is controlled by the point of highest barrier (lowest ψ_s) in the channel. It also implies that the *channel length* factor entering the subthreshold current expression is different from the *effective channel length* defined by the linear region characteristics, Eq. (4.66).