Physics of Semiconductor Devices

SECOND EDITION

S. M. Sze

Bell Laboratories, Incorporated
Murray Hill, New Jersey

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

UMC 1413
UMC v. AICP

Copyright © 1981 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

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

Library of Congress Cataloging in Publication Data:

Sze, S. M., 1936-
Physics of semiconductor devices.
"A Wiley-Interscience publication."
Includes index.
1. Semiconductors. I. Title.

TK7871.85.S988 1981 537.6'22 81–213
ISBN 0-471-05661–8 AACR2

Printed in the United States of America

32 33 34 35 36 37 38 39 40

UMC 1413
UMC v. AICP

Contents

INTRODUCTION 1

PART I SEMICONDUCTOR PHYSICS 5

Chapter 1 Physics and Properties of Semiconductors—A Résumé 7

1.1 Introduction, 7
1.2 Crystal Structure, 8
1.3 Energy Bands, 12
1.4 Carrier Concentration at Thermal Equilibrium, 16
1.5 Carrier Transport Phenomena, 27
1.6 Phonon Spectra and Optical, Thermal, and High-Field Properties of Semiconductors, 38
1.7 Basic Equations for Semiconductor Device Operation, 50

PART II BIPOLAR DEVICES 61

Chapter 2 p-n Junction Diode 63

2.1 Introduction, 63
2.2 Basic Device Technology, 64
2.3 Depletion Region and Depletion Capacitance, 74
2.4 Current–Voltage Characteristics, 84
2.5 Junction Breakdown, 96
2.6 Transient Behavior and Noise, 108
2.7 Terminal Functions, 112
2.8 Heterojunction, 122

UMC 1413
UMC v. AICP

Chapter 3 Bipolar Transistor 133

3.1 Introduction, 133
3.2 Static Characteristics, 134
3.3 Microwave Transistor, 156
3.4 Power Transistor, 169
3.5 Switching Transistor, 175
3.6 Related Device Structures, 181

Chapter 4 Thyristors 190

4.1 Introduction, 190
4.2 Basic Characteristics, 191
4.3 Shockley Diode and Three-Terminal Thyristor, 209
4.4 Related Power Thyristors, 222
4.5 Diac and Triac, 229
4.6 Unijunction Transistor and Trigger Thyristors, 234
4.7 Field-Controlled Thyristor, 238

PART III UNIPOLAR DEVICES 243

Chapter 5 Metal–Semiconductor Contacts 245

5.1 Introduction, 245
5.2 Energy-Band Relation, 246
5.3 Schottky Effect, 250
5.4 Current Transport Processes, 254
5.5 Characterization of Barrier Height, 270
5.6 Device Structures, 297
5.7 Ohmic Contact, 304

Chapter 6 JFET and MESFET 312

6.1 Introduction, 312
6.2 Basic Device Characteristics, 314
6.3 General Characteristics, 324
6.4 Microwave Performance, 341
6.5 Related Field-Effect Devices, 351

Chapter 7 MIS Diode and CCD 362

7.1 Introduction, 362
7.2 Ideal MIS Diode, 363
7.3 Si-SiO_2 MOS Diode, 379
7.4 Charge-Coupled Device, 407

UMC 1413
UMC v. AICP

Chapter 8 MOSFET 431

8.1 Introduction, 431
8.2 Basic Device Characteristics, 433
8.3 Nonuniform Doping and Buried-Channel Devices, 456
8.4 Short-Channel Effects, 469
8.5 MOSFET Structures, 486
8.6 Nonvolatile Memory Devices, 496

PART IV SPECIAL MICROWAVE DEVICES 511

Chapter 9 Tunnel Devices 513

9.1 Introduction, 513
9.2 Tunnel Diode, 516
9.3 Backward Diode, 537
9.4 MIS Tunnel Diode, 540
9.5 MIS Switch Diode, 549
9.6 MIM Tunnel Diode, 553
9.7 Tunnel Transistor, 558

Chapter 10 IMPATT and Related Transit-Time Diodes 566

10.1 Introduction, 566
10.2 Static Characteristics, 568
10.3 Dynamic Characteristics, 577
10.4 Power and Efficiency, 585
10.5 Noise Behavior, 599
10.6 Device Design and Performance, 604
10.7 BARITT and DOVETT Diodes, 613
10.8 TRAPATT Diode, 627

Chapter 11 Transferred-Electron Devices 637

11.1 Introduction, 637
11.2 Transferred-Electron Effect, 638
11.3 Modes of Operation, 651
11.4 Device Performances, 667

PART V PHOTONIC DEVICES 679

Chapter 12 LED and Semiconductor Lasers 681

12.1 Introduction, 681
12.2 Radiative Transitions, 682

UMC 1413
UMC v. AICP

12.3 Light-Emitting Diodes, 689
12.4 Semiconductor Laser Physics, 704
12.5 Laser Operating Characteristics, 724

Chapter 13 Photodetectors **743**

13.1 Introduction, 743
13.2 Photoconductor, 744
13.3 Photodiode, 749
13.4 Avalanche Photodiode, 766
13.5 Phototransistor, 782

Chapter 14 Solar Cells **790**

14.1 Introduction, 790
14.2 Solar Radiation and Ideal Conversion Efficiency, 791
14.3 *p-n* Junction Solar Cells, 799
14.4 Heterojunction, Interface, and Thin-Film Solar Cells, 816
14.5 Optical Concentration, 830

APPENDIXES **839**

A. List of Symbols, 841
B. International System of Units, 844
C. Unit Prefixes, 845
D. Greek Alphabet, 846
E. Physical Constants, 847
F. Lattice Constants, 848
G. Properties of Important Semiconductors, 849
H. Properties of Ge, Si, and GaAs at 300 K, 850
I. Properties of SiO_2 and Si_3N_4 at 300 K, 851

INDEX **853**

UMC 1413
UMC v. AICP

8

MOSFET

- INTRODUCTION
- BASIC DEVICE CHARACTERISTICS
- NONUNIFORM DOPING AND BURIED-CHANNEL DEVICES
- SHORT-CHANNEL EFFECTS
- MOSFET STRUCTURES
- NONVOLATILE MEMORY DEVICES

8.1 INTRODUCTION

The metal–oxide–semiconductor field-effect transistor (MOSFET) is the most important device for very-large-scale integrated circuits such as microprocessors and semiconductor memories. MOSFET is also becoming an important power device. It has many acronyms including IGFET (insulated-gate field-effect transistor) MISFET (metal–insulator–semiconductor field-effect transistor) and MOST (metal–oxide–semiconductor transistor). The principle of the surface field-effect transistor was first proposed in the early 1930s by Lilienfeld[1] and Heil.[2] It was subsequently studied by Shockley and Pearson[3] in the late 1940s. In 1960, Kahng and Atalla[4] proposed and fabricated the first MOSFET using a thermally oxidized silicon structure. The basic device characteristics have been subsequently studied by Ihantola and Moll,[5,6] Sah,[7] and Hofstein and Heiman.[8] The technology, application, and device physics have been reviewed by Wallmark and Johnson,[9] Richman,[10] and Brews.[11]

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 device. The MOSFET is a member of the family of field-effect transistors. The other members, JFETs and MESFETs, have already been considered in Chapter 6. Al-

UMC 1413
UMC v. AICP

though MOSFETs have been made with various semiconductors such as Ge,[12] Si, and GaAs,[13] and use various insulators such as SiO_2, Si_3N_4, and Al_2O_3, the most important system is the Si–SiO_2 combination. Hence most of the results in this chapter are obtained from the Si–SiO_2 system.

We first consider the basic device characteristics of the so-called long-channel MOSFET; that is, the channel length L is much longer than the sum of the source and drain depletion-layer widths $(W_S + W_D)$.* This serves as a foundation to understand short-channel, that is, $L \lesssim (W_S + W_D)$, and related MOSFET devices.

Figure 1 shows[14] the reduction of the minimum device dimension since the beginning of the integrated circuit era in 1959. Figure 1 also shows that the minimum dimension will shrink continuously; the 1-μm barrier for commercial devices may be overcome by 1990. The reduction of device dimensions is driven by the requirement that integrated circuits of high complexity be fabricated. The number of components per integrated-circuit

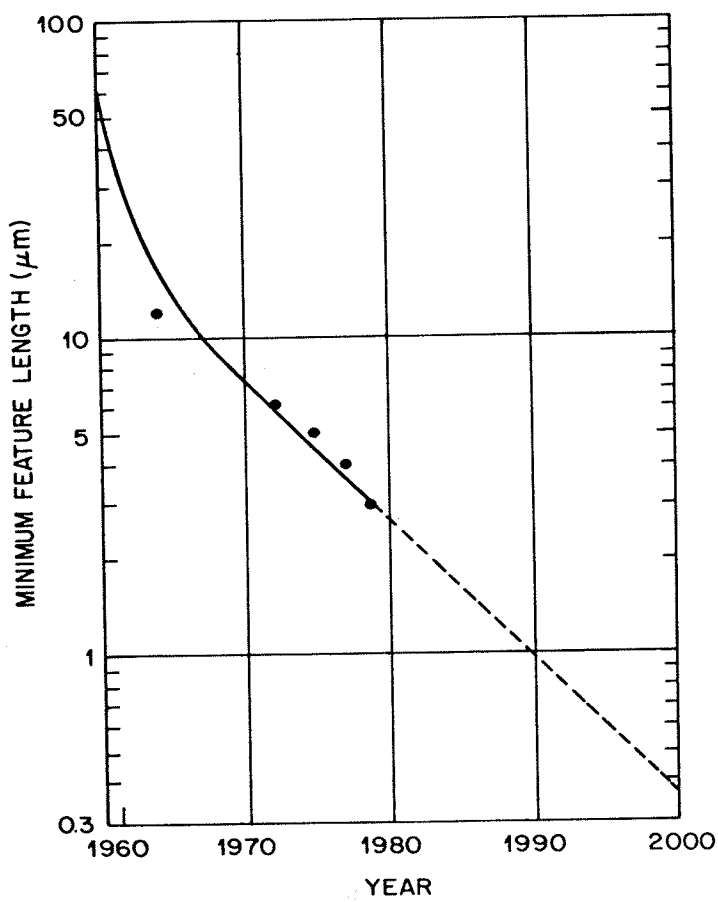

Fig. 1 The minimum device dimension in an integrated circuit as a function of the year for commercial devices. (After Ref. 14.)

*These terms will be defined in Section 8.2.

UMC 1413
UMC v. AICP

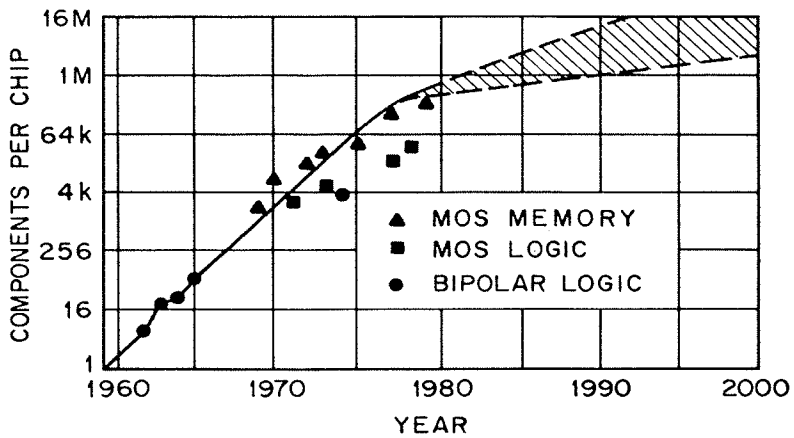

Fig. 2 Complexity of integrated circuits as a function of the year. (After Moore, Ref. 15.)

chip has grown exponentially[15] since 1959 (Fig. 2). The rate of growth is expected to slow down because of a lack of product definition and design. However, a complexity of 1 million or more devices per chip may be available around 1990 using 1-μm or submicron device geometries. As the channel length becomes shorter, one has to consider short-channel effects due to two-dimensional potential, high-field transport and oxide charging. Many device structures have been proposed to improve MOSFET performance. Some representative structures as well as the nonvolatile semiconductor memory, basically a MOSFET with a multilayer gate structure, will be discussed.

8.2 BASIC DEVICE CHARACTERISTICS

The basic structure of a metal–oxide–semiconductor field-effect transistor (MOSFET) is illustrated in Fig. 3. It is a four-terminal device and consists of a p-type semiconductor substrate into which two n^+ regions, the source and drain, are formed* (e.g., by ion implantation). The metal contact on the insulator is called gate; heavily doped polysilicon or a combination of silicide and polysilicon can also be used as the gate electrode. The basic device parameters are the channel length L, which is the distance between the two metallurgical n^+-p junctions; the channel width Z; the insulator thickness d; the junction depth r_j; and the substrate doping N_A. In a silicon integrated circuit, a MOSFET is surrounded by a thick oxide (called the field oxide to distinguish it from the gate oxide) to isolate it from adjacent devices.

The source contact will be used as the voltage reference throughout this

*This is an n-channel device; one may consider a p-channel device by exchanging p for n and reversing the polarity of the voltage.

UMC 1413
UMC v. AICP

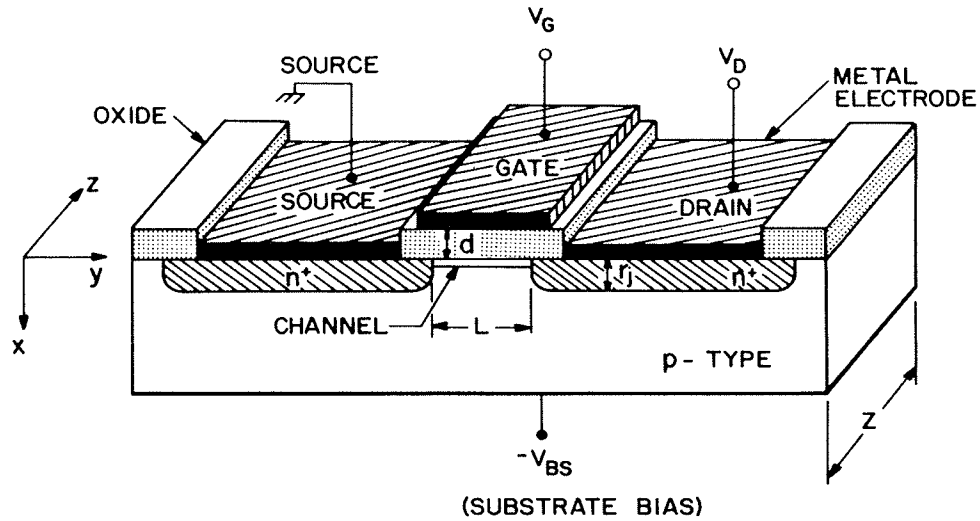

Fig. 3 Schematic diagram of a MOSFET. (After Kahng and Atalla, Ref. 4.)

chapter. When no voltage is applied to the gate, the source-to-drain electrodes correspond to two p-n junctions connected back to back. The only current that can flow from source to drain is the reverse leakage current.* When a sufficiently large positive bias is applied to the gate so that a surface inversion layer (or channel) is formed between the two n^+ regions, the source and the drain are then connected by a conducting-surface n channel through which a large current can flow. The conductance of this channel can be modulated by varying the gate voltage. The back-surface contact (or substrate contact) can have the reference voltage or be reverse-biased; the back-surface voltage will also affect the channel conductance.

8.2.1 Nonequilibrium Condition

When a voltage is applied across the source–drain contacts, the MOS structure is in a nonequilibrium condition; that is, the imref of the minority carriers (electrons, in the present case) is lowered from the equilibrium Fermi level. To show more clearly the band bending across the device, Fig. 4*a* shows[16] the MOSFET turned 90°. The two-dimensional, flat-band, zero-bias ($V_G = V_D = V_{BS} = 0$) equilibrium condition is shown in Fig. 4*b*. The equilibrium conditions under a gate bias that causes surface inversion are shown in Fig. 4*c*. The nonequilibrium condition with both drain and gate biases is shown in Fig. 4*d*, where we note the separation of the imrefs of electrons and holes; the hole imref E_{Fp} remains at the bulk Fermi level while the electron imref E_{Fn} (minority in the present case) is lowered

*This is the n-channel normally-off (enhancement-type) MOSFET. Other types will be discussed later.

UMC 1413
UMC v. AICP

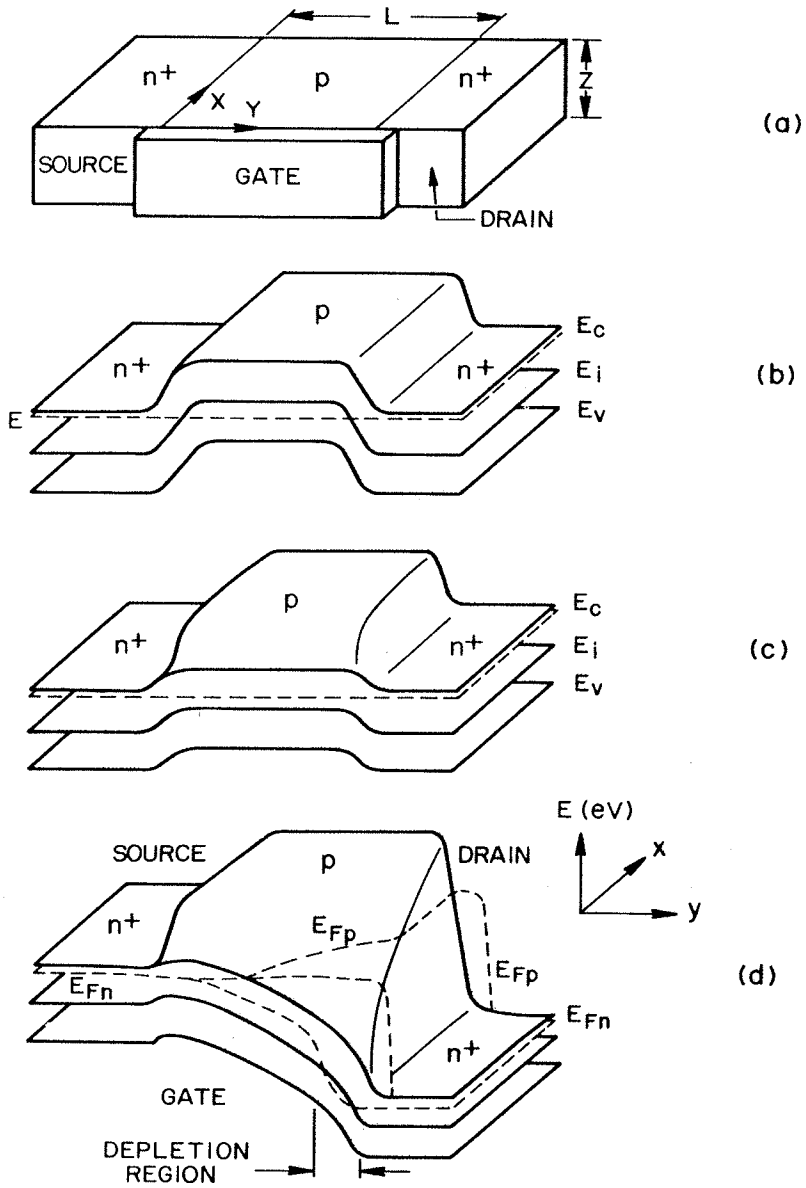

Fig. 4 Two-dimensional band diagram of an *n*-channel MOSFET. (*a*) Device configuration. (*b*) Flat-band zero-bias equilibrium condition. (*c*) Equilibrium condition under a gate bias. (*d*) Nonequilibrium condition under both gate and drain biases. (After Pao and Sah, Ref. 16.)

toward the drain contact. Figure 4*d* shows that the gate voltage required for inversion at the drain is larger than the equilibrium case in which $\psi_s(\text{inv}) \simeq 2\psi_B$. This is because the applied drain bias lowers the electron imref, and an inversion layer can be formed only when the potential at the surface crosses over the imref of the minority carrier.

Figure 5 shows a comparison[17] of the charge distribution and energy-band variation of an inverted *p* region for the equilibrium case and the nonequilibrium case at the drain. For the equilibrium case (discussed in Chapter 7), the surface depletion region reaches a maximum width W_m at

UMC 1413
UMC v. AICP

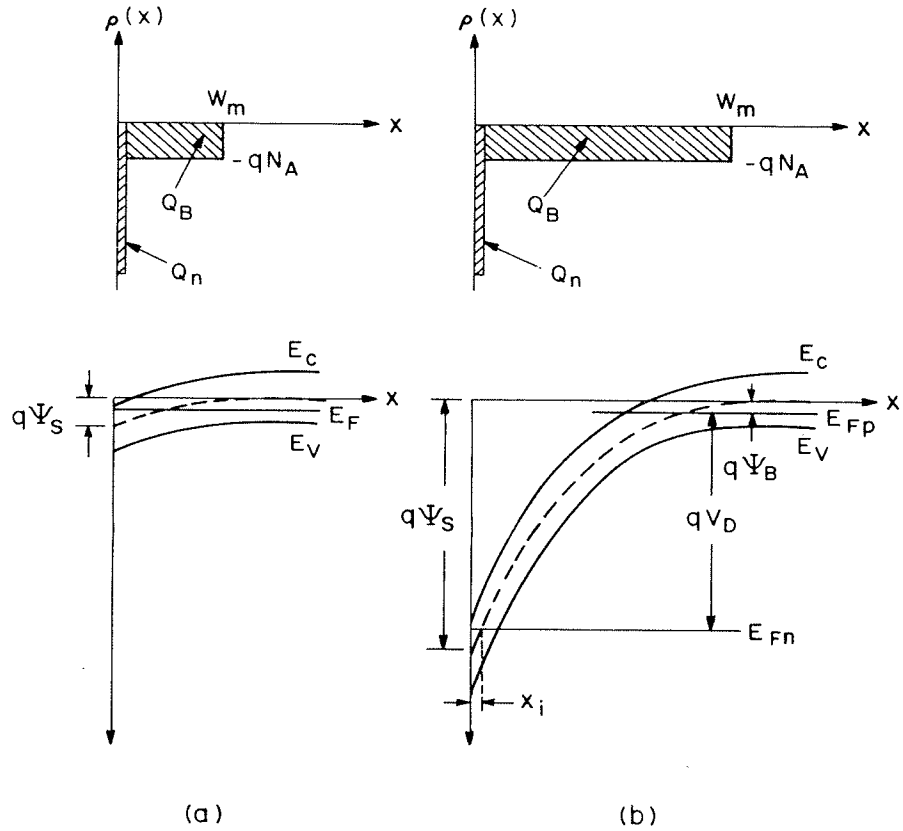

Fig. 5 Comparison of charge distribution and energy band variation of an inverted p region for (a) the equilibrium case and (b) the nonequilibrium case at the drain. (After Grove and Fitzgerald, Ref. 17.)

inversion. For the nonequilibrium case, the depletion-layer width is a function of the bias V_D, and the surface potential ψ_s at the onset of strong inversion is given, to a good approximation, by

$$\psi_s(\text{inv}) \simeq V_D + 2\psi_B \,. \tag{1}$$

The derivation for the characteristic of the surface-space charge under the nonequilibrium condition is similar to that in Chapter 7. The two assumptions are that (1) the imref for the majority carriers of the substrate does not vary with distance from the bulk to the surface, and (2) the imref for the minority carriers of the substrate is separated by the applied junction bias V_D from the imref for the majority carriers; that is, $E_{Fp} = E_{Fn} + qV_D$ for a p substrate. The first assumption introduces little error when the surface is inverted, because majority carriers are then only a negligible part of the surface space charge; the second assumption is correct under the inversion condition, because minority carriers are an important part of the surface-space-charge region when the surface is inverted.

Based on these assumptions, the one-dimensional Poisson equation for

UMC 1413
UMC v. AICP

the surface-space-charge region at the drain is given by

$$\frac{\partial^2 \psi}{\partial x^2} = -\frac{q}{\epsilon_s}(N_D^+ - N_A^- + p - n) \tag{2}$$

where

$$N_D^+ - N_A^- = n_{po} - p_{po}, \qquad p_{po} \simeq N_A$$
$$p = p_{po} e^{-\beta\psi}$$
$$n = n_{po} e^{\beta\psi - \beta V_D}, \qquad \beta \equiv q/kT. \tag{3}$$

Following the same approach as in Chapter 7, we obtain

$$\mathscr{E} = -\frac{\partial \psi}{\partial x} = \pm \frac{\sqrt{2}kT}{qL_D} F\left(\beta\psi, V_D, \frac{n_{po}}{p_{po}}\right) \tag{4}$$

and

$$Q_s = -\epsilon_s \mathscr{E}_s = \mp \frac{\sqrt{2}\epsilon_s kT}{qL_D} F\left(\beta\psi_s, V_D, \frac{n_{po}}{p_{po}}\right) \tag{5}$$

where

$$F\left(\beta\psi, V_D, \frac{n_{po}}{p_{po}}\right) \equiv \left[e^{-\beta\psi} + \beta\psi - 1 + \frac{n_{po}}{p_{po}} e^{-\beta V_D}(e^{\beta\psi} - \beta\psi e^{\beta V_D} - 1)\right]^{1/2} \tag{6}$$

and

$$L_D \equiv \left(\frac{kT\epsilon_s}{p_{po}q^2}\right)^{1/2}. \tag{7}$$

The surface charge per unit area after strong inversion is given by

$$Q_s = Q_n + Q_B \tag{8}$$

where

$$Q_B = -qN_A W_m = -\sqrt{2qN_A\epsilon_s(V_D + 2\psi_B)} \tag{9}$$

and Q_n, the charge due to minority carriers within the inversion layer, is

$$|Q_n| \equiv q\int_0^{x_i} n(x)\,dx = q\int_{\psi_s}^{\psi_B} \frac{n(\psi)\,d\psi}{d\psi/dx} \tag{10}$$

or

$$|Q_n| = q\int_{\psi_s}^{\psi_B} \frac{n_{po}e^{(\beta\psi - \beta V_D)}\,d\psi}{(\sqrt{2}kT/qL_D)F(\beta\psi, V_D, n_{po}/p_{po})} \tag{11}$$

where x_i denotes the point at which the intrinsic Fermi level intersects the imref for electrons. For the practical doping ranges in silicon, the value of x_i is quite small, of the order of 30 to 300 Å. Equation 11 is the basic formula for long-channel MOSFET, and can be evaluated numerically.

UMC 1413
UMC v. AICP

Under strong inversion conditions, a simplified expression for Q_n can be obtained from a charge-sheet model[18] and is given by

$$|Q_n| = \sqrt{2}qN_A L_D \left\{ \left[\beta\psi_s + \left(\frac{n_{po}}{p_{po}}\right) e^{(\beta\psi_s - \beta V_D)} \right]^{1/2} - (\beta\psi_s)^{1/2} \right\}. \tag{12}$$

This expression for Q_n is derived under the condition $V_{BS} = 0$. When a substrate reverse bias is applied, the depletion width increase, and the term βV_D in Eq. 12 is replaced by $\beta(V_D + V_{BS})$.

8.2.2 Linear and Saturation Regions

We shall first present a qualitative discussion of device operation. Let us consider that a voltage is applied to the gate, causing an inversion at the semiconductor surface, Fig. 6*a*. If a small drain voltage is applied, a current will flow from the source to the drain through the conducting channel. Thus the channel acts as a resistance, and the drain current I_D is proportional to the drain voltage V_D. This is the linear region. As the drain voltage increases, it eventually reaches a point at which the channel depth x_i at $y = L$ is reduced to zero; this is called the pinch-off point, Fig. 6*b*. Beyond the pinch-off point the drain current remains essentially the same, because for $V_D > V_{D\,\text{sat}}$, the voltage at Y remains the same, $V_{D\,\text{sat}}$. Thus the number of carriers arriving at point Y from the source, and hence the current flowing from source to drain, remains the same apart from a decrease in L to the value L' (Fig. 6*c*). Carrier injection from Y into the drain-depletion region is quite similar to the case of carrier injection from an emitter–base junction to the base–collector depletion region of a bipolar transistor.

We shall now derive the basic MOSFET characteristics under the following idealized conditions: (1) the gate structure corresponds to an ideal MOS diode as defined in Chapter 7; that is, there are no interface traps, fixed oxide charge, or work-function difference, and so on; (2) only drift current will be considered; (3) carrier mobility in the inversion layer is constant; (4) doping in the channel is uniform; (5) reverse leakage current is negligibly small; and (6) the transverse field ($\mathscr{E}_x$ in the x direction) in the channel is much larger than the longitudinal field ($\mathscr{E}_y$ in the y direction). The last condition corresponds to the so-called gradual channel approximation.

Under such idealized conditions, the total charge induced in the semiconductor per unit area Q_s at a distance y from the source is given by

$$Q_s(y) = [-V_G + \psi_s(y)]C_i \tag{13}$$

where $C_i \equiv \epsilon_i/d$ is the capacitance per unit area. The charge in the inversion layer is given by

$$\begin{aligned} Q_n(y) &= Q_s(y) - Q_B(y) \\ &= -[V_G - \psi_s(y)]C_i - Q_B(y). \end{aligned} \tag{14}$$

The surface potential $\psi_s(y)$ at inversion can be approximated by $2\psi_B +$

UMC 1413
UMC v. AICP

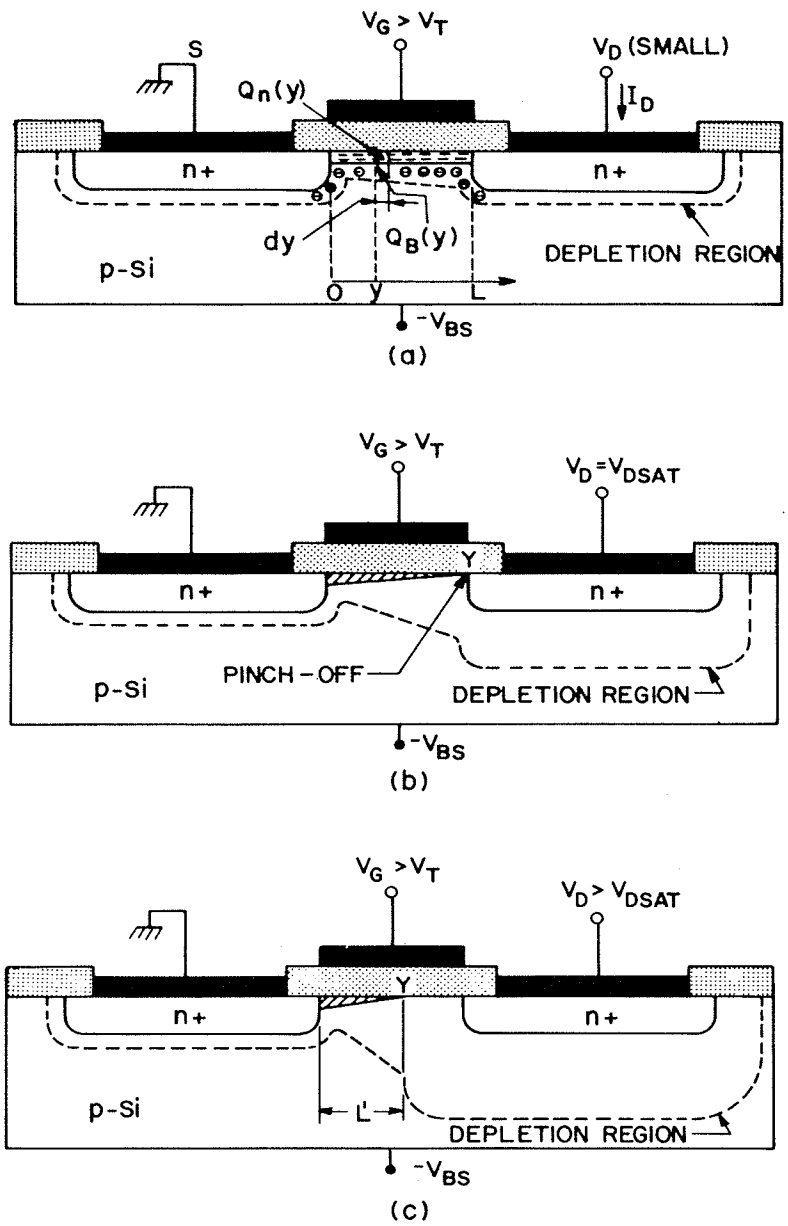

Fig. 6 (*a*) MOSFET operated in the linear region (low drain voltage). (*b*) MOSFET operated at onset of saturation. The point *Y* indicates the pinch-off point. (*c*) MOSFET operated beyond saturation and the effective channel length is reduced.

$V(y)$, where $V(y)$ is the reverse bias between point y and the source electrode (which is assumed to be grounded). The charge within the surface depletion region $Q_B(y)$ was given previously as

$$Q_B(y) = -qN_A W_m = -\sqrt{2\epsilon_s q N_A [V(y) + 2\psi_B]}. \tag{15}$$

Substituting Eq. 15 into Eq. 14 yields

$$Q_n(y) = -[V_G - V(y) - 2\psi_B]C_i + \sqrt{2\epsilon_s q N_A [V(y) + 2\psi_B]}. \tag{16}$$

UMC 1413
UMC v. AICP

The conductivity of the channel can be approximated by

$$\sigma(x) = qn(x)\mu_n(x). \tag{17}$$

The channel conductance is then given by

$$g = \frac{Z}{L}\int_0^{x_i} \sigma(x)\, dx. \tag{18}$$

For a constant mobility, the channel conductance becomes

$$g = \frac{qZ\mu_n}{L}\int_0^{x_i} n(x)\, dx = qZ\mu_n|Q_n|/L. \tag{19}$$

The channel resistance of an elemental section dy, Fig. 6a, is given by

$$dR = \frac{dy}{gL} = \frac{dy}{Z\mu_n|Q_n(y)|} \tag{20}$$

and the voltage drop across this elemental section is given by

$$dV = I_D\, dR = \frac{I_D\, dy}{Z\mu_n|Q_n(y)|} \tag{21}$$

where I_D is the drain current and is a constant independent of y. Substituting Eq. 16 into Eq. 21 and integrating from the source ($y = 0$, $V = 0$) to the drain ($y = L$, $V = V_D$) yields

$$I_D = \frac{Z}{L}\mu_n C_i \left\{\left(V_G - 2\psi_B - \frac{V_D}{2}\right)V_D - \frac{2}{3}\frac{\sqrt{2\epsilon_s qN_A}}{C_i}\left[(V_D + 2\psi_B)^{3/2} - (2\psi_B)^{3/2}\right]\right\} \tag{22}$$

for the present idealized case.

Equation 22 predicts that for a given V_G the drain current first increases linearly with drain voltage (the linear region), then gradually levels off, approaching a saturated value (the saturation region). The basic output characteristic of an idealized MOSFET is shown in Fig. 7. The dashed line indicates the locus of the drain voltage ($V_{D\,\mathrm{sat}}$) at which the current reaches a maximum value.

We shall now consider the above-mentioned two regions. For the case of small V_D, Eq. 22 reduces to

$$I_D \simeq \frac{Z}{L}\mu_n C_i\left[(V_G - V_T)V_D - \left(\frac{1}{2} + \frac{\sqrt{\epsilon_s qN_A/\psi_B}}{4C_i}\right)V_D^2\right] \tag{23}$$

or

$$I_D \simeq \left(\frac{Z}{L}\right)\mu_n C_i(V_G - V_T)V_D \qquad \text{for} \quad V_D \ll (V_G - V_T) \tag{23a}$$

where V_T (the threshold voltage) is given by

$$V_T = 2\psi_B + \frac{\sqrt{2\epsilon_s qN_A(2\psi_B)}}{C_i}. \tag{24}$$

UMC 1413
UMC v. AICP

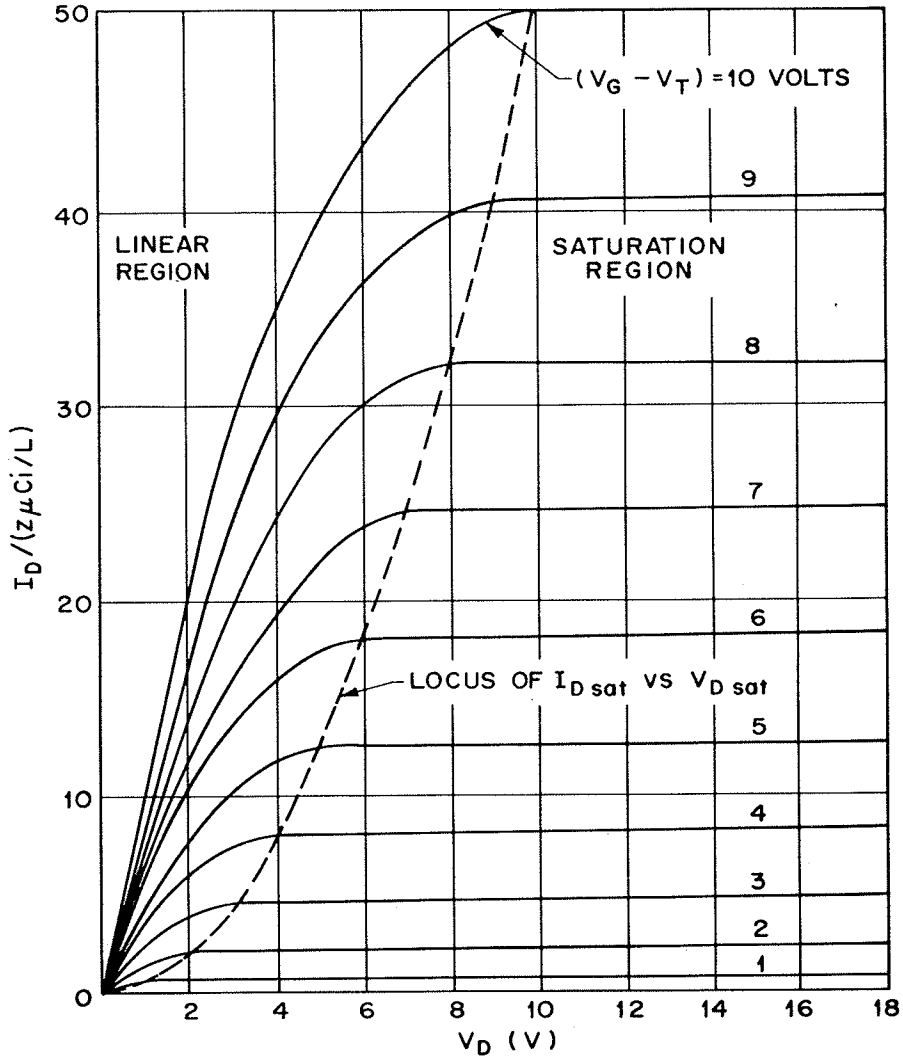

Fig. 7 Idealized drain characteristics (I_D versus V_D) of a MOSFET. The dashed line indicates the locus of the saturation drain voltage ($V_{D\,sat}$). For $V_D > V_{D\,sat}$, the drain current remains constant.

The calculated values of V_T as a function of semiconductor doping density and insulator thickness were shown in Chapter 7 for the Si–SiO_2 system. By plotting I_D versus V_G (for a given small V_D), the threshold voltage can be deduced from the linearly extrapolated value at the V_G axis. In the linear region, Eq. 23a, the channel conductance g_D and the transconductance g_m are given as

$$g_D \equiv \left.\frac{\partial I_D}{\partial V_D}\right|_{V_G=\text{const}} = \frac{Z}{L}\mu_n C_i (V_G - V_T) \tag{25}$$

$$g_m \equiv \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\text{const}} = \frac{Z}{L}\mu_n C_i V_D\,. \tag{26}$$

When the drain voltage is increased to a point such that the charge in the

UMC 1413
UMC v. AICP

inversion layer $Q(y)$ at $y = L$ becomes zero, the number of mobile electrons at the drain experiences a drastic fall-off. This point, called pinch-off, is analogous to the junction field-effect transistor. The drain voltage and the drain current at this point are designated as $V_{D\,\mathrm{sat}}$ and $I_{D\,\mathrm{sat}}$, respectively. Beyond the pinch-off point we have the saturation region. The value of $V_{D\,\mathrm{sat}}$ is obtained from Eq. 16 under the condition $Q_n(L) = 0$:

$$V_{D\,\mathrm{sat}} = V_G - 2\psi_B + K^2\left(1 - \sqrt{1 + 2V_G/K^2}\right) \tag{27}$$

where $K \equiv \sqrt{\epsilon_s q N_A}/C_i$. The saturation current $I_{D\,\mathrm{sat}}$ can be obtained by substituting Eq. 27 into Eq. 22:

$$I_{D\,\mathrm{sat}} \simeq \frac{mZ}{L}\,\mu_n C_i (V_G - V_T)^2. \tag{28}$$

where m is a function of doping concentration and approaches $\frac{1}{2}$ at low dopings.[11]

The threshold voltage V_T in the saturation region is the same as given by Eq. 24 for low substrate dopings and thin insulator layers. For higher dopings, V_T becomes V_G-dependent. The transconductance in the saturation region when Eq. 28 applies is given by

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D=\mathrm{const.}} = \frac{2mZ}{L}\,\mu_n C_i (V_G - V_T). \tag{29}$$

In previous discussions, we made many assumptions to bring out the most important characteristics of the MOSFET. We shall now remove the first two assumptions and consider the effects due to a nonideal gate MOS and diffusion current. The main effect of the fixed oxide charges and the difference in work functions is to cause a voltage shift corresponding to the flat-band voltage V_{FB}. This in turn causes a change in the threshold voltage V_T; in the linear region V_T becomes

$$\begin{aligned} V_T &= V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_s q N_A (2\psi_B)}}{C_i} \\ &= \left(\phi_{ms} - \frac{Q_f}{C_i}\right) + 2\psi_B + \frac{\sqrt{4\epsilon_s q N_A \psi_B}}{C_i}. \end{aligned} \tag{30}$$

When a substrate bias is applied, the threshold voltage becomes

$$V_T = V_{FB} + 2\psi_B + \sqrt{2\epsilon_s q N_A (2\psi_B + V_{BS})}/C_i \tag{31}$$

or

$$\begin{aligned} \Delta V_T &= V_T(V_{BS}) - V_T(V_{BS} = 0) \\ &= \frac{\sqrt{2\epsilon_s q N_A}}{C_i}\left(\sqrt{2\psi_B + V_{BS}} - \sqrt{2\psi_B}\right) \\ &= \frac{a}{\beta}\left(\sqrt{2\beta\psi_B + \beta V_{BS}} - \sqrt{2\beta\psi_B}\right) \end{aligned} \tag{32}$$

UMC 1413
UMC v. AICP

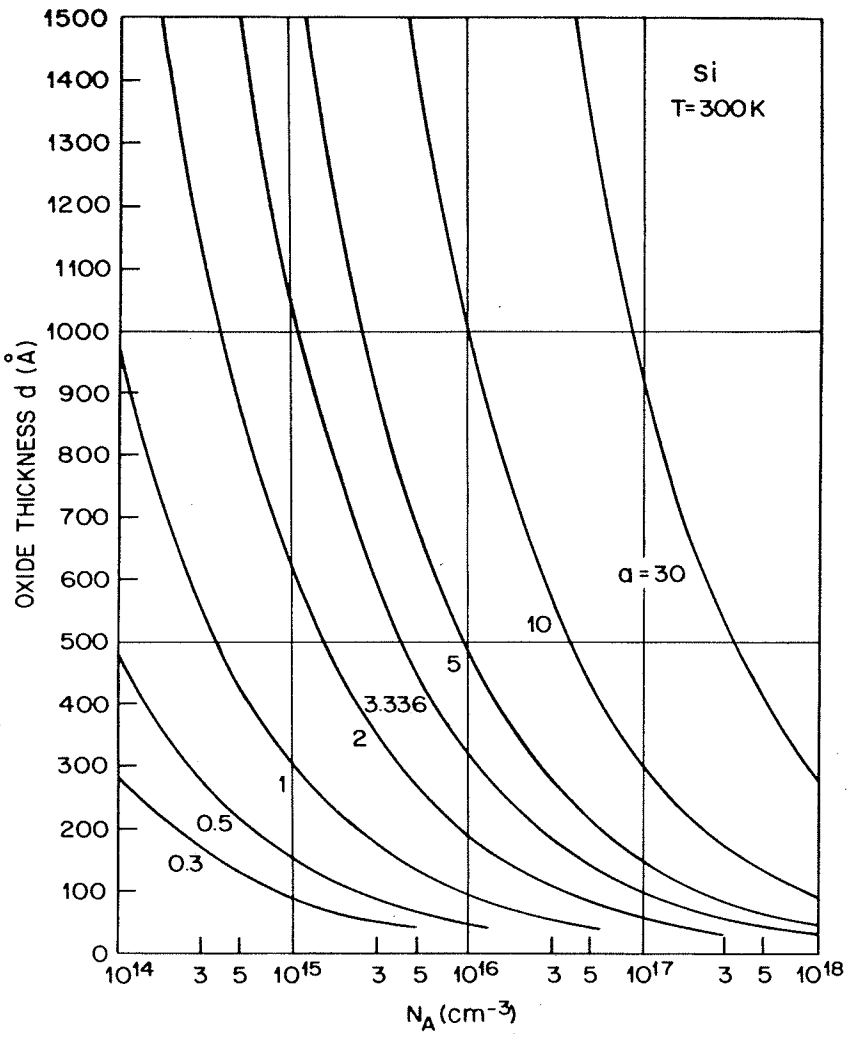

Fig. 8 Oxide thickness versus substrate doping for various a values. (After Brews, Ref. 19.)

where

$$a \equiv \sqrt{2}(\epsilon_s/L_D)/C_i = 2(\epsilon_s/\epsilon_i)(d/L_D). \tag{33}$$

In Fig. 8, oxide thickness versus substrate doping is plotted for given a values[19] using Eq. 33. The a values increase with increasing doping and oxide thickness.

Threshold voltage shift versus V_{BS} is plotted in Fig. 9 for various a values. As the a value increases, ΔV_T also increases. For a given a value, the resulting variation in ΔV_T is indicated by vertical bars for substrate dopings ranging from 10^{15} to 10^{17} cm^{-3} (Fig. 9). The primary influence upon ΔV_T is the choice of a itself; the influence of doping or oxide thickness upon ΔV_T, independent of a, is minor.

To consider the effect of the diffusion current component, we refer to Fig. 4 for the nonequilibrium condition. The drain current density including

UMC 1413
UMC v. AICP

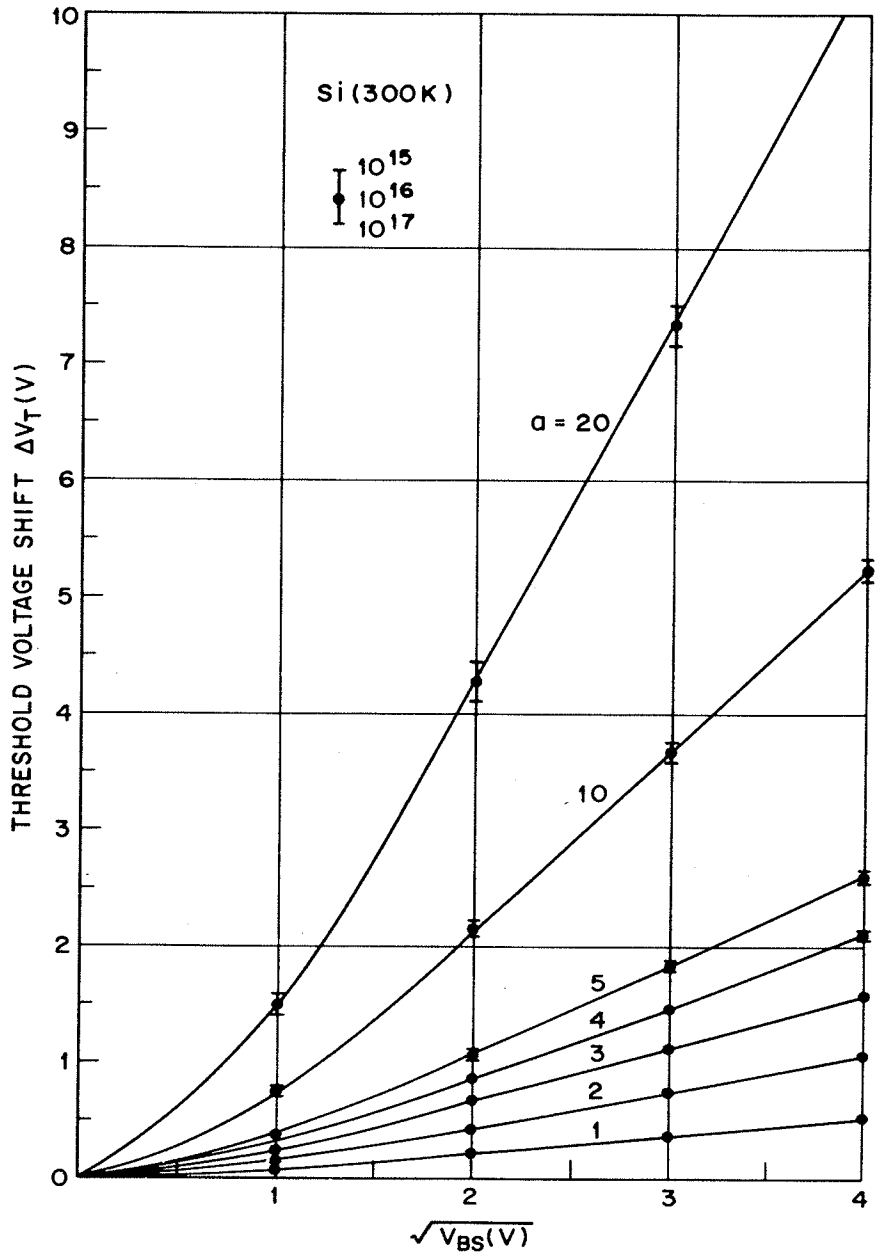

Fig. 9 Threshold voltage shift versus substrate reverse bias for various a values.

both drift and diffusion components is given by

$$J_D(x, y) = q\mu_n n \mathscr{E}_y + qD_n \nabla n$$

$$= -qD_n n(x, y) \nabla \psi_{Fn} \tag{34}$$

where ψ_{Fn} is the electron imref measured from the bulk Fermi level. The

UMC 1413
UMC v. AICP

total drain current based on the gradual-channel approximation is

$$
\begin{aligned}
I_D &= \int_0^{x_i} J_D(x, y) Z \, dx \\
&= \frac{1}{L} \int_0^L D_n q Z \left(\frac{\partial \psi_{Fn}}{\partial y} \right) \int_0^{x_i} n(x, y) \, dx \, dy \\
&= \frac{Z}{L} \frac{\epsilon_s \mu_n}{L_D} \int_0^{V_D} \int_{\psi_B}^{\psi_s} \frac{e^{\beta\psi - \beta V}}{F(\beta\psi, V, n_{po}/p_{po})} \, d\psi \, dV. \qquad (35)
\end{aligned}
$$

The gate voltage V_G is related to the surface potential ψ_s by

$$
\begin{aligned}
V_G' = V_G - V_{FB} &= -\frac{Q_s}{C_i} + \psi_s \\
&= \frac{2\epsilon_s kT}{C_i q L_D} F\left(\beta\psi_s, V, \frac{n_{po}}{p_{po}}\right) + \psi_s . \qquad (36)
\end{aligned}
$$

Equation 35 reduces to Eq. 22 for gate voltages well above threshold. Equation 22 however, becomes inaccurate for gate voltages near threshold, and near pinch-off. For a particular device with known physical dimensions, bulk impurity concentration, and effective mobility, Eq. 35 can be calculated numerically to give accurate results for the entire range of drain voltage from the linear region to the saturation region. Figure 10 demonstrates the current saturation phenomena very well, showing a typical drain characteristic for a long-channel MOSFET.[16]

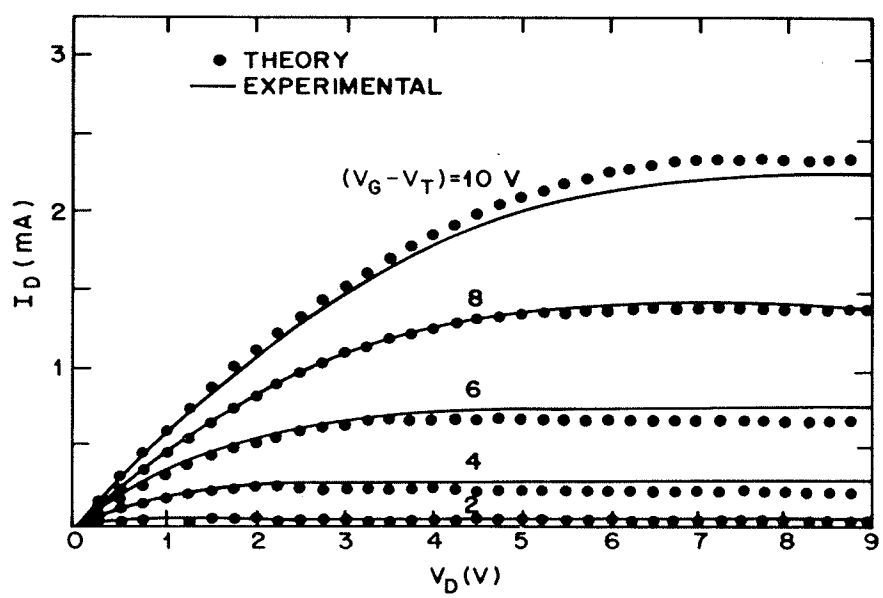

Fig. 10 Theoretical (dots) and experimental (solid lines) drain characteristics of a p-channel MOSFET having $d = 2000$ Å, $N_D = 4.6 \times 10^{14}$ cm^{-3}, and $\mu_p = 256$ cm^2/V-s. (After Pao and Sah, Ref. 16.)

UMC 1413
UMC v. AICP

8.2.3 Subthreshold Region

When gate voltage is below the threshold voltage and the semiconductor surface is in weak inversion, the corresponding drain current is called the subthreshold current.[20,21] The subthreshold region is particularly important for low-voltage, low-power applications, such as when the MOSFET is used as a switch in digital logic and memory applications, because the subthreshold region describes how the switch turns on and off.

In weak inversion, the drain current is dominated by diffusion and is derived in the same way as the collector current in a bipolar transistor with homogeneous base doping. Considering the MOSFET as an n-p-n (source–substrate–drain) bipolar transistor, we have

$$I_D = -qAD_n \frac{dn}{dy} = qAD_n \frac{n(0) - n(L)}{L} \tag{37}$$

where A is the cross section of current flow, and $n(0)$ and $n(L)$ are the electron densities in the channel at the source and the drain, respectively (Fig. 6a). These electron densities are given by

$$n(0) = n_{po} e^{\beta\psi_s} \tag{38a}$$

$$n(L) = n_{po} e^{\beta\psi_s - \beta V_D}. \tag{38b}$$

where ψ_s is the surface potential at the source. The area of current flow is given by the width Z of the device and the effective channel thickness normal to the semiconductor–insulator interface. Because of the exponential dependence of electron density on the potential ψ, the effective channel thickness corresponds to the distance in which ψ decreases by kT/q. Therefore, the effective channel thickness is $kT/q\mathscr{E}_s$ where $\mathscr{E}_s$ is the weak-inversion surface field given by

$$\mathscr{E}_s = -Q_B/\epsilon_s = \sqrt{2qN_A\psi_s/\epsilon_s}. \tag{39}$$

Substituting Eqs. 38 and 39 into 37 gives[18,22]

$$I_D = \mu_n \left(\frac{Z}{L}\right) \frac{aC_i}{2\beta^2} \left(\frac{n_i}{N_A}\right)^2 (1 - e^{-\beta V_D}) e^{\beta\psi_s} (\beta\psi_s)^{-1/2} \tag{40}$$

where we have used the relation $D_n = \mu_n kT/q$, and a is given by Eq. 33. The surface potential ψ_s is related to the gate voltage as follows:[18,19]

$$\psi_s = (V_G - V_{FB}) - \frac{a^2}{2\beta} \left\{ \left[1 + \frac{4}{a^2} (\beta V_G - \beta V_{FB} - 1) \right]^{1/2} - 1 \right\}. \tag{41}$$

Equation 40 indicates that in the subthreshold region the drain current varies exponentially with V_G, and for drain voltage V_D larger than $3kT/q$, the current becomes independent of V_D.

Equation 40 can be used to find the gate-voltage swing S, needed to

UMC 1413
UMC v. AICP

reduce the current by one decade. By definition,

$$\begin{aligned} S &\equiv \ln 10 \cdot dV_G/d(\ln I_D) \\ &= (kT/q) \ln 10 \cdot d(\beta V_G)/d(\ln I_D) \\ &= (kT/q) \ln 10 \cdot [1 + C_D(\psi_s)/C_i]\left\{1 - \left(\frac{2}{a^2}\right)[C_D(\psi_s)/C_i]^2\right\}. \end{aligned} \tag{42}$$

For $a \gg (C_D/C_i)$, the subthreshold swing becomes

$$S \simeq \frac{kT}{q} \ln 10 \cdot (1 + C_D/C_i). \tag{43}$$

The term in parentheses is the capacitive divider ratio $(C_i + C_D)/C_i$.

If there is a significant interface-trap density, the capacitance C_{it} associated with the interface traps is in parallel with the depletion-layer capacitance C_D. Using Eq. 43 and substituting $(C_D + C_{it})$ for C_D, we obtain

$$S \text{ (with interface traps)} = S \text{ (no interface traps)} \times \frac{1 + (C_D + C_{it})/C_i}{1 + C_D/C_i} \tag{44}$$

where $C_{it} = qD_{it}$ and D_{it} is the interface-trap density.

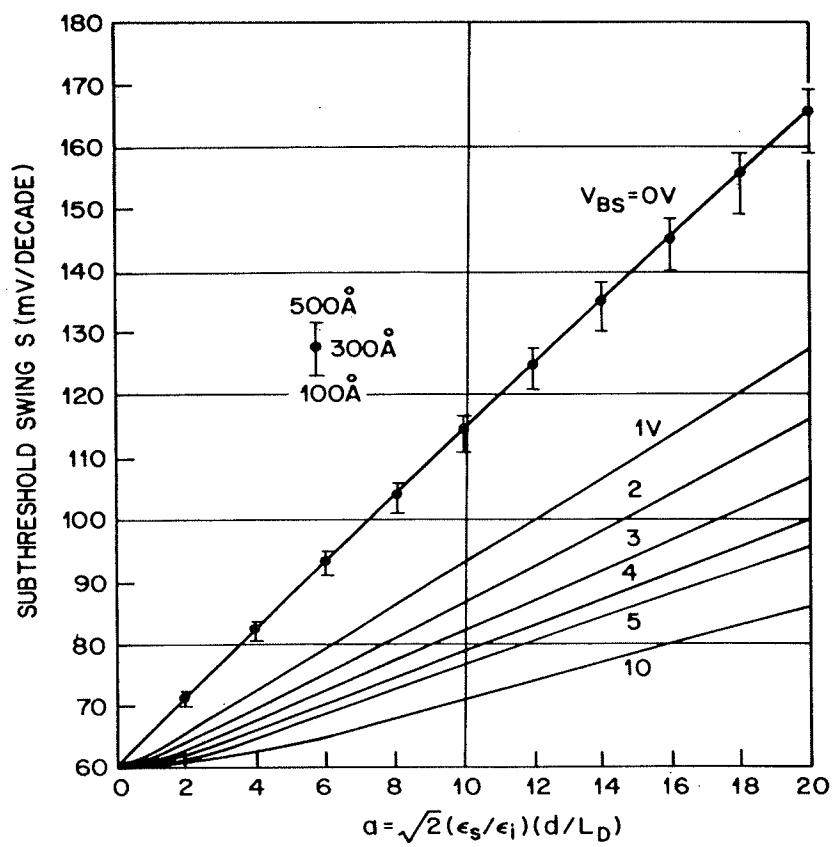

Fig. 11 Subthreshold swing versus a for various substrate reverse bias. (After Brews, Ref. 19.)

UMC 1413
UMC v. AICP

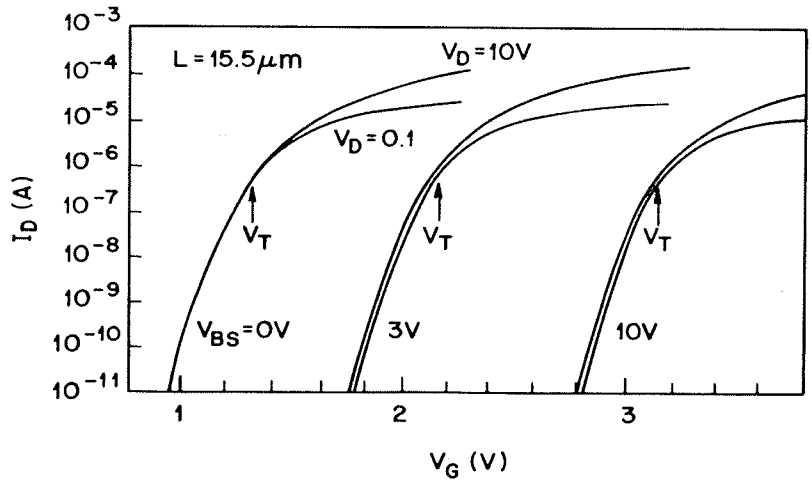

Fig. 12 Experimental subthreshold characteristics for a long-channel device ($L = 15.5\ \mu$m). (After Troutman, Ref. 23.)

When a substrate bias is applied, it increases the value of ψ_s. Consequently, the depletion-layer capacitance C_D is reduced and therefore S is reduced. Figure 11 shows the calculated subthreshold swing as a function of a for different substrate-reverse biases.[19] Again the primary influence upon S is the choice of a itself. Also, the first volt of the substrate bias results in the greatest reduction of S.

Experimental subthreshold characteristics for a long-channel (15.5 μm) MOSFET are shown[23] in Fig. 12 for three values of V_{BS}. As expected, for voltages below threshold voltage (i.e., below V_T as marked on the curves), current varies exponentially with gate voltage. For a given V_{BS}, the experimental curves for drain voltages of 0.1 V and 10 V show virtually no dependence on drain voltage in the subthreshold region. This important indication of long-channel behavior is predicted by Eq. 40. The MOSFET had a gate oxide of 570 Å and a substrate doping of $5.6 \times 10^{15}\ \text{cm}^{-3}$. The corresponding a value is 4. The calculated subthreshold swing S is 83 mV/decade for $V_{BS} = 0$, 67 mV/decade for $V_{BS} = 3$ V, and about 63 mV/decade for $V_{BS} = 10$ V (from Fig. 11). The calculated threshold voltage shift ΔV_T is 0.75 V for $V_{BS} = 3$ V and 1.7 V for $V_{BS} = 10$ V (from Fig. 9). These results are in excellent agreement with the measured values from Fig. 12.

8.2.4 Mobility Behavior

Because current flows in the inversion layer, mobility and drift velocity are expected to be influenced by the thickness of the inversion layer. When a very small longitudinal field $\mathscr{E}_y$ is applied ($\mathscr{E}_y$ is parallel to the current flow), the drift velocity varies linearly with $\mathscr{E}_y$, and the slope is the drift mobility ($v = \mu_n \mathscr{E}_y$). Experimental measurements on ⟨100⟩ p-type Si in-

UMC 1413
UMC v. AICP

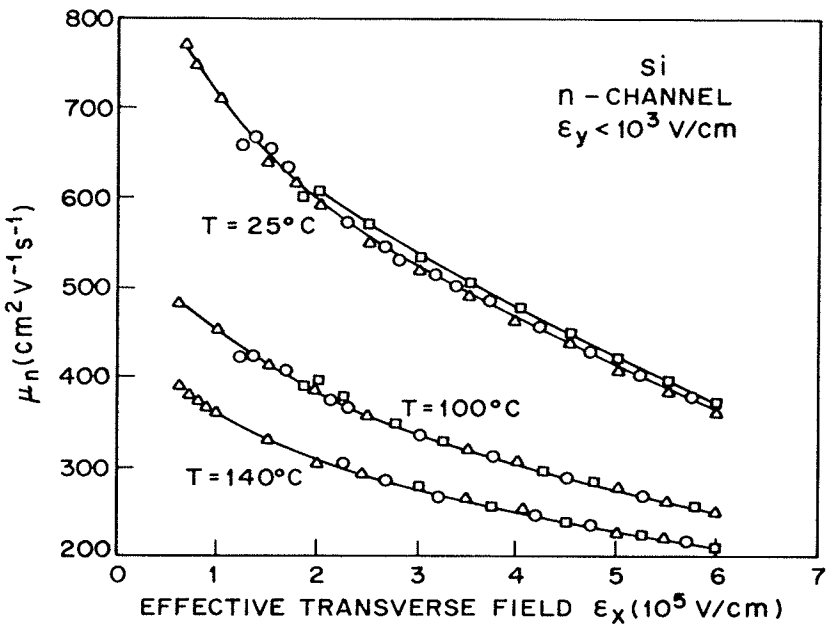

Fig. 13 Inversion layer mobility versus effective transverse field for three temperatures. (After Sabnis and Clemens, Ref. 24.)

version layers show that this mobility is a unique function of the transverse field $\mathscr{E}_x$, which is perpendicular to the current flow. This mobility is not a function of the surface processing or the doping density in the range $N_A < 10^{17}\ \text{cm}^{-3}$. The measured results are shown[24] in Fig. 13. At a given temperature, mobility decreases with increasing effective transverse field, defined as the field averaged over the electron distribution in the inversion layer, and is given by

$$(\mathscr{E}_x)_{\text{eff}} = \frac{1}{\epsilon_s}(Q_B + \tfrac{1}{2}Q_n). \tag{45}$$

When the longitudinal field increases, eventually velocity saturation occurs, similar to that of bulk silicon. The measured electron-drift velocity is shown[25,26] in Fig. 14. For a given transverse field ($\mathscr{E}_x$), the velocity is proportional to $\mathscr{E}_y$ at low longitudinal fields, and the proportionality constant is the mobility as plotted in Fig. 13. However, as $\mathscr{E}_y$ increases, the velocity tends to saturate. A general expression can be given for the drift velocity[27]

$$v_d = v_0\left[1 + \left(\frac{v_0}{v_c}\right)^2\left(\frac{v_0}{v_c} + G\right)^{-1} + \left(\frac{v_0}{v_s}\right)^2\right]^{-1/2} \tag{46}$$

where v_c, v_s, and G are fitting parameters, and

$$v_0 \equiv \mu_n(\mathscr{E}_x) \cdot \mathscr{E}_y\,. \tag{47}$$

That is, for a given transverse field $\mathscr{E}_x$, mobility is a unique function of $\mathscr{E}_x$.

UMC 1413
UMC v. AICP

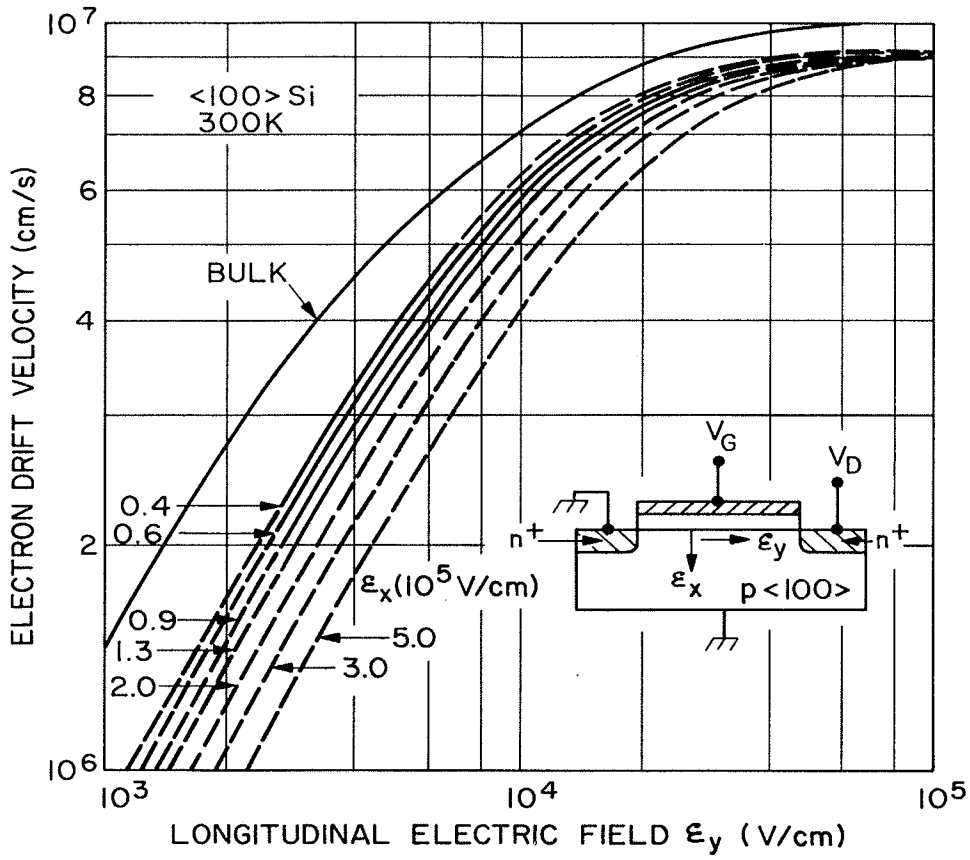

Fig. 14 Electron drift velocity versus longitudinal field for various transverse fields. The dashed curves are calculated and the solid portions indicate the regions where data were actually taken. (After Cooper and Nelson, Ref. 26).

In the limit $\mathscr{E}_y \to 0$, the drift velocity approaches $v_d = v_0 = \mu_n(\mathscr{E}_x)\mathscr{E}_y$. On the other hand, when $\mu_n\mathscr{E}_y$ is much greater than v_c and v_s, v_d is approximately equal to v_s, where v_s is a function of the transverse field.

Chapter 6 considered the effects of velocity saturation on device characteristics for the JFETs. Figure 15 shows similar results for a MOSFET. A comparison is made between the simulated current assuming a constant mobility (dashed lines) and the measured current from the same device having velocity saturation (solid lines).[28] Velocity saturation has two effects. First, saturation current is greatly reduced, especially for large gate voltages. Second, saturation current is linearly dependent on gate voltage, rather than nearly quadratic dependent as predicted by Eq. 28. Under the velocity saturation condition, the saturation current is given by

$$I_{D\,\mathrm{sat}} = ZC_i(V_G - V_T)v_s\,. \tag{48}$$

Therefore, the transconductance g_m becomes a constant:

$$g_m = (\partial I_{D\,\mathrm{sat}})/\partial V_G = ZC_i v_s\,. \tag{49}$$

UMC 1413
UMC v. AICP

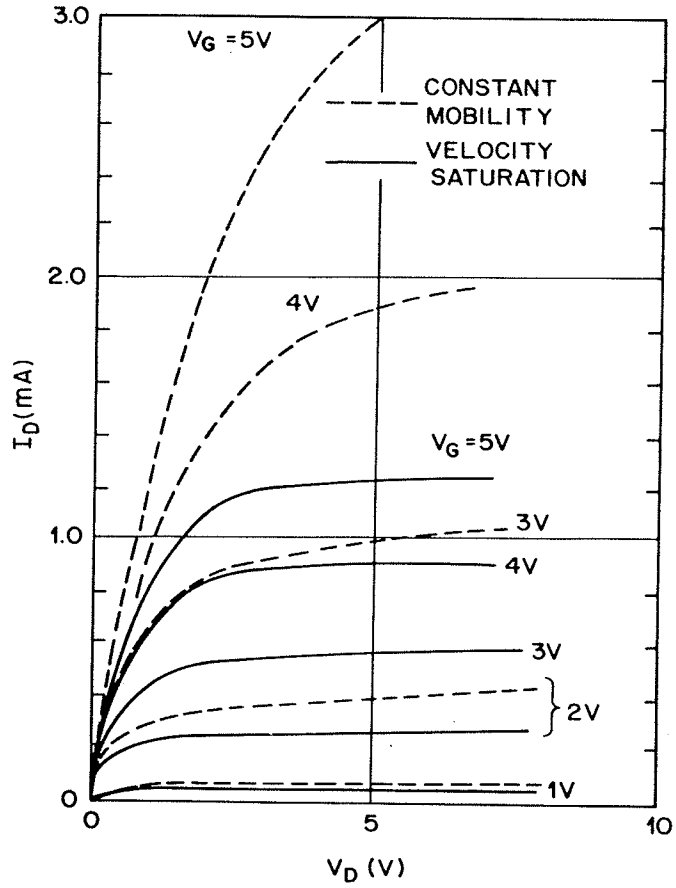

Fig. 15 Comparison of drain characteristics for constant mobility case (dashed lines) and field-dependent mobility (solid lines). (After Yamaguchi, Ref. 28.)

8.2.5 Temperature Dependence

Temperature affects device parameters and performance, especially mobility, threshold voltage, and subthreshold characteristics. The effective mobility in inversion layer has a T^{-2} power dependence on temperatures above 300 K at gate biases corresponding to strong inversion.[24]

In the linear region the threshold voltage is given by Eq. 30:

$$V_T = \phi_{ms} - \frac{Q_f}{C_i} + 2\psi_B + \frac{\sqrt{4\epsilon_s q N_A \psi_B}}{C_i}. \tag{50}$$

Because the work-function difference ϕ_{ms} and the fixed oxide charges are essentially independent of temperature, differentiating Eq. 50 with respect to temperature yields[29]

$$\frac{dV_T}{dT} = \frac{d\psi_B}{dT}\left(2 + \frac{1}{C_i}\sqrt{\frac{\epsilon_s q N_A}{\psi_B}}\right) \tag{51}$$

UMC 1413
UMC v. AICP

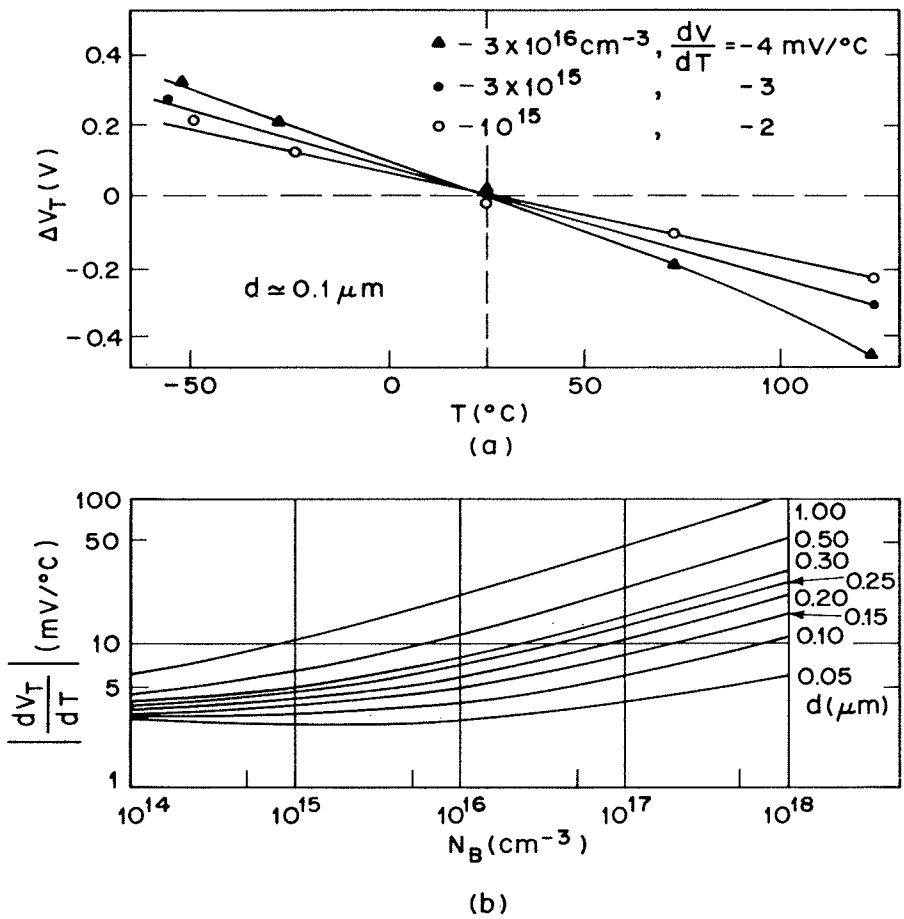

Fig. 16 (a) Experimental measurement of threshold voltage versus temperature. (*b*) dV_T/dT of a Si–SiO_2 system versus substrate doping with oxide thickness as a parameter. (After Vadasz and Grove, Ref. 29; Wang et al., Ref. 30.)

where

$$\frac{d\psi_B}{dT} \simeq \pm \frac{1}{T}\left[\frac{E_g(T=0)}{2q} - |\psi_B(T)|\right]. \tag{51a}$$

Figure 16*a* shows typical experimental measurements of threshold voltage near room temperature for the Si–SiO_2 systems.[29,30] The data can be represented by straight lines over this temperature range. Thus a representative figure for device behavior can be obtained by evaluating Eq. 51 at room temperature. Figure 16*b* shows the results of such calculations as a function of substrate doping for various values of oxide thickness. Also note that for a given oxide thickness, the quantity dV_T/dT generally increases with increased doping.

As temperature decreases, the MOSFET characteristics improve, especially in the subthreshold region. Figure 17 shows the transfer characteristics of a long-channel MOSFET ($L = 9\ \mu$m) with temperature as parameter.[31] Note that as temperature decreases from 296 K to 77 K, the

UMC 1413
UMC v. AICP

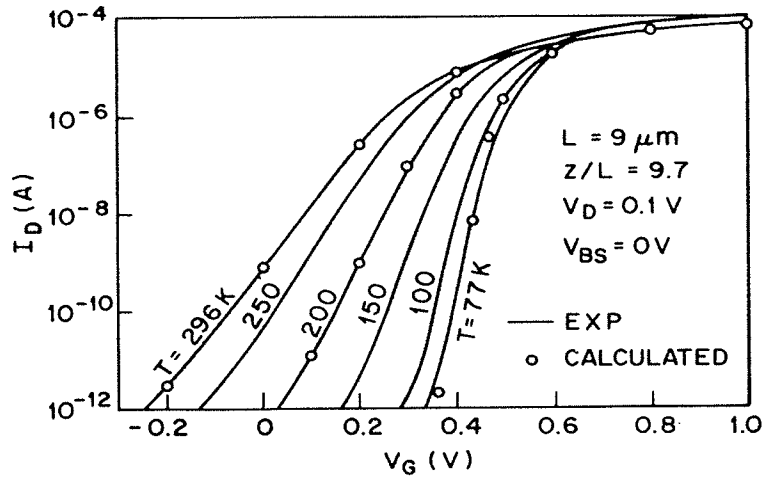

Fig. 17 Transfer characteristics for a long-channel device ($L = 9\ \mu m$) with temperature as a parameter. (After Gaensslen et al., Ref. 31.)

threshold voltage V_T increases from 0.25 V to about 0.5 V. This increase in V_T is similar to that shown in Fig. 16. The most important improvement is the reduction of the subthreshold swing S from 80 mV/decade at 296 K to 22 mV/decade at 77 K. Thus the improvement in the subthreshold swing at 77 K is about a factor of 4. This improvement comes mainly from the kT/q term in Eq. 42. Other improvements at 77 K include higher mobility, higher transconductance, higher threshold conductivity, lower power consumption, lower junction leakage current, and lower metal-line resistance. The major disadvantage is that the MOSFET must be immersed in a suitable inert coolant (e.g., liquid nitrogen) and low-temperature setup requires additional equipment.

8.2.6 Types of MOSFETs

The MOSFET is ideally a transadmittance amplifier with an infinite input resistance and a current generator at the output. In practice, however, we have other circuit parameters. An equivalent circuit is shown in Fig. 18 for the common-source connection.[32] The differential transconductance g_m was discussed previously. The input conductance G_{in} is caused by leakages through the thin gate insulator. For a thermally grown silicon dioxide layer, the leakage current between the gate and the channel is very small, of the order of 10^{-10} A/cm^2; thus the input conductance is negligible. The input capacitance C_{in} is equal to $\partial Q_M/\partial V_G$, where Q_M is the total charge on the gate.[16] In practical devices, the insulator layer and the metal gate may extend somewhat above the source and drain regions. This fringe effect will be the most important contribution to the feedback capacitance C_{fb}. The output conductance G_{out} is equal to the drain conductance. The output capacitance consists mostly of the two p-n junction capacitances connected in series through the semiconductor bulk. In the linear region, from

UMC 1413
UMC v. AICP

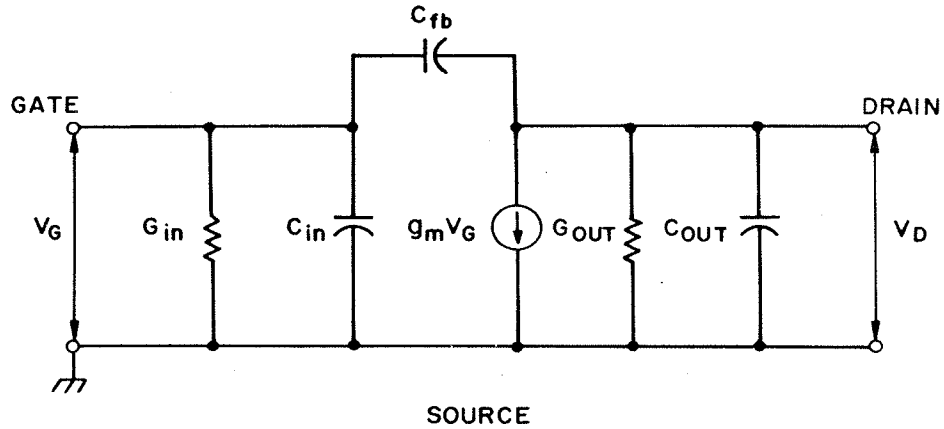

Fig. 18 Equivalent circuit of MOSFET for common-source configuration. (After Ihantola and Moll, Ref. 6.)

Eq. 26 and the fact that $C_{in} \simeq ZLC_i$, the maximum operating frequency is given by

$$f_m = \frac{\omega_m}{2\pi} = \frac{g_m}{2\pi C_{in}} \simeq \frac{\mu_n V_D}{2\pi L^2}. \tag{52}$$

In the saturation region, f_m is obtained from Eq. 49:

$$f_m \simeq \frac{v_s}{2\pi L}. \tag{53}$$

The corresponding transit time for velocity saturation is

$$\tau = \frac{L}{v_S}. \tag{54}$$

For $L = 1\ \mu$m and $v_s = 10^7$ cm/s, the transit time is only 10 ps. However, in a typical ring oscillator with 1 μm-channel MOSFETs, the measured delay time is usually an order of magnitude longer than 10 ps. Thus the delay is mainly caused by the parasitic resistance and capacitance around the device.

There are basically four different types of MOSFET, depending on the types of inversion layer. If at zero gate bias, the channel conductance is very low, we must apply positive voltage to the gate to form the n-channel. This type is the normally-off (enhancement) n-channel MOSFET. If an n-channel exists at zero bias, we must apply a negative bias to the gate to deplete carriers in the channel to reduce channel conductance. This type is called the normally-on (depletion) n-channel MOSFET. The n-channel enhancement and depletion-mode MOSFETs are shown in Fig. 19*a*. Similarly we have the p-channel normally-off (enhancement) and normally-on (depletion) MOSFET (Fig. 19*b*).

The electrical symbol, transfer characteristics, and output characteristics of the four types are shown[33] in Fig. 20. Note that for the normally-off

UMC 1413
UMC v. AICP

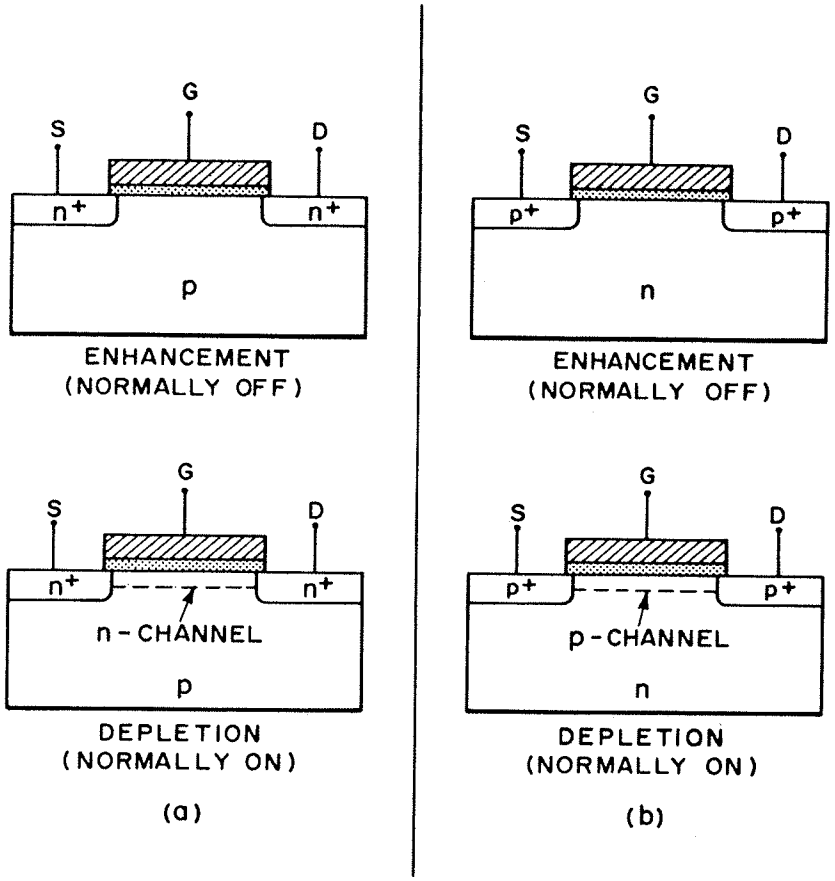

Fig. 19 Basic types of MOSFETs. (*a*) *n*-channel. (*b*) *p*-channel.

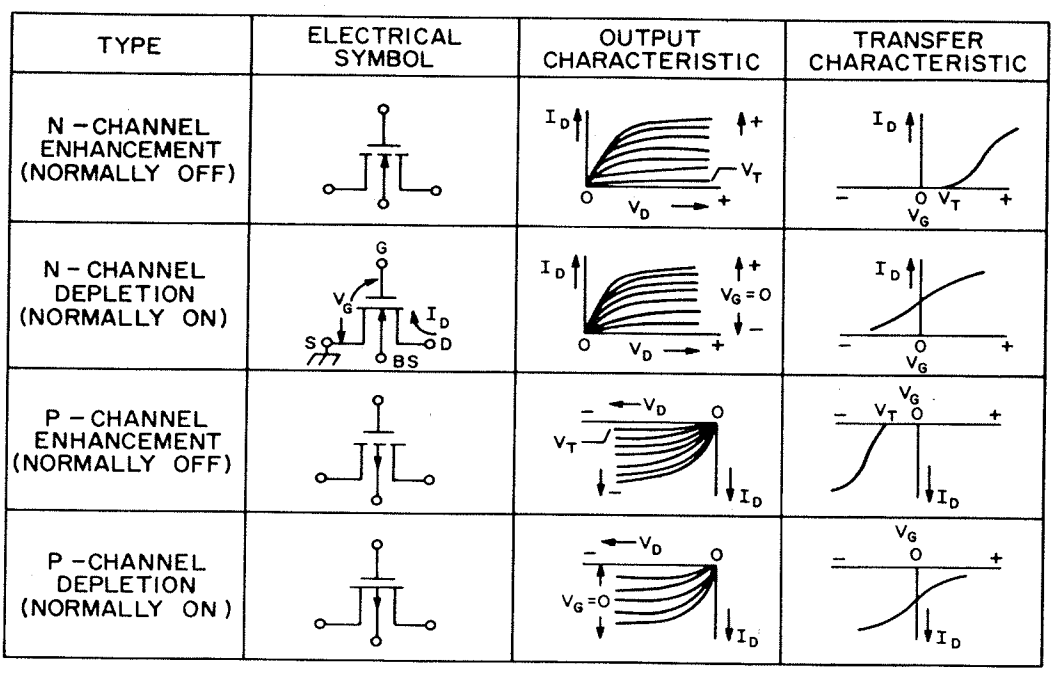

Fig. 20 Electric symbol, transfer characteristics, and output characteristics of the four types of MOSFET. (After Gallagher and Corak, Ref. 33.)

UMC 1413
UMC v. AICP

n-channel device, a positive gate bias larger than the threshold voltage V_T must be applied before a substantial drain current flows. For the normally-on n-channel device, a large current can flow at $V_G = 0$, and the current can be increased or decreased by varying the gate voltage. The discussion above can be readily extended to p-channel devices by changing polarities.

8.3 NONUNIFORM DOPING AND BURIED-CHANNEL DEVICES

In Section 8.2 doping concentration in the channel is assumed to be constant. In practical devices, however, the doping is generally nonuniform, even for doped substrates that are initially uniform, because the thermal oxidation causes impurity redistribution. Moreover, in modern MOSFET technology, ion implantation is used extensively to improve device performance. For example, ion implantation is used for (1) a self-aligned source and drain to reduce overlap capacitances, (2) a shallow dopant at the Si–SiO_2 interface for threshold voltage adjustment, (3) a channel implant on a lightly doped substrate to reduce punch-through between source and drain, and (4) a buried-channel device by incorporating within the surface region impurities of the type opposite to that of the substrate impurities.

The impurity profiles $N(x)$ in ion-implanted devices resemble a Gaussian distribution with the maximum concentration at a projected range R_p and with a standard deviation ΔR_p:

$$N(x) = \frac{D_I}{\sqrt{2\pi}\,\Delta R_p} \exp\left[-\frac{(x - R_p)^2}{2(\Delta R_p)^2}\right] \tag{55}$$

where D_I is the ion dose per unit area (Fig. 21).[34] Both projected range and

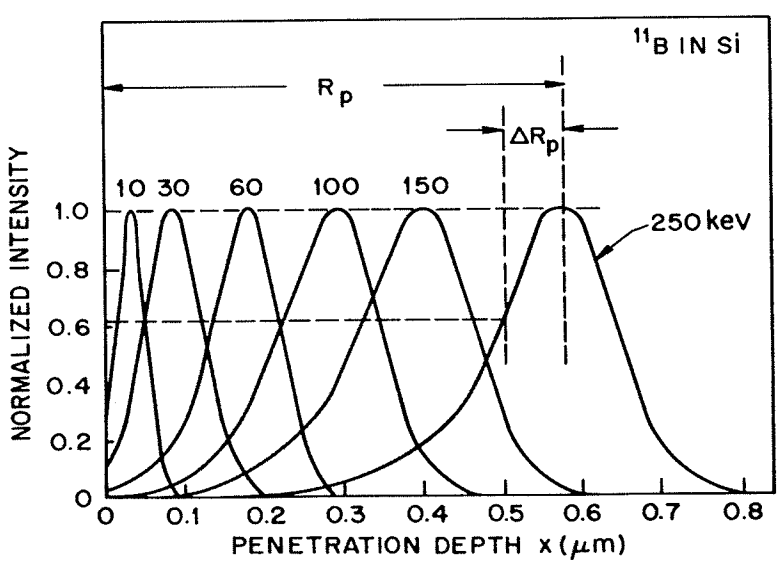

Fig. 21 Normalized range distribution of boron in silicon for different implantation energies. (After Wittmack, Maul, and Schulz, Ref. 34.)

UMC 1413
UMC v. AICP